Materials & Design 185 (2020) 108247

Contents lists available at ScienceDirect

Materials & Design

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

# Robust data-driven approach for predicting the configurational energy of high entropy alloys☆

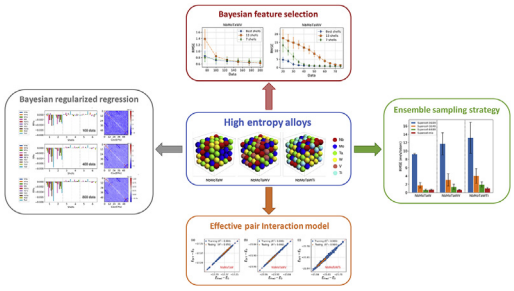
Jiaxin Zhang [a, **, 1], Xianglin Liu [b, *, 1], Sirui Bi [c], Junqi Yin [a], Guannan Zhang [d], Markus Eisenbach [a]

[a] Center for Computational Sciences, Oak Ridge National Laboratory, USA
[b] Materials Science and Technology Division, Oak Ridge National Laboratory, USA
[c] Department of Civil Engineering, Johns Hopkins University, USA
[d] Computer Science and Mathematics Division, Oak Ridge National Laboratory, USA

## HIGHLIGHTS

- A data-driven framework is proposed for predicting the configurational energy of high entropy alloys.
- Accuracy and robustness of the model are improved via physical feature selection with Bayesian information criterion.
- Uncertainty of the Bayesian regression model is quantified, with robust performance demonstrated.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Article history:*
Received 17 August 2019
Received in revised form
24 September 2019
Accepted 30 September 2019
Available online 10 October 2019

*Keywords:*
High entropy alloys
Uncertainty quantification
Bayesian regression
Bayesian information criterion

## ABSTRACT

High entropy alloys (HEAs) are promising next-generation materials due to their various excellent properties. To understand these properties, it's necessary to characterize the chemical ordering and identify order-disorder transitions through efficient simulation and modeling of thermodynamics. In this study, a robust data-driven framework based on Bayesian approaches is proposed for the accurate and efficient prediction of configurational energy of high entropy alloys. The recently proposed effective pair interaction (EPI) model with ensemble sampling is used to map the configuration and its corresponding energy. Given limited data calculated by first-principles calculations, Bayesian regularized regression not only offers an accurate and stable prediction but also effectively quantifies the uncertainties associated with EPI parameters. Compared with the arbitrary truncation of model complexity, we further conduct a physical feature selection to identify the truncation of coordination shells in EPI model using Bayesian information criterion. The results achieve efficient and robust performance in predicting the

☆ This manuscript has been co-authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US Government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

* Corresponding author.
** Corresponding author.
*E-mail addresses:* jiaxin.zhanguq@gmail.com (J. Zhang), xianglinliu01@gmail.com (X. Liu).

[1] These two authors contributed equally to this work.

https://doi.org/10.1016/j.matdes.2019.108247

First-principles calculations
Machine learning

configurational energy, particularly given small data. The developed methodology is applied to study a series of refractory HEAs, i.e. NbMoTaW, NbMoTaWV and NbMoTaWTi where it is demonstrated how dataset size affects the confidence when data is sparse.



## 1. Introduction

High entropy alloys (HEAs) consisting of four or more principal elements have been widely studied due to their exceptional mechanical properties [1–5]. The increased number of elements expands the potential candidates for next-generation materials [6–9]. Typically, the material properties are inherently linked to the chemical ordering, much efforts have been therefore devoted to analyzing the degree of chemical ordering and to identify the order-disorder phase transitions [8,10–12]. Due to the expensive time costs in experimental research, computational simulations, typically first-principles calculations, are playing an increasingly central role in the investigation of various properties of HEAs [13–16].

First-principles density functional theory (DFT) methods have been established as a powerful and reliable tool in computational material science and have enabled critical advancements in materials properties and performance discovery [17,18]. However, even with the increasing numerical efficiency and growing computing power (parallel and GPU computing), it is still difficult to address the challenge of DFT calculations in relatively large supercells (thousands of atoms) and intensive sampling (huge number of configurations) [19]. To characterize the order-disorder phase transition, a straightforward way is to combine the DFT method with Monte Carlo simulations. However, this "brute-force" method is computationally so intensive that it is often impractical for complex systems [20]. Consequently, it is quite necessary to establish an approximate configurational energy model fitted to DFT data and feed this accurate enough and efficient "surrogate" model into Monte Carlo simulations for modeling thermodynamics and order-disorder phase transitions.

A widely used approach is the cluster expansion (CE) method [21–23], which uses a discrete sum representation of material properties, for example, configurational energies, in terms of lattice site configuration and site effective cluster interactions (ECIs), such as site pairs, triplets, quadruplets and higher order interactions. The fundamental challenge in constructing a CE model is to determine the ECIs in an efficient and robust way. The commonly used fitting algorithm minimizes the overall differences between the CE fitted energies and DFT calculated energies with respect to different input configurations. Practically, CE has to be truncated and many advanced methods have been proposed to improve its efficiency and accuracy. These methods include compressive sensing [24], Bayesian method [25], cluster basis set selection [26–28], machine learning [29] and regularization [30,31]. Although CE provides an effective approach for constructing an effective Hamiltonian and feeding it to Monte Carlo simulations to investigate order-disorder phase transitions. Yet, its application to multicomponent systems is still intractable due to the rapidly increasing combinatorial number of interatomic interactions between chemical elements [28,32]. Therefore, fitting a CE for multicomponent alloys, i.e. HEAs, becomes extremely difficult [13].

The recent developments in machine learning present exciting opportunities and challenges for various scientific fields [33,34]. Benefiting from advanced learning algorithms and large databases from high-throughput computations, machine learning has been widely applied to materials research and discovery [35–38]. Some examples of successful applications include discovering complex materials behavior [39–41], accurate prediction of phase transitions and prediction [42–44], accelerated material design and prediction of material properties [45–48], modeling of various physical quantities, for instance, interatomic potentials [49–51] and atomic forces [50,52]. Compared to the success of these fields, few studies have been conducted in the context of machine learning for the modeling of thermodynamics of HEAs [44,53]. This is due to the inherent challenges originating from the extremely large configuration space associated with multicomponent alloys. In many cases, only small datasets can be obtained from expensive DFT calculations due to limited computational resources. The challenge gives rise to the issue of uncertainty quantification in model inference and certified predictions [54–59]. The learned model also faces the additional challenges to capture the underlying physics with important features and cover the overall configurational space in an accurate and robust scheme [53,60].

In this work, we develop an efficient and robust Bayesian framework to fit an accurate, feature-selected efficient Hamiltonian which is employed in subsequent Monte Carlo simulations for modeling thermodynamics and order-disorder phase transitions. Bayesian regularized regression is employed to deal with the unstable prediction due to sparse data, and moreover, to effectively quantify the uncertainties associated with the EPI parameters. To investigate the impact of model complexity, we conduct physical feature selection using a Bayesian information criterion that allows for effective truncation of the coordination shells given a specific dataset. We demonstrate that the accuracy and reliability of predictions with feature selection are significantly higher than the prediction with an arbitrary truncation, especially when data are limited. The first section of this paper presents a brief overview of classical cluster expansion and the proposed effective pair interaction model [53]. Then we propose a robust data-driven framework which consists of Bayesian regularized regression and Bayesian feature selection to effectively reduce the model complexity. In the second section, we apply the proposed robust algorithm to three refractory HEAs, i.e. NbMoTaW, NbMoTaWV and NbMoTaWTi [61]. We show that an ensemble random sampling performs better in the prediction of configurational energy compared to using only a single supercell. Due to the limitation of dataset size, we carry out uncertainty quantification and correlation analysis of model parameters in terms of a Bayesian framework. Moreover, the effect of physical feature selection is carefully investigated for these three HEAs. Finally, the conclusions of the current work are summarized.

## 2. Theory and algorithm

### 2.1. Re-visited cluster expansion method

The cluster expansion (CE) method [21–23] is a widely used approach for thermodynamic simulation of binary alloys due to its versatility and simplicity. Specifically, a binary alloy $XY$ of $N$ sites can be represented as a vector of occupation configurations, $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_k, \ldots, \sigma_N\}$ where "spin" variable $\sigma_k$ takes a value of $-1$ or $+1$ depending on the occupant (atom X or Y) of site $k$. A property of this binary crystal that depends on $\boldsymbol{\sigma}$ can be formulated as a polynomial expansion in terms of occupation configurations

$$F(\boldsymbol{\sigma}) = NV_0 + \sum_{\beta} V_{\beta}^{(n)} \Phi_{\beta}^{(n)}(\boldsymbol{\sigma}) \tag{1}$$

where the expansion coefficients $V_{\beta}^{(n)}$ are called the effective cluster interactions (ECIs) which are independent of the configurations and determined by the crystal structure and chemistry of the binary alloy, $V_0$ is a constant that represents the empty cluster and $\Phi_{\beta}^{(n)}(\boldsymbol{\sigma})$ is the $n$-site cluster function, defined as the product of basis function $\Theta_{\beta_k}(\sigma_k)$, which is given by:

$$\Phi_{\beta}^{(n)}(\boldsymbol{\sigma}) = \prod \Theta_{\beta_k}(\sigma_k) \tag{2}$$

Note that the cluster functions in Eq. (2) form a complete orthonormal basis on the configuration space $\boldsymbol{\sigma}$. When all possible cluster functions are considered in the CE model, Eq. (1) is an exact expression. However, a more practical way for CE is a truncated summation over a finite number of cluster functions considering negligible long range interactions. Typically, the energy is primarily determined by short-range interactions, it is therefore natural to represent the energy as a summation of interactions whose strength diminishes with increasing range, which is given by:

$$E(\boldsymbol{\sigma}) = \sum_{i} V_i \Phi_i(\sigma_i) + \sum_{ij} V_{ij} \Phi_{ij}(\sigma_i, \sigma_j) + \sum_{ijk} V_{ijk} \Phi_{ijk}(\sigma_i, \sigma_j, \sigma_k) + \cdots \tag{3}$$

where the $V_i$, $V_{ij}$ and $V_{ijk}$ represent the interaction strength of point clusters, pair clusters and triplet clusters, and can be determined by the DFT calculated total energies of different configurations in a variety of supercells. One can therefore utilize CE as an efficient Hamiltonian in Monte Carlo simulations to reveal order-disorder phase transitions.

However, it is often a challenging task to fit the ECIs of CE in multicomponent crystalline solids because of the number of terms scales as $N^M$ for an $M$-body terms for $N$ species [13]. The series need many terms and the number of terms grows rapidly with the diameter of the cluster. As additional terms are added, the series coefficients may converge poorly given limited number of configurations.

### 2.2. Effective pair interaction model

Conventional CE method is generally difficult when applied to HEAs, we thus propose to use an Ising-like model with only effective pair interactions (EPIs) without considering high-order interactions [53]. Fig. 1 shows a prototype square lattice with effective pair interactions. We define a series of short-range pair interactions in terms of the pair distance, for example, the nearest-neighbor, the next nearest-neighbor and so on. The local energy of site $i$ is made up of all the pair interactions involving $i$. Therefore, the effective Hamiltonian at lattice site $i$ can be expressed as:

$$H(i) = J_0 + \sum_{j \neq i} J_m^{X(i)Y(j)} c_j \tag{4}$$

where $J_m^{X,Y}$ is the interatomic pair potential between element $X$ and $Y$, $X(i)$ is referred to as element $X$ at site $i$, $m$ is the number of coordination shellsseparating between $i$ and $j$, $c_j$ is the occupation parameter of site $j$, and $J_0$ is the concentration dependent part, which can be discarded for a given composition. Summing up the Hamiltonian over all atomic sites yields the total energy, which is given by:

$$\tilde{E}(\boldsymbol{\sigma}) = NJ_0 + N \sum_{X,Y,m} J_m^{X,Y} \sigma_m^{X,Y} \tag{5}$$

where $\sigma_m^{X,Y}$ is the percentage of $XY$ bonds in the $m$-th coordination shells. Considering an $n$-component alloy system, the total number of different chemical bonds in $m$-th shell is $n(n+1)/2$ but there are $n$ constraints from the concentration of each element for a fixed chemical composition. As a result, the number of independent variables in an $n$-component alloy system is

$$N_m = \frac{n(n+1)}{2} - n = \frac{n(n-1)}{2} \tag{6}$$

which consists of the short-range order (SRO) parameters that exist at $m$-th shell for an $n$-component alloy. The Warren-Cowley SRO parameters is defined as

$$\alpha_m^{X,Y} = 1 - \frac{P_m^{X|Y}}{c_A} \tag{7}$$

where $c_A$ is the concentration of element $X$, and $P_m^{X|Y}$ is the probability of finding element $X$ at the $m$-th neighbor shell of element $Y$. $\alpha_m^{XY}$ is a crucial parameter to characterize the different chemical configurations. $\alpha_m^{XY} > 0$ signifies the preference to form $XY$ bonds at the $m$-th shell, $\alpha_m^{XY} < 0$ indicates the opposite and $\alpha_m^{XY} = 0$ for each $m$ suggests to a completely random system. In fact, there is an effective pair interaction (EPI) corresponding to each SRO parameter. Consequently, Eq. (5) can be written as

$$E = N \sum_{X \neq Y,m} J_m^{X,Y} P_m^{X|Y} \tag{8}$$

where $P_m^{X|Y}$ is closely related to the SRO parameter in Eq. (7). For example, for the four-component Nb–Mo–Ta–W refractory HEAs, there are a total of ten different bonds for each coordination shell but only six independent bonds, namely Nb–Mo, Nb–Ta, Nb–W, Mo–Ta, Mo–W and Ta–W. Given a specific configuration of multicomponent HEAs Nb–Mo–Ta–W, it is not difficult to calculate the $P_m^{Nb|Mo}$, $P_m^{Nb|Ta}$, $P_m^{Nb|W}$, $P_m^{Mo|Ta}$, $P_m^{Mo|W}$ and $P_m^{Ta|W}$ at the $m$-th neighbor shell. The corresponding interatomic pair coefficients $V_m^{Nb|Mo}$, $V_m^{Nb|Ta}$, $V_m^{Nb|W}$, $V_m^{Mo|Ta}$, $V_m^{Mo|W}$ and $V_m^{Ta|W}$ at the $m$-th neighbor shell in Eq. (8) can be determined by linear regression using $P_m^{X|Y}$ as the features [53]. The cost of building an EPI model comes primarily from the cost of generating the training dataset from DFT calculations. This therefore gives rise to two important questions: 1) how to conduct an accurate and robust prediction that minimizes the error of energy for a given training dataset and 2) how to determine the number of physical feature $m$ when data are sparse. This leads to a cluster selection problem of finding the optimal set of clusters in a robust way.

### 2.3. Bayesian regularized regression

To obtain an EPI for a specific HEA one must determine the interatomic pair potential $J_m^{X,Y}$ in Eq. (8), which can be cast into a matrix form,

$$\boldsymbol{E} = \boldsymbol{JP} + \boldsymbol{\varepsilon} \tag{9}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ are independent and identically distributed (i.i.d.) variables that follow $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$. $\boldsymbol{P}$ is a matrix containing the probability quantities of the training data where each element in row $i$ at the $m$-th shell is defined as

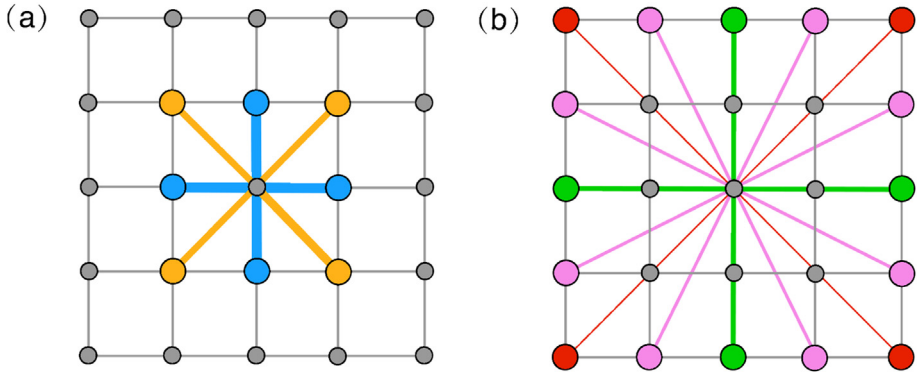



**Fig. 1.** Square lattice with effective pair interaction highlighted. (a) The nearest-neighbor pair is marked in blue, while the next nearest-neighbor pair is marked in yellow; (b) the pair marked in green, pink and red correspond to the 3rd, 4th and 5th neighbor respectively. Equivalent interacted pairs (same distance) are marked in the same color.

$$\boldsymbol{P}^i_m = P^{X|Y}_m(\boldsymbol{\sigma}_i), \quad i = 1, \ldots, N_\sigma \tag{10}$$

where $N_\sigma$ is the number of training data (configurations). It is necessary to note that there are $N_m = n(n-1)/2$ column vectors for each shell and $\boldsymbol{P}$ is therefore a $N_\sigma \times N_m$ matrix. $\boldsymbol{E}$ is a column vector in which the $i$th element is the physical quantity $E_i$ (for example, total energy) of the configuration $\boldsymbol{\sigma}_i$ and $\boldsymbol{J}$ is a column vector in which $m$-th shell is $J_m$ which also includes $N_m = n(n-1)/\ 2$ elements. Determining the EPIs by solving the linear system given by Eq. (10) is equivalent to finding the parameter vector $\boldsymbol{J}$, which minimizes the residual sum of squared errors (RSS) $||\boldsymbol{JP} - \boldsymbol{E}||_2^2$ using an ordinary least squares (OLS) method. Typically, the OLS method often has low bias but larger variance. A solution determined by OLS method usually performs well in an overdetermined system. Due to the large computational cost in DFT calculations, the linear system in Eq. (10) is, however, often underdetermined and therefore leads to an ill-posed problem. Another drawback associated with the OLS method is its susceptibility to possible overfitting [30], which refers to an EPI whose values are over-tuned to predict physical quantity in training dataset but are losing the predictability for the new configurations that have not been considered before. Meanwhile, the nearsightedness of physical interactions in CE suggests a sparsity property for $\boldsymbol{J}$ [24].

Regularization is an effective way to counteract overfitting and achieve sparse solutions by adding a regularization term in the form of $\ell_1$ or $\ell_2$ norm. For $\ell_1$ regularization, the optimal EPI values $\widehat{\boldsymbol{J}}$ can be obtained by

$$\widehat{\boldsymbol{J}} = \underset{\boldsymbol{J}}{\operatorname{argmin}} \boldsymbol{JP} - \boldsymbol{E}_2^2 + \lambda_1 ||\boldsymbol{J}||_1 \tag{11}$$

where $\lambda_1$ is a penalty parameter that determines the amount of regularization. The primary benefit of $\ell_1$ regularization is its promotion of sparsity, which is achieved by feature selection with a set of EPI values set to zero. However, in principle, this shrinkage sometimes incorrectly forces the EPI parameters to zero, consequently leads to an unstable prediction, specifically under the case of a sparse training dataset. Instead, this study uses the $\ell_2$ penalty for both fitting and penalization of the EPI coefficients. Thus, the solution becomes

$$\widehat{\boldsymbol{J}} = \underset{\boldsymbol{J}}{\operatorname{argmin}} \boldsymbol{JP} - \boldsymbol{E}_2^2 + \lambda_2 \boldsymbol{J}_2^2 \tag{12}$$

which is the most popular technique for improving prediction accuracy by penalizing large regression coefficients to reduce overfitting. Unlike $\ell_1$ regularization, $\ell_2$ regularization tends to contain all physical interaction information by only shrinking the size of EPI coefficients rather than setting most of them to zero. It therefore gives rise to challenges in optimally determining the $\ell_2$ regularization parameter and physically identifying the important features. Moreover, it is critical to quantify the uncertainties associated with the prediction and EPI coefficients, particularly given the small size of the training dataset.

In this paper, we propose a Bayesian view of regression with feature selection to address these challenges. Bayesian regression assumes the parameters $\boldsymbol{J}$ and $\sigma^2$ in Eq. (9) to be random variables, therefore the likelihood function can be written as:

$$p\left(\boldsymbol{E}\middle|\boldsymbol{J}, \boldsymbol{P}, \sigma^2\right) \propto \left(\sigma^2\right)^{-n/2} \exp\left( -\frac{1}{2\sigma^2}(\boldsymbol{E} - \boldsymbol{JP})^T(\boldsymbol{E} - \boldsymbol{JP})\right) \tag{13}$$

Bayesian regression can be also used to take $\ell_2$ regularization into consideration in the estimation procedure. Instead of identifying the optimal $\lambda_2$ in a hard sense, Bayesian regression treats the regularization parameter $\lambda_2$ as a random variable that can be estimated from the training data. This can be achieved by introducing a hierarchical model for the hyper-parameters of the model. In the Bayesian setting, the target total energy $E$ is assumed to be a Gaussian distribution, which is given by:

$$p(\boldsymbol{E}|\boldsymbol{J}, \boldsymbol{P}, \lambda_2) = \mathcal{N}(\boldsymbol{E}|\boldsymbol{JP}, \lambda_2) \tag{14}$$

and the prior for the EPI coefficient $\boldsymbol{J}$ is given by a Gaussian distribution

$$p\left(\boldsymbol{J}|\xi\right) = \mathcal{N}\left(\boldsymbol{J}|0, \xi^{-1}\boldsymbol{I}_p\right) \tag{15}$$

Consequently, the $\ell_2$ regularization in Eq. (12) is equivalent to finding a maximum a posterior (MAP) estimation [62] given a Gaussian prior over $\boldsymbol{J}$ with precision $\xi^{-1}$. Typically, a MAP estimation of the posterior distribution is obtained by Markov Chain Monte Carlo (MCMC) algorithm, which is often computationally intensive and difficult to converge for high dimensional problem. In this work, we consider a conjugate prior for which the posterior distribution can be derived analytically. To this end, the priors over $\lambda_2$ and $\xi$ are selected to be a gamma distribution

$$\lambda_2 \sim \mathcal{G}(\alpha_1, \alpha_2), \quad \xi \sim \mathcal{G}(\beta_1, \beta_2) \tag{16}$$

where $\alpha_1, \alpha_2, \beta_1$ and $\beta_2$ are the hyperparameters of the gamma priors over $\lambda_2$ and $\xi$. Here we select $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 10^{-8}$ to be non-informative priors. All three random variables $\boldsymbol{J}$, $\lambda_2$ and $\xi$ are estimated jointly using a maximum likelihood estimate during the fit of the regression model. Note that Bayesian regularized regression performs more robust for ill-posed problems in training.

### 2.4. Bayesian feature selection

To construct the effective Hamiltonian in HEAs, an EPI model is employed to identify the coordination shells as the essential physical features. For each shell, there are $N_m = n(n-1)/2$ independent sub-features that are determined by the number of element species, for instance, $N_m = 6$ for four-component Nb−Mo−Ta−W HEAs and $N_m = 10$ for five-component Nb−Mo−Ta−W−V HEAs. Liu and Zhang [53] have shown that the first two shells associated with nearest neighbor pairs and the next nearest neighbor pairs have a more significant impact on the accuracy of the prediction but long-range pair interactions with a larger $m$ show weak influence. In practice, the truncation of coordination shells $m$, needs to be carefully examined and determined for each specific material. In this work, a Bayesian model selection method is applied to identify a "better" model complexity among a finite set of candidate EPI models.

The Bayesian information criterion (BIC) is widely used for feature selection and it measures the efficiency of the parameterized model for predicting the data, which is defined as

$$\mathrm{BIC} = -2\log(\widehat{L}) + k\log(n_d) \tag{17}$$

where $\widehat{L}$ is the maximal value of the model likelihood function, i.e. $\widehat{L} = p(\boldsymbol{d}|\boldsymbol{J}^*, M)$, where $\boldsymbol{J}^*$ are the parameter values that maximize the likelihood function. $n_d$ is the number of observed data $\boldsymbol{d}$, $k$ is the number of parameters (features) of the model $M$. BIC can be derived by an efficient approximation using Laplace's approach to approximate the evidence $p(\boldsymbol{d}|M)$, which is defined by Bayesian inference [63].

$$p(\boldsymbol{J}|\boldsymbol{d}, M) = \frac{p(\boldsymbol{d}|\boldsymbol{J}, M)p(\boldsymbol{J}|M)}{p(\boldsymbol{d}|M)} = \frac{p(\boldsymbol{d}|\boldsymbol{J}, M)p(\boldsymbol{J}|M)}{\int p(\boldsymbol{d}|\boldsymbol{J}, M)p(\boldsymbol{J}|M)d\boldsymbol{J}} \tag{18}$$

BIC can be extended for linear regression under the assumption that the model errors $\boldsymbol{\varepsilon}$ are i. i.d. Random variables that follow a Gaussian distribution $N(0, \sigma^2)$. The likelihood of $\boldsymbol{\varepsilon}$ can be written as

$$L = \prod_{i=1}^{n_d} \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{(\boldsymbol{E}_i - \boldsymbol{J}\boldsymbol{P}_i)^2}{2\sigma^2} \right) = \frac{1}{(2\pi)^{\frac{n_d}{2}}\sigma^{n_d}} \exp\left( -\frac{\sum_{i=1}^{n_d}(\boldsymbol{E}_i - \boldsymbol{J}\boldsymbol{P}_i)^2}{2\sigma^2} \right) \tag{19}$$

Note that $\sum_{i=1}^{n_d}(\boldsymbol{E}_i - \boldsymbol{J}\boldsymbol{P}_i)^2 = \boldsymbol{J}\boldsymbol{P} - \boldsymbol{E}_2^2$ is the RSS of $\boldsymbol{\varepsilon}$. Taking the derivative of $L$ with respect to σ and equating it to zero yields the maximal value of $L$ and the corresponding log of $\widehat{L}$ is $\log(\widehat{L}) = -n_d/2\log(\mathrm{RSS}/n_d)$. As a result, the BIC in terms of RSS is given by

$$\mathrm{BIC}_{\mathrm{RSS}} = n_d\log\left(\frac{\mathrm{RSS}}{n_d}\right) + k\log(n_d) \tag{20}$$

This $\mathrm{BIC}_{\mathrm{RSS}}$ can be used to identify an appropriate number of the physical feature, the coordination shells in the EPI model, according to the intrinsic complexity present in a particular dataset. When selecting from a set of candidate models with different numbers of features, the one with the lowest $\mathrm{BIC}_{\mathrm{RSS}}$ value is preferred.

### 2.5. Robust data-driven algorithm procedure

With the constituents outlined in the previous sections, the proposed methodology is summarized here and a flowchart is provided in Fig. 2.

• Step 1: *Data collection* - An ensemble sampling strategy is used to combine the DFT data calculated with different sizes of the supercell. This benefits from incorporating different long-range order and short-range order datasets.

• Step 2: *Feature identification* - Given a set of random configurations of a $n$-component alloy, the first step is to determine the $N_m = n(n-1)/2$ independent pair interactions according to the EPI model. Then the probabilities $P_m^{X|Y}$, as physical features, defined in Eq. (7), are calculated for the $m$-th coordination shells.

• Step 3: *Feature selection* - To achieve an accurate prediction and reduce overfitting, feature selection using BIC in Eq. (20) is employed here to determine the truncated number of coordination shells $m$ for a specific HEA. Note that each shell also includes $N_m = n(n-1)/2$ sub-features, which are either fully retained or truncated as a whole in the $m$-th shell.

• Step 4: *Bayesian regularized regression* - Bayesian regression with $\ell_2$ regularization performs a robust prediction for the configurational energy $\boldsymbol{E}_i$ given a specific configuration $\boldsymbol{\sigma}_i$. Under the assumption of Gaussian distribution with conjugate prior, the uncertainty of the EPIs parameters are efficiently quantified, particularly given limited DFT data due to its prohibitively computational cost.

• Step 5: *Error evaluation and model update* - The performance of the predicted model can be assessed by k-fold cross validation (k = 5) with the root-mean-square error (RMSE) metric $\varepsilon_{\mathrm{R}}$, which is defined as:

$$\varepsilon_{\mathrm{R}} = \left( \frac{1}{n_d} \sum_{i=1}^{n_d} \left( E_i^{DFT} - E_i^{Pred} \right)^2 \right)^{1/2} \tag{21}$$

where $E^{DFT}$ is the true value of energy from DFT and $E^{Pred}$ is the predicted energy using the proposed methodology. If the RMSE metric $\varepsilon_{\mathrm{R}}$ is smaller than a specific threshold $\bar{\varepsilon}$, for example, $\bar{\varepsilon} = 1$ meV, the predictive accuracy is acceptable, otherwise additional DFT data are required and the algorithm returns to Step 1 to update the model and further improve the performance.

• Step 6: *Monte Carlo calculations of thermodynamics* - If the prediction is accurate and stable enough, we can use this fitted model as a surrogate in Monte Carlo simulations to model thermodynamics and order-disorder phase transitions. This step is not the focus of this paper and the interested reader can find more discussions in the recent review literature [12,13]

## 3. Results and discussions

### 3.1. Prediction of configurational energy

In this work, we systematically investigate three HEAs, namely NbMoTaW, NbMoTaWV and NbMoTaWTi. The locally self-consistent multiple scattering (LSMS) method [64] is used here for the calculation of the total energy, with supercells of 16, 32, 64 and 128 for NbMoTaW and 20, 40, 80 and 160 for NbMoTaWV and NbMoTaWTi respectively. Fig. 3 shows the BCC supercell lattice of NbMoTaW (128 atoms), NbMoTaWV and NbMoTaWTi (160 atoms). For each supercell size, 200 configurations are randomly drawn and the corresponding energy is calculated by the DFT method. Three smaller supercells with a total of 600 configurations are selected as the training dataset for Bayesian regularized regression and the largest supercell with 200 configurations are chosen for testing purpose. Six coordination shells in EPI model are chosen for this case. Fig. 4 shows a comparison of predicted energy with DFT calculated energy for three HEAs and the corresponding training and testing RMSE are illustrated in Table 1. For these three HEAs, the testing RMSEs $\varepsilon_R \sim 0.6$ meV show that the learned model is accurate and robust for a system described by a relatively large supercell size.

Typically, with the increasing of supercell size, the configurational systems tend to transition from ordered to disordered state.

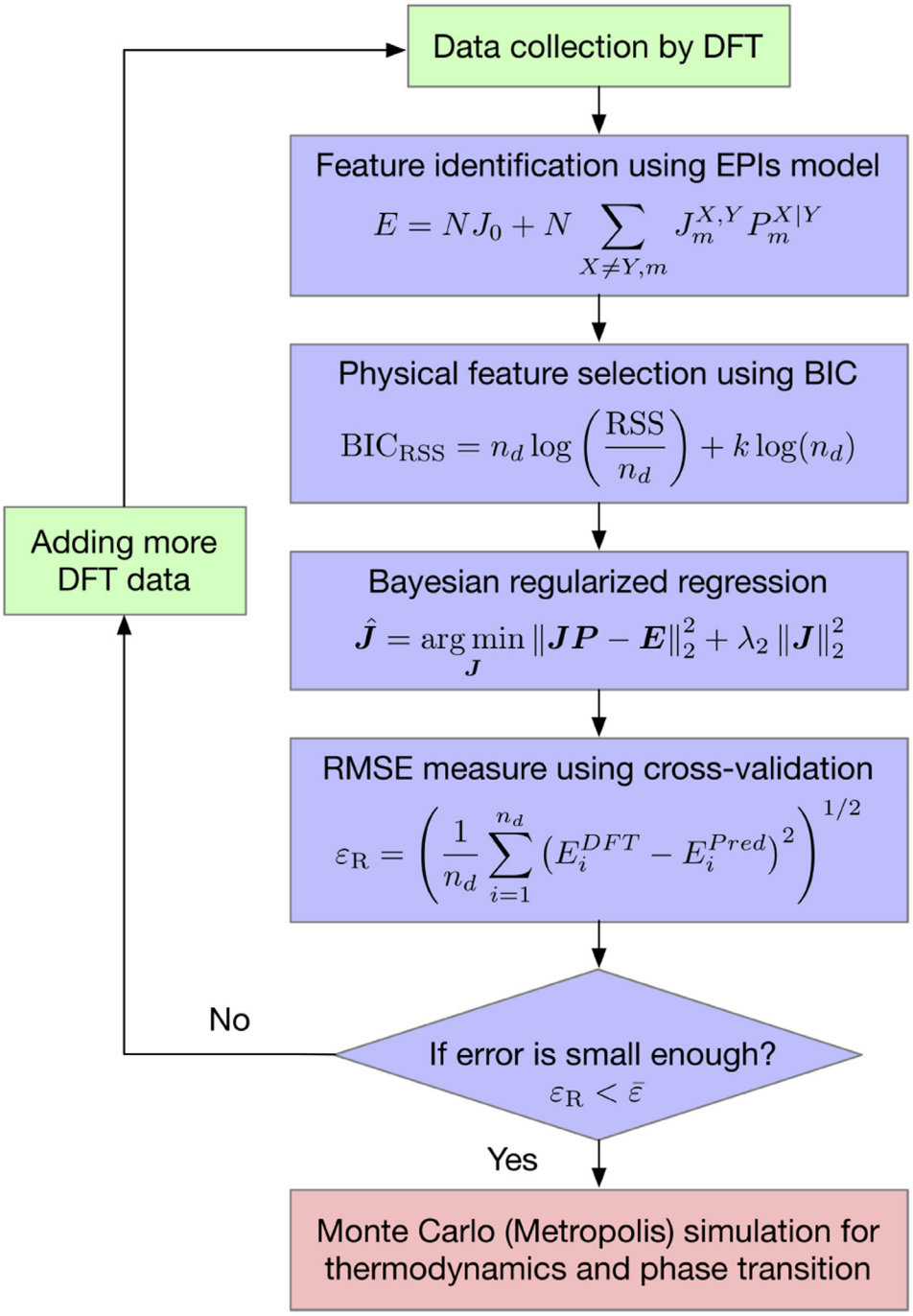



**Fig. 2.** Flowchart for a robust data-driven algorithm using Bayesian framework.

Due to the periodic boundary condition, the configurations drawn from smaller supercells often include different long-range order, while the samples obtained with larger supercells contain various degrees of short-range order. As a result, it is highly possible that a random system with large supercell is not well represented by only using samples generated from small supercells, which are commonly used because of their efficiency, but may result in loss of physical information in the thermodynamic limit. To conduct a robust data-driven approach, we adopt an ensemble random sampling strategy that combines the data from different supercells. This simple, yet efficient, technique aims to obtain a training dataset with different degrees of order and disorder such that the data are more representative.

The benefit of the ensemble sampling strategy can be seen in Fig. 5. All 200 configurations from relatively large supercell (128 for NbMoTaW and 160 for NbMoTaWV and NbMoTaWTi) are selected as the testing dataset. Total 150 configurations randomly drawn from each smaller supercell (16, 32, 64 for NbMoTaW and 20, 40, 80 for NbMoTaWV and NbMoTaWTi) are selected as the training dataset. The ensemble sampling strategy considers an ensemble of datasets, which is achieved by randomly generating 50 configurations using Latin hypercube sampling [65] from each of the three supercells. In terms of these four cases, 100 random trials are carried out to estimate the standard deviation (error bar in Fig. 5) of the testing RMSE results. It can be easily seen that the mean and standard deviation of RMSE results are significantly larger when

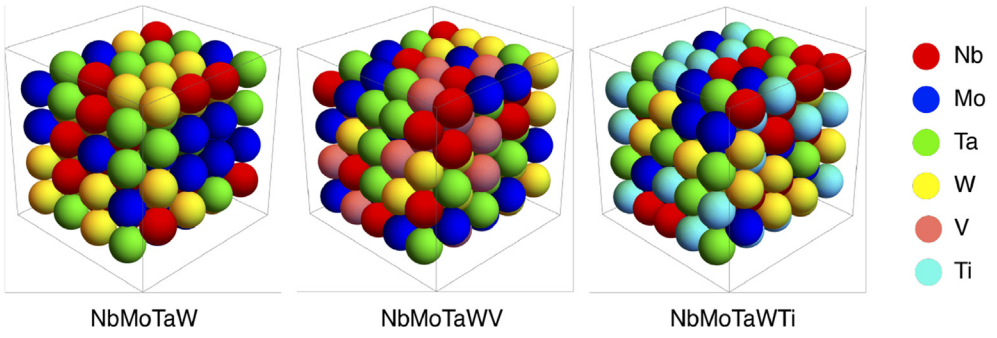



**Fig. 3.** Bcc supercells of refractory HEAs. (a) NbMoTaW with 128 atoms, (b) NbMoTaWV with 160 atoms and (c) NbMoTaWTi with 160 atoms.

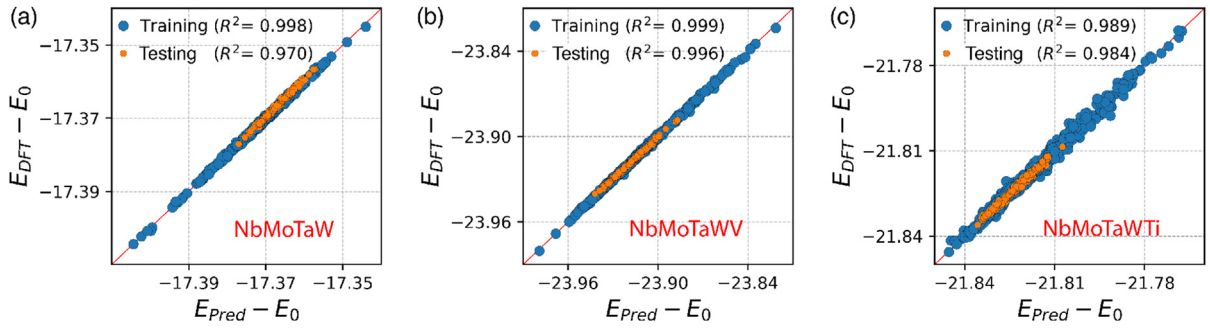



**Fig. 4.** Comparison of DFT calculated energy with predicted energy using Bayesian regularized regression for (a) NbMoTaW, (b) NbMoTaWV and (c) NbMoTaWTi.

**Table 1**
Training and testing RMSE accuracy of configurational energy for three HEAs.

| HEA | NbMoTaW | NbMoTaWV | NbMoTaWTi |
|---|---|---|---|
| Training $\varepsilon_R$ (meV) | 0.335 | 0.710 | 1.400 |
| Testing $\varepsilon_R$ (meV) | 0.632 | 0.647 | 0.665 |

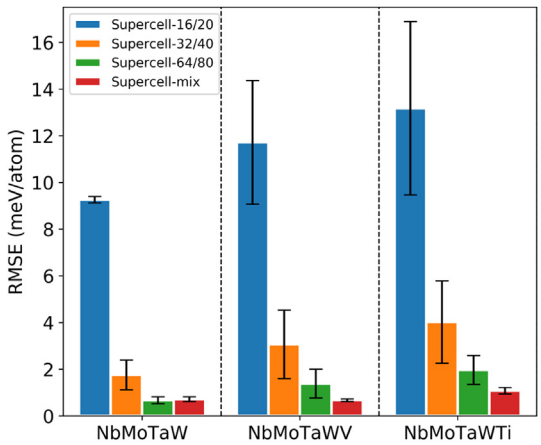



**Fig. 5.** Testing performance comparison between ensemble sampling strategy and sampling drawn from only single supercell. Blue, orange and green bars: testing results using 150 configurations only from one specific supercell (16, 32, 64 for NbMoTaW and 20, 40, 80 for NbMoTaWV and NbMoTaWTi respectively). Red bar: testing results using an ensemble of 150 configurations that consists of three 50 data drawn from each supercell.

only the smallest supercell is used. The results underscore the fact that the configuration space in a random system is not well covered by training data only drawn from small supercells. The performance in supercell of system 32 (40) and 64 (80) are better than that those for 16 (20)-atom supercells, showing that more degrees of short-range and long-range order are captured by these training data. The ensemble sampling strategy (red bars) demonstrates a minimal RMSE mean ( $<1$ meV) and standard deviation which outperforms the other three cases in terms of the accuracy and stability. This robust strategy plays a substantial role in the data-driven modeling of the configurational energy such that the subsequent Monte Carlo simulation based on this efficient Hamiltonian can safely explore the whole region of configuration space.

### 3.2. Uncertainty in effective pair interaction bonds

The coordination shells, as the physical features, have a pivotal influence on the EPI model and their impact can be analyzed from the EPI parameters, as shown in Fig. 6 - Fig. 8. For all the three refractory HEAs, the first two shells, involving the nearest and next-nearest neighbor interactions are dominant, while the 3rd to 6th shells present a less essential role. A comparison between NbMoTaW and the other two HEAs, NbMoTaWV and NbMoTaWTi shows that the EPIs of NbMoTaW, as shown in Fig. 6 (c), are relatively stable and short-ranged due to small magnitude associated with the long-ranged shells, while the NbMoTaWV and NbMoTaWTi, as shown in Fig. 7 (c) and Fig. 8 (c) are more frustrated and long-ranged, with significant contributions from up to the 6th shell.

Moreover, three different dataset sizes, $n_d = 100, 400$ and 800 are shown here to investigate the effect of data on the EPI parameters (chemical bonds). The ensemble sampling strategy is used herein such that we collect 25, 100 and 200 configurations from each of the four supercell sizes, i.e. 16, 32, 64 and 128 for NbMoTaW and 20, 40, 80 and 160 for NbMoTaWV and NbMoTaWTi. Given a small dataset size, for example, $n_d = 100$, the trend of EPI parameters is consistent - the first two shells are dominant, but the values still have a discrepancy from larger dataset size, for example, $n_d = 800$. In other words, the uncertainties associated with the EPI parameters are primarily

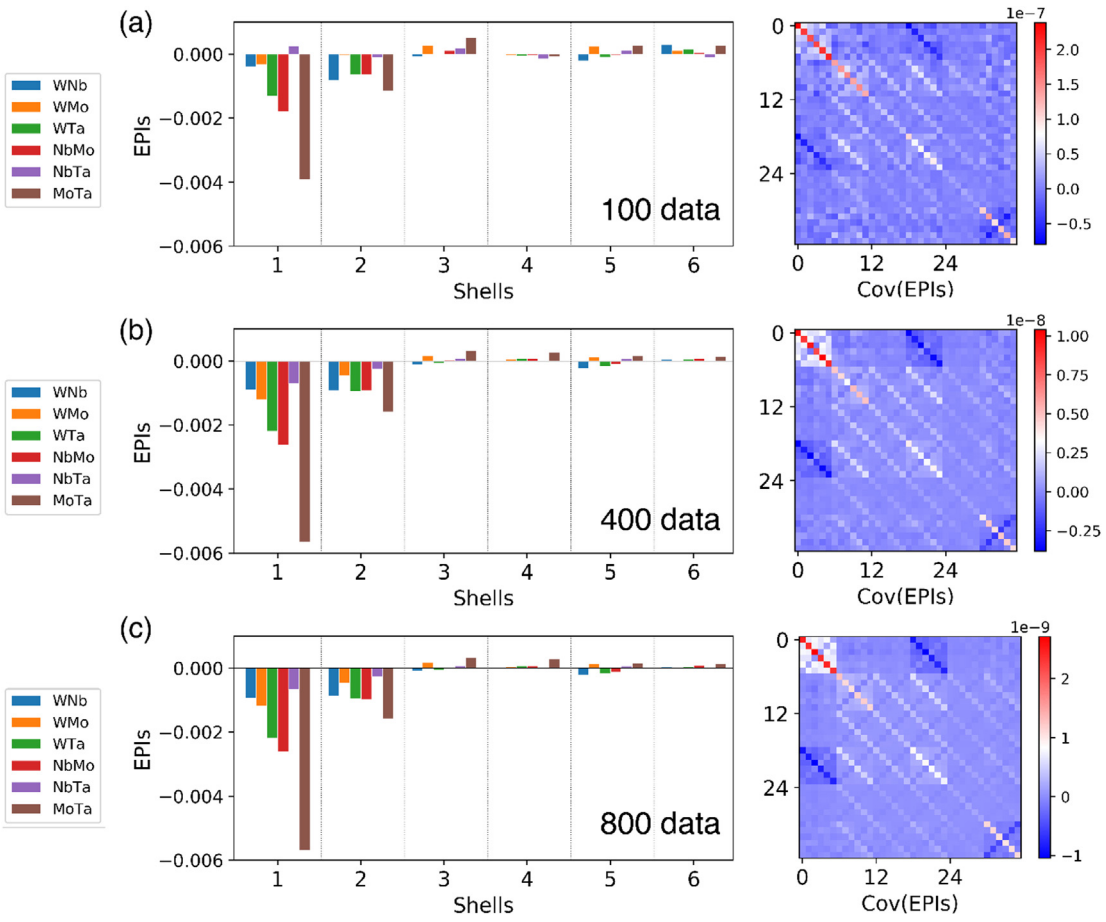



**Fig. 6.** Effective pair interaction (EPI) parameters and their uncertainties that are quantified by the variance-covariance matrix given different sizes of training dataset for NbMoTaW.

caused by a lack of data. This is reflected in the variance-covariance matrix of EPI parameters, as shown in Figs. 6–8 (a). It is easily seen that the variance in 1st shells is the largest, followed by the 2nd shell and the 6th shell, which are larger than the other shells. The covariance values trending to zero demonstrates that there is no strong correlation among each of the shells, which is in agreement with the independence assumption.

As the dataset size increases, from $n_d = 100$ to $n_d = 400$, the EPI parameters become more stable and almost the same as the case of $n_d = 800$. The corresponding variance is also significantly reduced by more than one order of magnitude, as shown in Figs. 6–8 (b). Furthermore, the variance-covariance matrix presents an increasingly clear "pattern" with the increasing of dataset size and we therefore can observe: a) the pattern of the variance-covariance matrix is divided by the identified physical features (coordination shells); b) the EPI parameters in the nearest-neighbor shell shows a stronger positive correlation, while the 6th shell illustrates a relatively large uncertainty with a negative correlation among the EPI parameters; c) the EPI parameters between the 1st shell and 4th shell have a negative correlation. Even though most of the EPI parameters are physically independent as we expected, the EPI parameters in a specific shell are not completely independent, particularly for the nearest-neighbor shell, and there is still a slight positive or negative correlation between different shells. By inspecting the positions of the neighboring atoms, it is easy to see that one reason for the correlation is that some atoms (such as the first and the forth neighbors) are related by supercell lattice vectors, especially in the case of small supercells.

### 3.3. Effect of physical feature selection

Due to the lack of data, it is desirable to quantify the uncertainties of the EPI parameters via a Bayesian method as discussed above and how these uncertainties will be propagated to the prediction of configurational energies and finally affect thermodynamic quantities. It therefore gives rise to a critical issue for understanding how to systemically reduce the uncertainty and improve the robustness and reliability of model predictions when only limited data can be collected. To deal with this challenge, we introduce the Bayesian information criterion (BIC) to conduct physical feature selection and investigate the effect of the coordination shell number $m$ on the prediction accuracy. Fig. 9 - Fig. 11 show the BIC values for the different numbers of shells given various sizes of the training dataset. Using an ensemble sampling strategy, we randomly collect data (from 20 configurations to 200 configurations) from each of four supercells by 100 random trials and then estimate the mean and standard deviation of BIC values for each specific number of shells. For NbTaMoW in Fig. 9 (a), $m = 2$, corresponding to the smallest BIC value, represents the "best" number of shells given only 20 configurations, while a larger number of $m$ leads to overfitting. As more data are collected into the model, the best number of shells gradually increases and converges towards $m = 9$. When relatively large datasets ($n_d = 150$) are available, the mean of BIC values with $m > 9$ are quite close and the variation of BIC in each $m$ is much smaller than the case of small dataset size. NbTaMoWV as shown in Fig. 10 has a similar overall trend for NbMoTaW. Also, NbTaMoWTi tends to a smaller number of shells given limited data but it displays a smaller converged

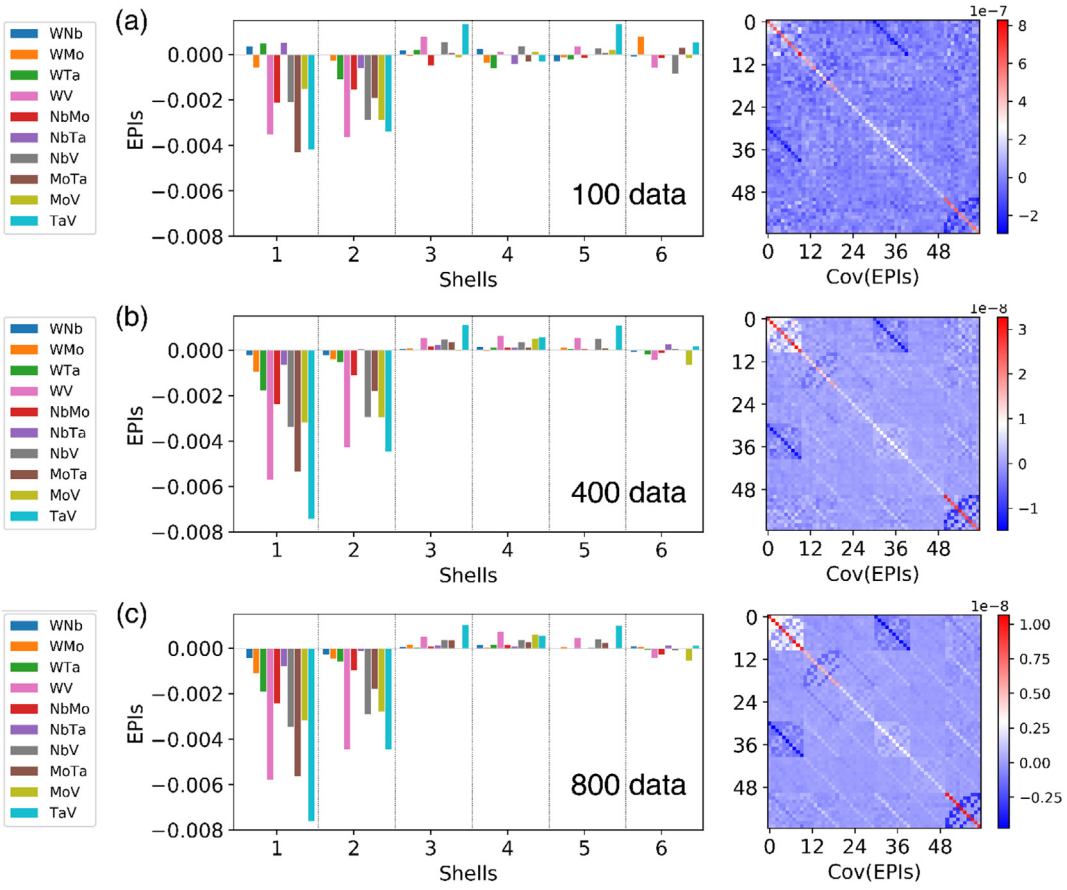



**Fig. 7.** Effective pair interaction (EPI) parameters and their uncertainties that are quantified by the variance-covariance matrix given different sizes of training dataset for NbMoTaWV.

number of shells $m = 6$, which differs from $m = 9$ for NbTaMoW and NbTaMoWV.

Table 2 provides a complete list of best number of shells given a specific dataset size (from 20 to 200 configurations for each supercell) for these three refractory HEAs. If we divide the data into three categories, such as small size (total number of configurations $n_t = 4 \times n_d n_t < 100$), medium size ($100 < n_t < 400$) and large size ($n_t > 400$), we can conclude the following:

- Small size: select a small number of shells $m = 2$ or $m = 3$ to avoid overfitting
- Medium size: identify $m = 5$ or $m = 6$ in a reliable choice to avoid underfitting
- Large size: determine a reasonable large number of feature $6 \leq m \leq 9$ due to the bias-variance tradeoff [66].

Next, we examine the effect of feature selection on the testing RMSE. As shown in Fig. 12, we compare the RMSE results with three numbers of shells that include $m = 13$, $m = 7$ and $m = m_{\text{best}}$ where $m_{\text{best}}$ is referred to as the best number of shells. It is easy to observe that the RMSE for $m = 13$ ($\varepsilon_R = 4.5$ meV) and $m = 7$ ($\varepsilon_R = 3.4$ meV) are significantly larger than the results using the best number of shells that yields $\varepsilon_R = 1.5$ meV given $n_d = 20$ training configurations. The RMSE mean value for $m = 13$ and $m = 7$ is reduced as dataset size increases (see Fig. 12 (a)) but the variations (standard deviation) are still substantially large. When a relatively large dataset is collected, as shown in Fig. 12 (b), the RMSE of $m = 13$ and $m = m_{\text{best}}$ gradually converges to a small value, $\varepsilon_R = 0.21$ meV, while $m = 7$ converges to $\varepsilon_R = 0.35$ meV and can not be further decreased. In other words, the shorter cutoff (for example $m = 7$) for the physical feature does not fully capture all the physical information and underestimates the system complexity of HEAs.

Fig. 13 and Fig. 14 show the RMSE results for NbMoTaWV and NbMoTaWTi respectively. Given a small size of data ($n_d < 25$), the RMSE of $m = 13$ and $m = 7$ are nearly more than $\varepsilon_R = 10$ meV, which may lead to a large bias for Monte Carlo simulation of thermodynamics. Nevertheless, the best shells with the same data show a relatively small ($\varepsilon_R < 5$ meV) and reliable RMSE estimate. For the medium dataset size, the RMSE of $m = 13$ and $m = 7$ are reduced but still larger than that of the best shells, which can be observed from Figs. 13 (a) and Fig. 14(a). When relatively large data is considered, $m = 13$ and $m = 7$ in the model for NbMoTaWTi eventually converge towards the minimal value at the best shells, while $m = 7$ in NbMoTaWV still shows slight underfitting issue because the best number identified by BIC is $m = 9$, greater than the arbitrary choice of $m = 7$. Through careful analysis and comparison of these three HEAs, we established that the model with the best number of shells identified by feature selection demonstrates a highly accurate and robust performance on either small or relatively large data. Rather than an arbitrary selection of physical feature, the reliable feature selection can effectively reduce the risk of underfitting and overfitting during model predictions.

Finally, we would like to comment on the experimental implications of our results. One possible result due to the elemental pair interactions is to form short range order. However, despite the evidence [67,68] provided by X-ray absorption fine structure (EXAFS), identifying SRO in experiments is still very challenging. Another possible consequence is the formation of second phases. For example, a recent study [69] of CrVTaW revealed Cr- and V-rich precipitates using atom probe tomography (APT) and transmission

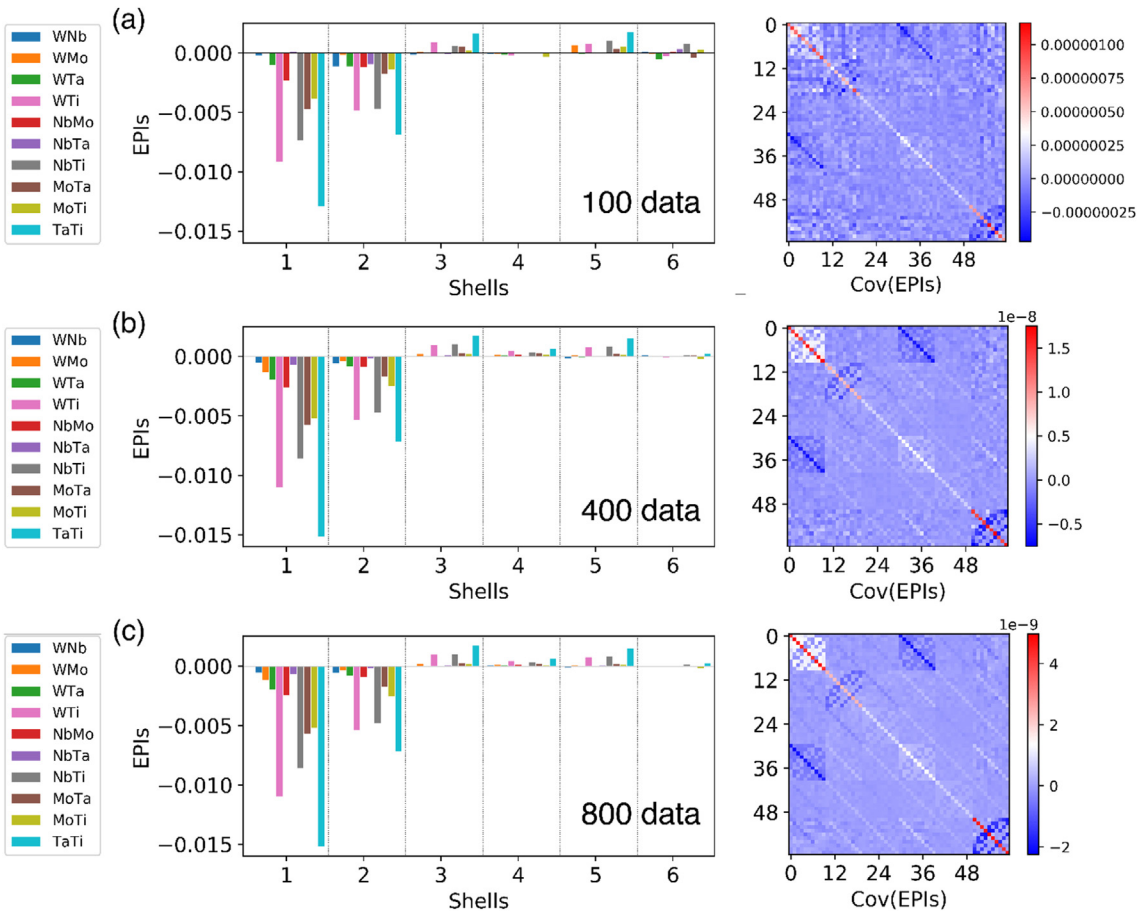



**Fig. 8.** Effective pair interaction (EPI) parameters and their uncertainties that are quantified by the variance-covariance matrix given different sizes of training dataset for NbMoTaTi.

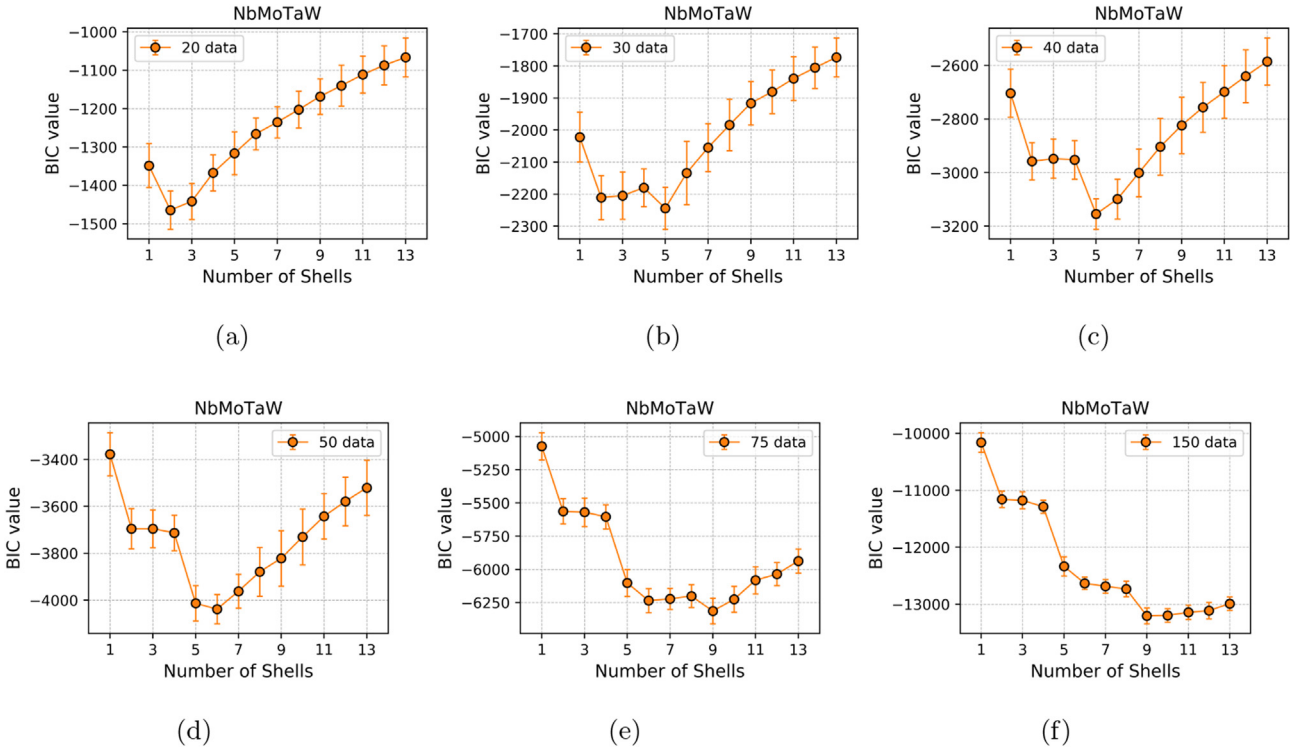



**Fig. 9.** Bayesian information criterion results for different numbers of coordination shells in NbMoTaW given a specific training dataset size, including (a) 20 configurations, (b) 30 configurations, (c) 40 configurations, (d) 50 configurations, (e) 75 configurations and (f) 150 configurations.

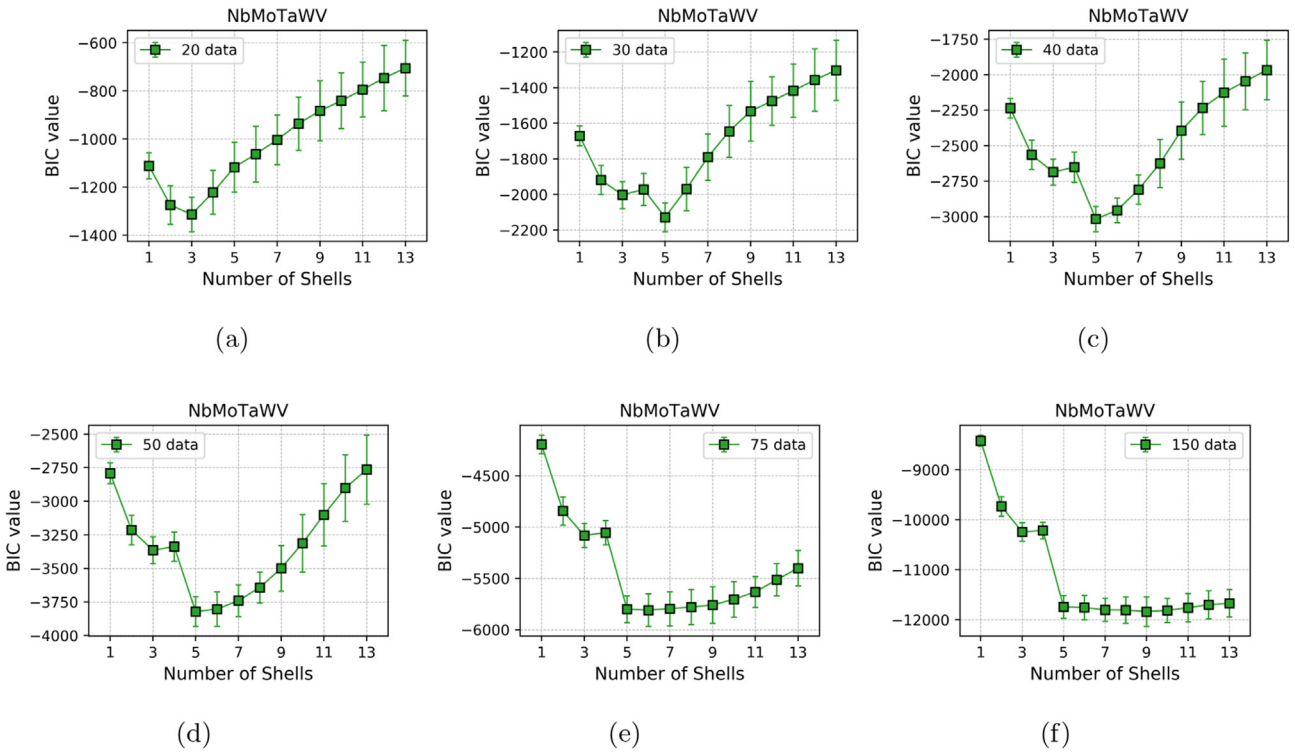



**Fig. 10.** Bayesian information criterion results for different numbers of coordination shells in NbMoTaWV given a specific training dataset size, including (a) 20 configurations, (b) 30 configurations, (c) 40 configurations, (d) 50 configurations, (e) 75 configurations and (f) 150 configurations.

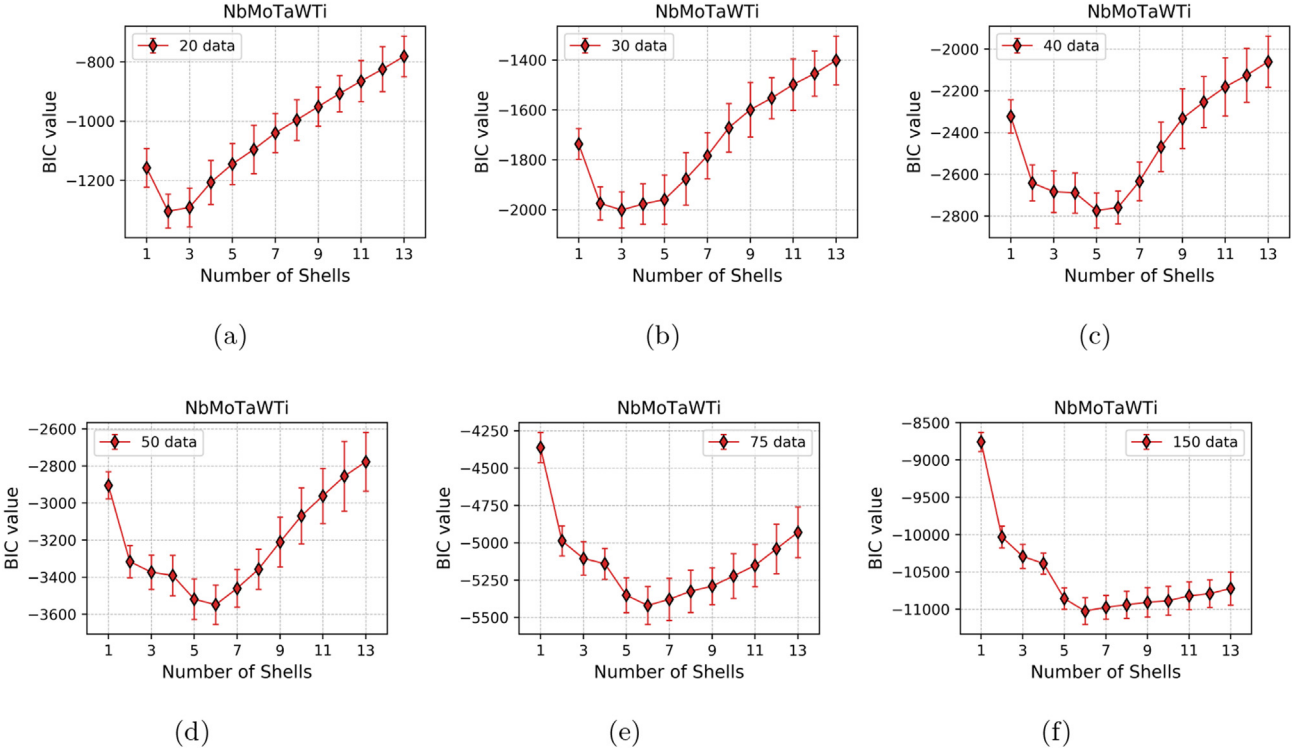



**Fig. 11.** Bayesian information criterion results for different numbers of coordination shells in NbMoTaWTi given a specific training dataset size, including (a) 20 configurations, (b) 30 configurations, (c) 40 configurations, (d) 50 configurations, (e) 75 configurations and (f) 150 configurations.

electron microscopy (TEM), which agrees with our finding that adding V leads to stronger pair interactions. On the other hand, there is also a thermodynamic simulation of refractory HEAs [70], where the preference of forming TaMo bond and separation of V are identified.

**Table 2**
The best number of coordination shells given a specific size of training dataset.

| Each supercell | NbMoTaW | NbMoTaWV | NbMoTaWTi |
|---|---|---|---|
| 20 | 2 | 3 | 2 |
| 25 | 2 | 3 | 3 |
| 30 | 5 | 5 | 3 |
| 35 | 5 | 5 | 5 |
| 40 | 5 | 5 | 5 |
| 45 | 5 | 5 | 6 |
| 50 | 6 | 5 | 6 |
| 55 | 6 | 5 | 6 |
| 60 | 6 | 5 | 6 |
| 65 | 6 | 5 | 6 |
| 70 | 9 | 5 | 6 |
| 75 | 9 | 6 | 6 |
| 100 | 9 | 9 | 6 |
| 125 | 9 | 9 | 6 |
| 150 | 9 | 9 | 6 |
| 175 | 9 | 9 | 6 |
| 200 | 9 | 9 | 6 |

components inherent in multicomponent systems. The Bayesian regression algorithm is employed to establish an efficient Hamiltonian through a set of random configurations with the corresponding energies calculated using the DFT method and to quantify the uncertainties and correlations of effective pair interaction parameters. To improve the accuracy and reliability of prediction, we further perform Bayesian feature selection for dealing with the truncation of the model, specifically given the lack of data. For all three HEAs, NbTaMoW, NbTaMoWV and NbTaMoWTi, we have demonstrated highly accurate and robust performance in predicting the configurational energy. The proposed method is therefore a powerful tool for studying the thermodynamics and order-disorder phase transitions through the subsequent Monte Carlo simulations.

Specifically, we find that a small and single supercell is unable to explore the various order and disorder in multicomponent systems sufficiently. We therefore apply an ensemble sampling strategy which naturally incorporates chemical configurations of different short-range and long-range order. This method performs a sam-

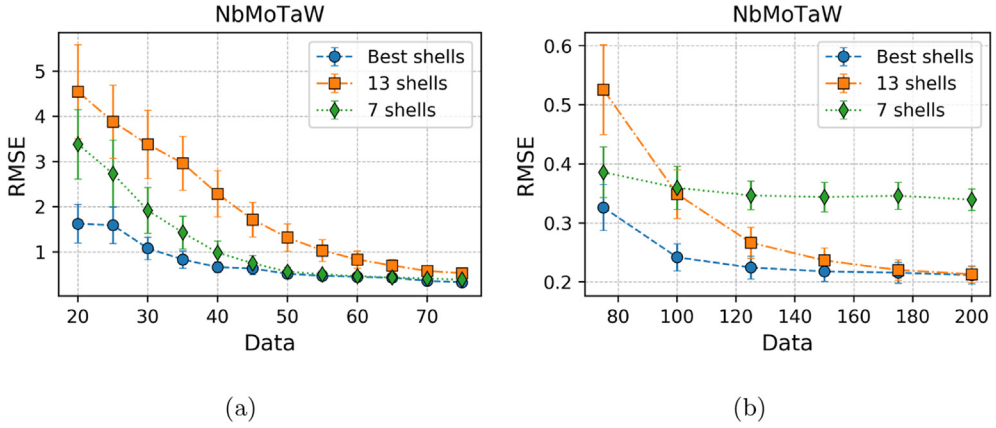



**Fig. 12.** RMSE results of NbMoTaW with three selections of coordination shell number given different sizes of training dataset (a) $n_d \leq 75$ and (b) $n_d \geq 75$. (note: (b) has different y-axis scale from (a)).

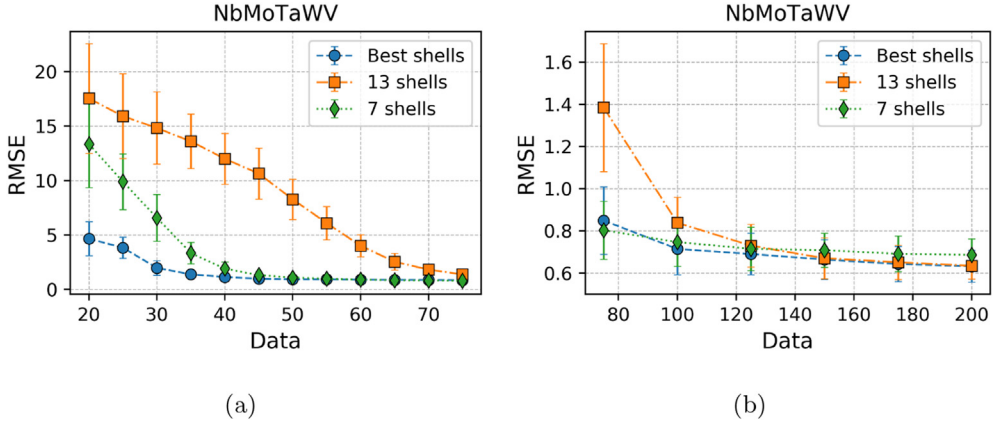



**Fig. 13.** RMSE results of NbMoTaWV with three selections of coordination shell number given different sizes of training dataset(a) $n_d \leq 75$ and (b) $n_d \geq 75$ (note: (b) has different y-axis scale from (a)).

## 4. Conclusions

In this work, we develop a systematical and robust Bayesian framework to discover efficient and accurate modeling of configurational energies from a data-driven perspective. A short-range pair interaction model is used here to characterize the physical feature and it is well-suited to deal with a large number of pling capability well-suited to a huge configuration space and it has demonstrated to be a simple, yet robust, scheme to enhance the representativity of data. Using this strategy, the resulting RMSE of three HEAs show a very small mean error ($\varepsilon < 1$ meV) with a tiny standard deviation that is significantly lower than the other cases that use only single supercells.

Also, we note that the chemical bonds in the EPI model show a

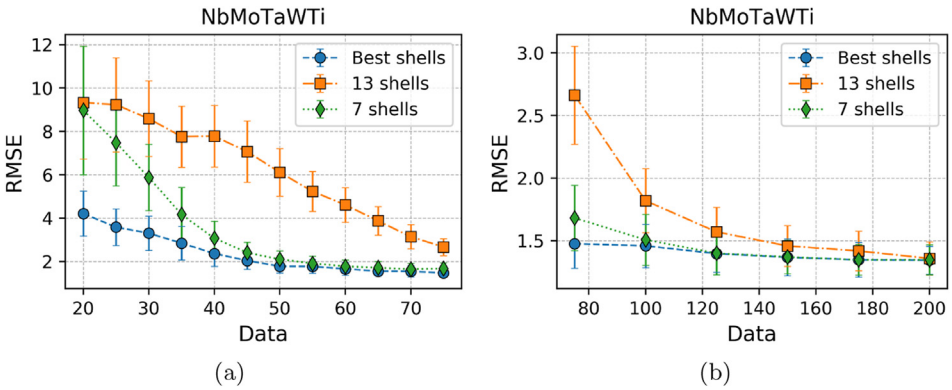



**Fig. 14.** RMSE results of NbMoTaWTi with three selections of coordination shell number given different sizes of training dataset(a) $n_d \leq 75$ and (b) $n_d \geq 75$ (note: (b) has different y-axis scale from (a)).

frustrated behavior if the dataset size is too small ($n_t = 100$). This can also be observed from the uncertainty quantification of EPI bonds, for instance, the nearest-neighbor shell displays a high level of variation and frustration. However, as more data are obtained, for example, from $n_t = 100$ to $n_t = 400$, the bonds tend to become stable and the uncertainties are reduced. This is due to the finite supercell sizes, in which some neighboring atoms are related to the supercell lattice vectors. We also notice a certain degree of correlation within the nearest-neighbor shell even though they are essentially assumed to be independent.

Finally, the impact of feature selection and the effect of dataset size are carefully discussed. For each material, the best truncated number of coordination shells is slightly different given a specific dataset but they all demonstrate a similar trend. We therefore provide a general suggestion according to the dataset size: a) $m = 2 \sim 3$ for small size ($n_t < 100$), $m = 5 \sim 6$ for medium size ($100 \leq n_t \leq 400$) and $m = 6 \sim 9$ for relatively large size ($n_t > 400$). Using this feature selection scheme, we have demonstrated an accurate and robust performance for all three HEAs while avoiding underfitting or overfitting that is often occurrence in machine learning modeling.

## Data availability statement

The data of this study will be made available on request.

## CRediT authorship contribution statement

**Jiaxin Zhang:** Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing - original draft. **Xianglin Liu:** Conceptualization, Data curation, Validation, Writing - review & editing. **Sirui Bi:** Validation, Investigation, Visualization, Writing - review & editing. **Junqi Yin:** Resources, Writing - review & editing. **Guannan Zhang:** Supervision, Writing - review & editing. **Markus Eisenbach:** Writing - review & editing, Supervision, Project administration, Funding acquisition.

## Acknowledgements

This work of J. Z. was supported by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory. X. L. and M. E. were supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Science and Engineering Division. This research used resources of the Oak Ridge Leadership Computing Facility, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

## References

[1] J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau, S.-Y. Chang, Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes, Adv. Eng. Mater. 6 (2004) 299–303.
[2] O.N. Senkov, G. Wilks, J. Scott, D.B. Miracle, Mechanical properties of nb25mo25ta25w25 and v20nb20mo20ta20w20 refractory high entropy alloys, Intermetallics 19 (2011) 698–706.
[3] Y. Zhang, T.T. Zuo, Z. Tang, M.C. Gao, K.A. Dahmen, P.K. Liaw, Z.P. Lu, Microstructures and properties of high-entropy alloys, Prog. Mater. Sci. 61 (2014) 1–93.
[4] Z. Li, K.G. Pradeep, Y. Deng, D. Raabe, C.C. Tasan, Metastable high-entropy dual-phase alloys overcome the strength–ductility trade-off, Nature 534 (2016) 227.
[5] M.A. Tunes, V.M. Vishnyakov, Microstructural origins of the high mechanical damage tolerance of nbtamow refractory high-entropy alloy thin films, Mater. Des. 170 (2019) 107692.
[6] M.-H. Tsai, J.-W. Yeh, High-entropy alloys: a critical review, Mater. Res. Lett. 2 (2014) 107–123.
[7] D.B. Miracle, O.N. Senkov, A critical review of high entropy alloys and related concepts, Acta Mater. 122 (2017) 448–511.
[8] M.C. Gao, J.-W. Yeh, P.K. Liaw, Y. Zhang, High-entropy Alloys: Fundamentals and Applications, Springer, 2016.
[9] A. Raza, H.J. Ryu, S.H. Hong, Strength enhancement and density reduction by the addition of al in crfemov based high-entropy alloy fabricated through powder metallurgy, Mater. Des. 157 (2018) 97–104.
[10] M. Widom, W.P. Huhn, S. Maiti, W. Steurer, Hybrid Monte Carlo/molecular dynamics simulation of a refractory metal high entropy alloy, Metall. Mater. Trans. A 45 (2014) 196–200.
[11] B.S. Murty, J.-W. Yeh, S. Ranganathan, P. Bhattacharjee, High-entropy Alloys, Elsevier, 2019.
[12] M. Eisenbach, Z. Pei, X. Liu, First-principles study of order-disorder transitions in multicomponent solid-solution alloys, J. Phys. Condens. Matter 31 (2019) 273002.
[13] M. Widom, Modeling the structure and thermodynamics of high-entropy alloys, J. Mater. Res. 33 (2018) 2881–2898.
[14] M.C. Gao, P. Gao, J.A. Hawk, L. Ouyang, D.E. Alman, M. Widom, Computational modeling of high-entropy alloys: structures, thermodynamics and elasticity, J. Mater. Res. 32 (2017) 3627–3641.
[15] Y. Ikeda, B. Grabowski, F. Körmann, Ab initio phase stabilities and mechanical properties of multicomponent alloys: a comprehensive review for high entropy alloys and compositionally complex alloys, Mater. Char. 147 (2019) 464–511.
[16] M.B. Kivy, M.A. Zaeem, S. Lekakh, Investigating phase formations in cast alfeconicu high entropy alloys by combination of computational modeling and experiments, Mater. Des. 127 (2017) 224–232.
[17] Y. Ye, Q. Wang, J. Lu, C. Liu, Y. Yang, High-entropy alloy: challenges and prospects, Mater. Today 19 (2016) 349–362.
[18] D. Ma, B. Grabowski, F. Körmann, J. Neugebauer, D. Raabe, Ab initio thermodynamics of the cocrfemnni high entropy alloy: importance of entropy contributions beyond the configurational one, Acta Mater. 100 (2015) 90–97.
[19] W. Huang, A. Urban, Z. Rong, Z. Ding, C. Luo, G. Ceder, Construction of ground-state preserving sparse lattice models for predictive materials simulations, NPJ Comput. Mater. 3 (2017) 30.
[20] S.N. Khan, M. Eisenbach, Density-functional monte-carlo simulation of cuzn order-disorder transition, Phys. Rev. B 93 (2016), 024203.
[21] R. Kikuchi, A theory of cooperative phenomena, Phys. Rev. 81 (1951) 988.
[22] J.M. Sanchez, F. Ducastelle, D. Gratias, Generalized cluster description of multicomponent systems, Phys. A Stat. Mech. Appl. 128 (1984) 334–350.
[23] A. van de Walle, G. Ceder, Automating first-principles phase diagram calculations, J. Phase Equilibria 23 (2002) 348.

[24] L.J. Nelson, G.L. Hart, F. Zhou, V. Ozoliņš, et al., Compressive sensing as a paradigm for building physics models, Phys. Rev. B 87 (2013), 035125.
[25] T. Mueller, G. Ceder, Bayesian approach to cluster expansions, Phys. Rev. B 80 (2009), 024103.
[26] V. Blum, G.L. Hart, M.J. Walorski, A. Zunger, Using genetic algorithms to map first-principles results to model Hamiltonians: application to the generalized ising model for alloys, Phys. Rev. B 72 (2005) 165113.
[27] G.L. Hart, V. Blum, M.J. Walorski, A. Zunger, Evolutionary approach for determining first-principles Hamiltonians, Nat. Mater. 4 (2005) 391.
[28] A. Seko, Y. Koyama, I. Tanaka, Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations, Phys. Rev. B 80 (2009) 165122.
[29] A.R. Natarajan, A. Van der Ven, Machine-learning the configurational energy of multicomponent crystalline solids, NPJ Comput. Mater. 4 (2018) 56.
[30] J.H. Chang, D. Kleiven, M. Melander, J. Akola, J.M. Garcia-Lastra, T. Vegge, Clease: a versatile and user-friendly implementation of cluster expansion method, J. Phys. Condens. Matter 31 (2019) 325901.
[31] M. Ångqvist, W.A. Muñoz, J.M. Rahm, E. Fransson, C. Durniak, P. Rozyczko, T.H. Rod, P. Erhart, Icet—a python Library for Constructing and Sampling Alloy Cluster Expansions, Advanced Theory and Simulations, 2019, p. 1900015.
[32] C. Jiang, B.P. Uberuaga, Efficient ab initio modeling of random multicomponent alloys, Phys. Rev. Lett. 116 (2016) 105501.
[33] M.I. Jordan, T.M. Mitchell, Machine learning: trends, perspectives, and prospects, Science 349 (2015) 255–260.
[34] Y. LeCun, Y. Bengio, G. Hinton, Deep Learn. Nat. 521 (2015) 436.
[35] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (2018) 547.
[36] T. Mueller, A.G. Kusne, R. Ramprasad, Machine learning in materials science: recent progress and emerging applications, Rev. Comput. Chem. 29 (2016) 186–273.
[37] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: generative models for matter engineering, Science 361 (2018) 360–365.
[38] A. Solomou, G. Zhao, S. Boluki, J.K. Joy, X. Qian, I. Karaman, R. Arróyave, D.C. Lagoudas, Multi-objective bayesian materials discovery: application on the discovery of precipitation strengthened niti shape memory alloys through micromechanical modeling, Mater. Des. 160 (2018) 810–827.
[39] C. Kim, G. Pilania, R. Ramprasad, From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown, Chem. Mater. 28 (2016) 1304–1311.
[40] P. Raccuglia, K.C. Elbert, P.D. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, A.J. Norquist, Machine-learning-assisted materials discovery using failed experiments, Nature 533 (2016) 73.
[41] S. Feng, H. Zhou, H. Dong, Using deep neural network with small dataset to predict material defects, Mater. Des. 162 (2019) 300–310.
[42] J. Carrasquilla, R.G. Melko, Machine learning phases of matter, Nat. Phys. 13 (2017) 431.
[43] W. Huang, P. Martin, H.L. Zhuang, Machine-learning phase prediction of high-entropy alloys, Acta Mater. 169 (2019) 225–236.
[44] T. Kostiuchenko, F. Körmann, J. Neugebauer, A. Shapeev, Impact of lattice relaxations on phase transitions in a high-entropy alloy studied by machine-learning potentials, NPJ Comput. Mater. 5 (2019) 55.
[45] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, Accelerating materials property predictions using machine learning, Sci. Rep. 3 (2013) 2810.
[46] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, NPJ Comput. Mater. 2 (2016) 16028.
[47] X. Chen, H. Zhou, Y. Li, Effective design space exploration of gradient nanostructured materials using active learning based surrogate models, Mater. Des. 183 (2019) 108085.
[48] N.H. Paulson, M.W. Priddy, D.L. McDowell, S.R. Kalidindi, Data-driven reduced-order models for rank-ordering the high cycle fatigue performance of polycrystalline microstructures, Mater. Des. 154 (2018) 170–183.
[49] D. Dragoni, T.D. Daff, G. Csányi, N. Marzari, Achieving dft accuracy with a machine-learning interatomic potential: thermomechanics and defects in bcc ferromagnetic iron, Phys. Rev. Mater. 2 (2018), 013808.
[50] S. Chmiela, A. Tkatchenko, H.E. Sauceda, I. Poltavsky, K.T. Schütt, K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields, Sci. Adv. 3 (2017), e1603015.
[51] V.L. Deringer, G. Csányi, Machine learning based interatomic potential for amorphous carbon, Phys. Rev. B 95 (2017), 094203.
[52] Z. Li, J.R. Kermode, A. De Vita, Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces, Phys. Rev. Lett. 114 (2015), 096405.
[53] X. Liu, J. Zhang, M. Eisenbach, Y. Wang, Machine Learning Modeling of High Entropy Alloy: the Role of Short-Range Order, 2019 arXiv preprint arXiv: 1906.02889.
[54] J. Zhang, M.D. Shields, On the quantification and efficient propagation of imprecise probabilities resulting from small datasets, Mech. Syst. Signal Process. 98 (2018) 465–483.
[55] J. Zhang, M.D. Shields, On the quantification and efficient propagation of imprecise probabilities with copula dependence, Int. J. Approx. Reason. (2019) (in review).
[56] M. Aldegunde, N. Zabaras, J. Kristensen, Quantifying uncertainties in first-principles alloy thermodynamics using cluster expansions, J. Comput. Phys. 323 (2016) 17–44.
[57] J. Zhang, S. TerMaath, M.D. Shields, Imprecise Global Sensitivity Analysis Using Bayesian Multimodel Inference and Importance Sampling, Reliability Engineering & System Safety, 2019a (in review).
[58] J. Zhang, M. Shields, S. TerMaath, Probabilistic Modeling and Prediction of Out-Of-Plane Unidirectional Composite Lamina Properties, Mechanics of Advanced Materials and Structures, 2019 (in review).
[59] J. Kristensen, N.J. Zabaras, Bayesian uncertainty quantification in the evaluation of alloy properties with the cluster expansion method, Comput. Phys. Commun. 185 (2014) 2885–2892.
[60] P. Fernandez-Zelaia, V.R. Joseph, S.R. Kalidindi, S.N. Melkote, Estimating mechanical properties from spherical indentation using bayesian approaches, Mater. Des. 147 (2018) 92–105.
[61] X. Liu, J. Zhang, S. Bi, Y. Wang, G.M. Stocks, M. Eisenbach, Chemical Complexity in High Entropy Alloys: A Pair-Interaction Perspective, 2019, p. 10223, arXiv preprint arXiv:1907.
[62] J. Zhang, M.D. Shields, Efficient Monte Carlo resampling for probability measure changes from bayesian updating, Probabilistic Eng. Mech. 55 (2019) 54–66.
[63] J. Zhang, M.D. Shields, The effect of prior probabilities on quantification and propagation of imprecise probabilities resulting from small datasets, Comput. Methods Appl. Mech. Eng. 334 (2018) 483–506.
[64] Y. Wang, G. Stocks, W. Shelton, D. Nicholson, Z. Szotek, W. Temmerman, Order-n multiple scattering approach to electronic structure calculations, Phys. Rev. Lett. 75 (1995) 2867.
[65] M.D. Shields, J. Zhang, The generalization of Latin hypercube sampling, Reliab. Eng. Syst. Saf. 148 (2016) 96–108.
[66] P. Mehta, M. Bukov, C.-H. Wang, A.G. Day, C. Richardson, C.K. Fisher, D.J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Phys. Rep. (2019).
[67] F. Zhang, S. Zhao, K. Jin, H. Xue, G. Velisa, H. Bei, R. Huang, J. Ko, D. Pagan, J. Neuefeind, et al., Local structure and short-range order in a nicocr solid solution alloy, Phys. Rev. Lett. 118 (2017) 205501.
[68] Y. Tong, K. Jin, H. Bei, J. Ko, D.C. Pagan, Y. Zhang, F. Zhang, Local lattice distortion in nicocr, feconicr and feconicrmn concentrated alloys investigated by synchrotron x-ray diffraction, Mater. Des. 155 (2018) 1–7.
[69] O. El-Atwani, N. Li, M. Li, A. Devaraj, J. Baldwin, M. Schneider, D. Sobieraj, J. Wróbel, D. Nguyen-Manh, S.A. Maloy, et al., Outstanding radiation resistance of tungsten-based high-entropy alloys, Sci. Adv. 5 (2019), eaav2002.
[70] A. Fernández-Caballero, J. Wróbel, P. Mummery, D. Nguyen-Manh, Short-range order in high entropy alloys: theoretical formulation and application to mo-nb-ta-vw system, J. Phase Equilibria Diffusion 38 (2017) 391–403.