Machine Learning Force Field Aided Cluster Expansion Approach to Configurationally Disordered Materials: Critical Assessment of Training Set Selection and Size Convergence

Jun-Zhong Xie, Xu-Yuan Zhou, Dong Luan, and Hong Jiang*

Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

E-mail: jianghchem@pku.edu.cn

Phone: (010) 62765970

Abstract

Cluster expansion (CE) is a powerful theoretical tool to study the configuration-dependent properties of substitutionally disordered systems. Typically, a CE model is built by fitting a few tens or hundreds of target quantities calculated by first-principles approaches. To validate the reliability of the model, a convergence test of cross-validation (CV) score to the training set size is commonly conducted to verify the sufficiency of training data. However, such test only confirms the convergence of the predictive capability of the CE model within the training set and it is unknown whether the convergence of the CV score would lead to robust thermodynamic simulation results such as order-disorder phase transition temperature T_c. In this work, using carbon defective MoC_{1-x} as a model system and aided by the machine-learning force field technique, a training data pool with about 13000 configurations has been efficiently obtained and used to generate different training sets of the same size randomly.

By conducting parallel Monte Carlo simulations with the CE models trained with different randomly selected training set, the uncertainty in calculated T_c can be evaluated at different training set size. It is found that the training set size that is sufficient for the CV score to converge still leads to a significant uncertainty in the predicted T_c, and that the latter can be considerably reduced by enlarging the training set to that of a few thousand configurations. This work highlights the importance of considering large training set for building the optimal CE model that can achieve robust statistical modeling results, and the facility provided by the machine-learning force field approach to efficiently produce adequate training data.

1 Introduction

Cluster expansion (CE)[1] is a classic method widely used to study the configuration-dependent properties of alloyed systems.[2–5] The essence of the CE method is to represent the target configuration-dependent property as an expansion of cluster functions with the effective cluster interactions (ECIs) as the coefficients. The representation is in principle exact if all possible clusters are considered, while appropriate truncation[6–8] into a finite cluster set is necessary in practice. In this work we are mainly concerned with the total energy of a given occupation configuration in its relaxed structure. Applications of the CE method to other properties, such as the band gap,[9,10] ion diffusion barrier,[11] tensor properties,[12] and configurational electronic entropy,[13] have also been actively explored. With a selected set of clusters, ECIs are usually determined with the Connolly-Williams structure inversion method[14] by fitting the energies of tens or hundreds of different configurations calculated by first-principles methods. Once validated for its accuracy, the CE model can be used for efficient search of ground state configurations, characterizing ordering behaviors at finite temperature and establishing the phase diagram of alloyed systems.[2]

Although conceptually well established and practically widely used, the CE method, especially regarding what is the optimal strategy to build CE models, has continuously attracted a lot of interest in the recent decades.[6–8,15–28] The main challenge is to build an accurate (unbiased) and robust (with low variance) CE model based on a limited number of training data obtained from computationally demanding first-principles calculations. The cross-validation (CV) technique is widely used for building CE models,[6] in which the whole training data are divided into groups with each group used for testing in turn and the remaining data for training, and one obtains a quantitative measurement of the predictive accuracy of a given CE model, the so-called CV score. For a given set of training data, minimizing the CV score with respect to different cluster selection can usually lead to accurate CE models without suffering from the overfitting problem. One can also validate the sufficiency of the training data by checking the convergence of the CV score as a function of the training set size. This CV-based strategy suffers from the two major difficulties: 1) the minimization of the CV score over all possible selections of clusters is an NP-hard problem,[20] and 2) the CV-score optimized CE model depends quite sensitively on the training set selection. To overcome the first difficulty, various techniques have been proposed, such as hierarchical cluster selection,[5,6] genetic algorithm,[8,15] compressive sensing,[20] and Bayesian inference.[18,21] There have been also several different schema to address the training set selection issue, some emphasizing the importance of including ground state structures,[6,29] and others emphasizing the importance of covering different regions of the configuration space.[17,30]

We note that a large part of the difficulties of the CE method that many methodological developments have tried to tackle can be attributed to the limited number of training data that are obtained from expensive first-principles calculations. Recent developments in machine-learning (ML) force field (FF) methods,[31–33] which can predict the energy of complex systems as accurately as the first principles method that is used to train the force field, but with dramatically reduced computational cost,[32,33] provide a novel framework to address those difficulties faced by the CE method. In particular, in this work we combine the deep potential molecular dynamics (DeePMD) approach,[34–36] which provides a state-of-the-art MLFF implementation based on the deep neural network and active learning techniques,[36] with the cluster expansion method to address the long-standing challenges of the latter related to insufficient training data. We use the carbon defective face-centered cubic α-MoC (α-MoC$_{1-x}$) as a model system. α-MoC has attracted a lot of interest in recent years because of its intriguing catalytic properties.[37–40] The defective nature of experimentally prepared

α-MoC is well established,[41,42] but how carbon vacancies are distributed and their roles played in determining catalytic properties of α-MoC$_{1-x}$ are far from clear. In this work we mainly use α-MoC$_{1-x}$ as a test-bed to showcase novel features of the MLFF-aided CE approach to configurationally disordered materials, and a comprehensive theoretical investigation of structural and thermodynamic properties of α-MoC$_{1-x}$ surface that are relevant to heterogeneous catalysis is scientifically interesting by itself, and will be presented elsewhere.

The paper is organized as the followings. In the next section, we briefly present main ingredients of the theoretical approach used in this work including the cluster expansion method and the deep potential neural network force field, and give some important computational details. In Sec.3, we first present some validation on the accuracy of the MLFF obtained for the α-MoC system, and present our main findings regarding the convergence of the CE models with respect to training set size and selection aided by the large data pool obtained from the MLFF calculations. Sec.4 summarizes the main findings of this work and conclude with some general remarks.

2 Theoretical Methods and Computational Details

2.1 Cluster expansion method

We first give a brief overview of the cluster expansion method using a generic binary alloy A$_{1-x}$B$_x$ as an example. The more systematic formulations for general multi-component and multi-sublattice cases can be found in Refs. 1,3,43,44. For an alloyed system with N sites that can be occupied by either A or B, a configuration $\boldsymbol{\sigma}$ characterizing the occupation of all sites is defined by specifying a spin-like variable to each site, S_i, which is equal to -1 (+1) if the site i is occupied by A (B). The energy per site of a given configuration $\boldsymbol{\sigma} = (S_1, S_2, \cdots, S_N)$ in its locally relaxed structure can be exactly mapped onto the following Ising-like Hamiltonian

$$E(\boldsymbol{\sigma}) = N \sum_{\alpha} m_\alpha J_\alpha \bar{\Pi}_\alpha(\boldsymbol{\sigma}). \tag{1}$$

Here the summation is taken over all symmetrically distinct clusters, with each cluster defined as a particular set of sites denoted by α. J_α is the effective cluster interaction (ECI) associated with α, m_α is the multiplicity per site due to the translational and rotational symmetry of the underlying lattice, and $\bar{\Pi}_\alpha(\boldsymbol{\sigma})$ denotes the cluster function defined as

$$\bar{\Pi}_\alpha(\boldsymbol{\sigma}) = \frac{1}{Nm_\alpha} \sum_{(i_1<i_2<\cdots<i_{g_\alpha})\in\alpha} S_{i_1}S_{i_2}\cdots S_{i_g}, \tag{2}$$

with g_α denoting the number of sites in the cluster, also termed as the order of the cluster. The expansion is exact if all clusters up to the N-body term are considered in the summation,[1,3] but in practice it has to be truncated to consider only two-, three- and sometimes also four-body terms within a certain spatial cutoff, denoted as D_{cut}. ECIs of a truncated CE model are usually determined by fitting the energies of a few tens or hundreds of representative supercell configurations in their locally optimized structures calculated by DFT,[2,5,6,45] known as the Connolly-Williams method or structure inversion method.[14] For a given training set, the clusters included in the CE model are usually selected by using the k-fold cross-validation (CV) technique.[46] The whole training set is divided into k roughly equal-size groups, denoted as S_j with j going from 1 to k. Each group is used for testing in turn and the rest of the data for training. For each candidate set of clusters, the k-fold CV score is calculated by

$$S_{\text{CV}} = \sqrt{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j}\sum_{i\in S_j}(E_i - \hat{E}_{(i,j)})^2}, \tag{3}$$

where E_i is the energy of structure i calculated by DFT, n_j is the number of structures in S_j, and $\hat{E}_{(i,j)}$ is the predicted energy of structure i by using the CE model with ECIs obtained from a least-squares fitting of the data excluding those from the j-th group. With k equal to the size of the training set, one obtains the leave-one-out cross-validation (LOOCV) score, which is widely used in characterizing the predictive capability of CE models.[6] We have compared the effects of using different variants of CV, and obtained essentially the same results. Therefore we will use the

LOOCV, abbreviated as CV henceforth, through the paper.

Once an accurate CE model is built, it can be used to calculate the energy of any configuration efficiently. One of the most important uses of the CE approach is to calculate statistical thermodynamical properties of alloyed systems, for which the central quantities are cluster correlation functions (CCFs), defined as ensemble averaged cluster functions (Eq. 2) at finite temperature T,

$$\langle\bar{\Pi}_\alpha\rangle_T = \frac{\sum_{\boldsymbol{\sigma}} \bar{\Pi}_\alpha(\boldsymbol{\sigma}) \mathrm{e}^{-E(\boldsymbol{\sigma})/k_B T}}{\sum_{\boldsymbol{\sigma}} \mathrm{e}^{-E(\boldsymbol{\sigma})/k_B T}}, \tag{4}$$

which can be typically evaluated by Monte Carlo simulation in a large supercell.[47] The short-range order (SRO) in configurationally disordered materials can be directly revealed by comparing the calculated CCFs to those of a fully random alloy $\mathrm{A}_{1-x}\mathrm{B}_x$, which correspond to the infinite temperature limit of Eq. 4, and can be expressed analytically as[48]

$$\langle\bar{\Pi}_\alpha\rangle_\infty = \langle\prod_i^{g_\alpha} S_i\rangle = \prod_i^{g_\alpha}\langle S_i\rangle = (2x-1)^{g_\alpha}. \tag{5}$$

In particular, for the two-body clusters, the clustering trend of A-A and B-B (A-B) pairs can be indicated by CCFs being greater (smaller) than those of the fully random state.[2] The evolution of CCFs as a function of temperature can be also used to characterize the order-disorder phase transition.[6]

2.2 Deep potential force field model

The generation of large training data pool is facilitated by exploiting the recently developed ML inter-atomic potential model, DeePMD,[35,36,49] as implemented in the DeePMD-kit package,[49] termed as the DP model henceforth. Following the protocol established by Behler and Parrinello,[50] the DP model represents the total energy E of a given structure, denoted by x, as a summation of atomic energies E_i that are dependent on the local environment of each atom

$$E(\mathbf{x}) = \sum_i E_i = \sum_i E_{s_i}(\mathcal{R}_i; \mathbf{w}_{s_i}). \tag{6}$$

Here s_i denotes the chemical species of the i-th atom, and $\mathcal{R}_i$ denotes a set of structural descriptors that characterize the local chemical environment of the i-th atom within a certain cutoff radius R_{cut} and in the meanwhile preserve the translational, rotational and permutational symmetry.[49] The atomic energy E_i is related to $\mathcal{R}_i$ by a deep neural network (DNN) function with parameters denoted as $\mathbf{w}_{s_i}$. To determine the NN parameters encoded in $\mathbf{w}_s$, the following loss function was minimized[34]

$$L\left(p_\epsilon, p_{\text{f}}, p_\xi\right) = p_\epsilon \Delta\epsilon^2 + \frac{p_{\text{f}}}{3n} \sum_{\text{i}} |\Delta \boldsymbol{F}_{\text{i}}|^2 + \frac{p_\xi}{9} \|\Delta\xi\|^2, \tag{7}$$

where $\Delta\epsilon$, $\Delta\boldsymbol{F}_{\text{i}}$ and $\Delta\xi$ represent the root mean square errors in energy, force and virial, respectively. More details about the methodological aspects of DeePMD can be found in Refs.[35,49]

In practice, we used the DFT data obtained during structural optimization of different occupation configurations to train a preliminary DP force field, and then used the “on-the-fly” active learning algorithm[36] implemented in the Deep Potential GENerator (DP-GEN) scheme[51] to refine the force field iteratively. The workflow of DP-GEN includes three main steps: exploration, labeling and training. The general idea of DP-GEN is to efficiently explore the structural space with the force field trained by available data (exploration), find structures that the current force field exhibits significant prediction uncertainty, conduct DFT calculations for those newly found structures, and add them to the training data to obtain the next generation of force field models with improved predictive accuracy. This process is iterated until a sufficient accuracy is reached.

2.3 Computational details

All DFT calculations were conducted with the plane wave based periodic DFT method implemented in the Vienna Ab Initio Simulation Package (VASP).[52,53] The core-valence interaction is described with the projector augmented wave (PAW) method.[54,55] The electron exchange and correlation are treated within the generalized gradient approximation (GGA) in the Perdew-Burke-Ernzerhof (PBE) functional.[56] The energy cut-off for the plane wave basis is set to 450 eV and the convergence criterion for the electronic relaxation loop was set as 10^{-5} eV. Electron occupation is determined by the Gaussian smearing technique with a smearing width of σ = 0.05 eV. A Gamma-centered $\mathbf{k}$-mesh of $3\times3\times3$ was used to sample the Brillioun zone in DFT calculations.

For the DP force field training, we consider 94 different configurations of carbon vacancies randomly generated within a $2\times2\times2$ supercell of α-MoC$_{1-x}$ with x falling in the range of [0.0, 0.50] that covers the vacancy concentration typically found in experimentally prepared samples.[42] Geometric optimization was carried out by DFT, and from the structures generated during relaxation, about 2400 structures were selected as the initial training set. When running DP-GEN calculations, four DP force field models were trained with the same training data and different random seeds for NN parameters. One of the four NN FF models is used to run molecular dynamics (MD) simulation by LAMMPS program[57] to explore the structural space, and the deviation between the forces predicted by four models is used as the indicator for structures that need to be calculated by DFT and included in the training set in the next iteration. The temperature for the MD simulation gradually rises from 10 K to 1000 K in 27 iterations in order to explore the region far away from local energy minima in the structural space in a well controlled manner. About 9300 structures are labeled during the DP-GEN active learning process, and totally about 11700 DFT data are used to train the final DP force field that is employed to generate the training data pool for subsequent CE model building.

Both the construction of CE models and Monte Carlo simulation were accomplished by using Alloy Theoretical Automatic Toolkit (ATAT) packages.[45,58] In the Monte Carlo simulation, we started with a randomly generated α-MoC$_{1-x}$ supercell with $x = 1/3$ that contains 13824 Mo atoms and 9216 C atoms totally, which is chosen in terms of the typical composition in experimentally prepared α-MoC$_{1-x}$ samples.[41] The initial simulation temperature was set as 2500 K and decreased 50 K per step. At each temperature, the numbers of equilibration and sampling steps of Monte Carlo simulation were automatically determined by using the algorithm developed in Ref.

47, with the target convergence for the statistically averaged energy set to 10^{-4} eV.

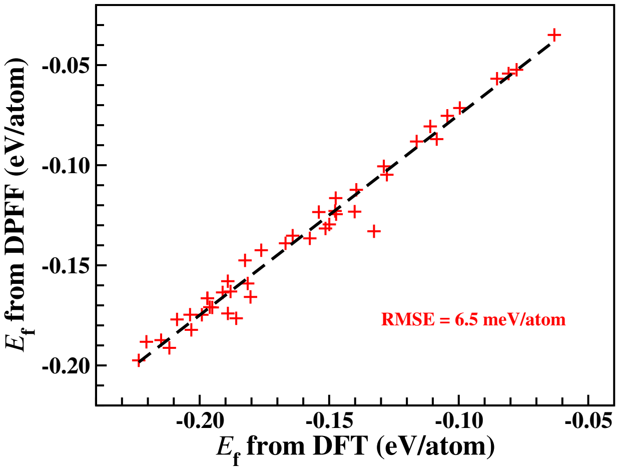

Figure 1: Comparison of the formation energies predicted by the DP force field against DFT results of 44 testing configurations.

3 Results and discussions

3.1 Validation of the NN-FF and construction of the training data pool

All CE models to be considered in the following discussion are built by using the DP force field to calculate the energy of any configuration with both internal coordinates and lattice constants fully relaxed. While the process of building the DP force field itself involves multiple cycles of training-validation steps, we further verify the accuracy of the final DP force field by considering 44 new configurations in a $2 \times 2 \times 2$ α-MoC$_{1-x}$ supercell that are not used in the training process. For each configuration, the structure is relaxed by the DP force field and DFT respectively, and the final total energy is used to calculate the formation energy as

$$E_{\mathrm{f}}(\sigma) = E_{\mathrm{MoC}_{1-x}}(\sigma) - (1-x)E_{\mathrm{MoC}} - xE_{\mathrm{Mo}}, \tag{8}$$

where E_{MoC} and E_{Mo} are the total energy of the stoichiometric α-MoC structure and the metal Mo in the face centered cubic (FCC) structure, respective. Since the FCC Mo, corresponding to 100%

vacancy concentration, is not included in the DP training set, we can expect that the DP approach is incapable to accurately predict its total energy. Therefore we use E_{Mo} and E_{MoC} calculated by DFT when calculating the DP formation energy. In addition, there is usually a nearly constant deviation, i.e. independent of configuration and geometric structure, between DP-predicted total energies and DFT ones, which, however, has no physical effects on subsequent use of the force field in statistical modeling. To account for that factor, we correct all DP total energies by a constant of 25 meV/atom. A comparison of E_{f} from DP and DFT is shown in Fig. 1. The root of mean squared error (RMSE) for the discrepancy between DP and DFT results is only 6.5 meV/atom, indicating that the DP force field we have obtained for α-MoC$_{1-x}$ is able to produce results with an accuracy comparable to that of DFT calculations.

Using the validated DP force field, it becomes feasible to build a training data pool with more than ten thousand different configurations. In particular, we produced about 14000 different configurations with randomly distributed vacancies and the composition parameter x falling in the range of [0.0, 0.5] based on the $2\times2\times2$ supercell of α-MoC. Each configuration is structurally optimized using the DP force field by the LAMMPS program[57] facilitated by the DeePMD-kit package.[49] For a further insurance of the accuracy of the DP results, we use the strategies suggested in the DP-GEN method[36,51] and check the model deviation of all configurations by comparing the energies of all relaxed structures calculated by four DP force field models with the identical NN architecture, trained with the same DFT data but with different random initial seeds for NN hyper-parameters (i.e. $\mathbf{w}_s$). As shown in Fig. S1, the energy uncertainties for most relaxed structures are within 3 meV/atom, and only a rather small number of configurations exhibit large model deviation, mainly located in the high energy region in the configuration space (as shown in Fig. S2). All those significant outliers are eliminated, and we obtain a training data pool with about 13000 configurations, which is used to generate training set for CE models in the following discussion.

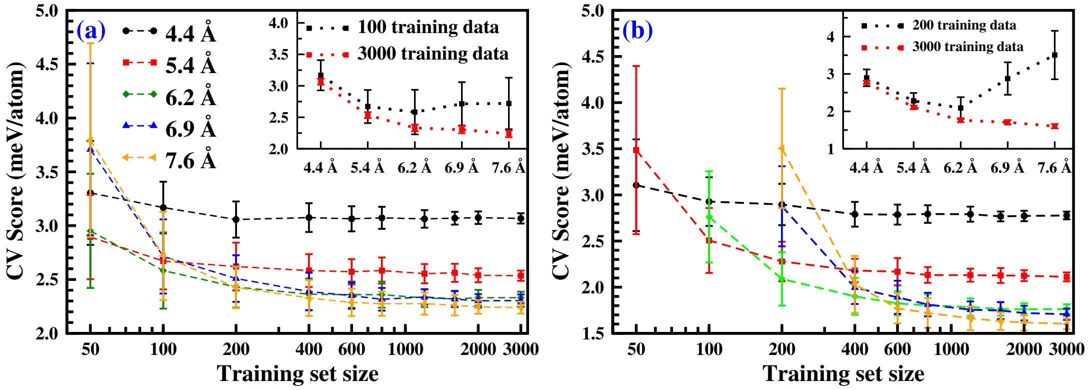

Figure 2: CV scores as a function of training set size in the three-body series (a) and four body cluster series (b) of the CE models. The insets shows the CV scores as a function of the CE model complexity with a given training set size. The error bars are evaluated by the variance of the CV scores in fitting 64 different randomly selected training sets with the same data size from the training data pool.

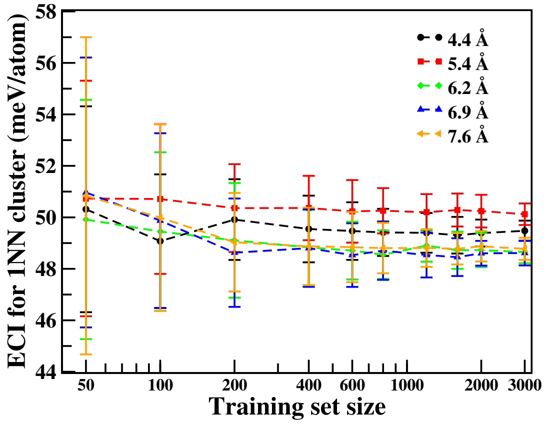

Figure 3: The convergence tendency of 1NN ECI value to the training data size. The mean values and standard errors are evaluated with 64 fittings with different training data of the three-body model series.

3.2 Systematic CV score and ECIs convergence test to training set size

With a training data pool two orders of magnitude larger in size than those in typical CE studies, we can systematically check the CV score convergence with respect to the training set size. In addition, we can also explore the effects of random selection of training data in a more unambiguous manner, since the similarity between the different training sets of the same size is negligible when randomly sampling the the large data pool. The latter issue has never been carefully addressed, to the best of our knowledge, because of demanding cost of DFT calculations of a large number of configurations.

In order to investigate the convergence behavior of CE models with different complexity, the diameter of the largest cluster considered, denoted as $D_{\rm cut}$, is taken as a measurement of the CE model complexity. For given $D_{\rm cut}$, all the clusters to certain order whose diameters are within the range of $D_{\rm cut}$ are included in the CE model. The larger $D_{\rm cut}$ means the more complex CE model with more clusters. We have built two series of CE models. In the first series, three-body model series, we consider all clusters up to three-body ones within $D_{\rm cut}$, and the resultant CE models contain 6, 12, 18, 27, 31 clusters for $D_{\rm cut}$=4.4, 5.4, 6.2, 6.9 and 7.6 Å, denoted as T4.4, T5.4, T6.2, T6.9, and T7.6, respectively. In the second series, four-body model series, we further consider four-body clusters, and the corresponding CE models contain 9, 28, 53, 115, and 144 clusters, denoted as F4.4, F5.4, F6.2, F6.9 and F7.6, respectively.

For a given CE model and training set size $N_{\rm train}$, 64 different randomly selected training sets are used to conduct least-squares fitting, and the mean and variance of the CV score are calculated to characterize the accuracy of the CE model. Fig. 2 shows the convergence of the mean CV score ($S_{\rm CV}$) as a function of training set size $N_{\rm train}$ with the error bars indicating the uncertainty of $S_{\rm CV}$ due to random selection of training set. In general, $S_{\rm CV}$ decrease as $N_{\rm train}$ increases. It can be regarded as converged when the decrease of the mean CV score gained by increasing $N_{\rm train}$ becomes negligible with respect to the error bar. For the purpose of quantitative analysis, the CV score is considered converged if the reduction of the mean $S_{\rm CV}$ is within 0.1 meV/atom when increasing $N_{\rm train}$ up to 3000 in the following discussion. For the three-body model series in Fig.2(a), the simplest T4.4 and T5.4 models converge with only 200 training data, and for T6.2, T6.9, and T7.6 models, 400 training data are required to reach convergence. A similar trend is observed in four-body model series in Fig.2(b). The simpler F4.4 and F5.4 models achieve convergence with 400 training data, while the more complex models F6.9 and F7.6 with over 100 clusters converge with 1200 training data.

From another perspective, for a given training set size, the CV score can be taken as a function of CE model complexity characterized by $D_{\rm cut}$, as shown in the inset of Fig. 2(a) and (b). When the training set size ($N_{\rm train}$=100 or 200) is small, the CV score exhibits a minimum at about $D_{\rm cut} = 6.2$, indicating the occurrence of over-fitting when $D_{\rm cut} > 6.2$. However, when the training data is large enough ($N_{\rm train} = 3000$), over-fitting can be avoided and the CV score decreases monotonically as $D_{\rm cut}$ increases.

Besides the convergence of the mean value of $S_{\rm CV}$, what is also important is its variance caused by random selection of training set, as indicated by the error bars shown in Fig.2. It is noteworthy that the error bars of $S_{\rm CV}$ are quite significant when the training set is small, i.e. a few hundred. With a given set of clusters, the magnitude of the error bars decreases as the training set size increases. As expected, for a given training set size, the more complex model shows a greater uncertainty in the CV score. Generally speaking, the uncertainty of $S_{\rm CV}$ can be regarded as negligible only when $N_{\rm train}$ reaches a few thousand.

For the application of CE models for subsequent statistical simulation, it is more relevant to check how ECIs in the CE model are affected by the training set selection. The uncertainty in ECIs is expected to have significant effects on thermodynamic properties obtained from the CE model. As an illustration, Fig. 3 shows the mean and the variance of the ECI corresponding to the first nearest neighboring (1NN) cluster in the three-body model series as a function of the training set size. For a given cluster selection ($D_{\rm cut}$), the mean of the ECI is nearly constant with increasing $N_{\rm train}$. But a small training set size gives rise to a large error bar of ECI that characterises the certainty with respect to random selection of training set, which can be as large as about 10 meV/atom with $N_{\rm train}$ = 50 or 100. The uncertainty decreases significantly as the training set size increases, and reduces to about 1 meV/atom when $N_{\rm train} = 3000$. Comparing CE models of different complexity, we can see that the ECI of 1NN changes by about 1 meV/atom when $D_{\rm cut}$ increases from 4.4 Å to 5.4 Å, and from 5.4 Å to 6.2 Å, and remains nearly constant when further increasing $D_{\rm cut}$, which is consistent with the trend observed in the CV score as a function of the model complexity.

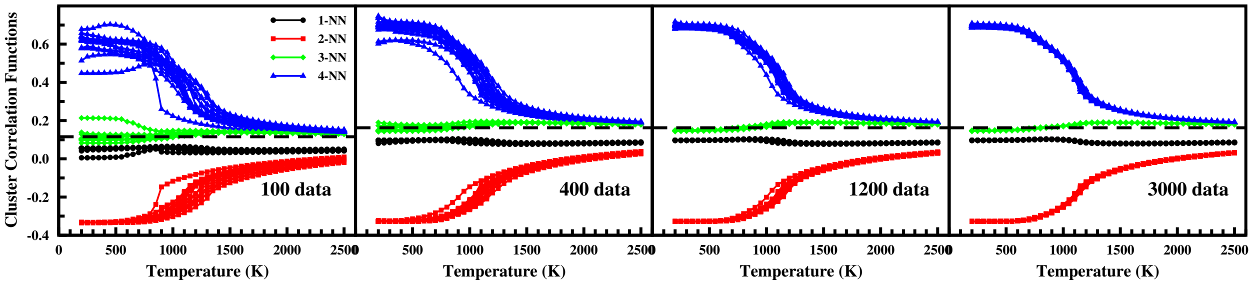

Figure 4: The CCFs of 1NN and 2NN clusters as a function of temperature in Monte Carlo simulation using T6.2. The simulations are performed with CE model trained with 100, 400, 1200, 3000 data. For given training data size, the results of 16 parallel simulations are combined and the CCFs of a given cluster in different simulations are shown with the same color. The dash lines represent the CCF for two-body clusters in a total random configuration.

3.3 Effects of training set selection on order-disorder transition

The CE method is widely used to study order-disorder properties of alloyed systems as a function of temperature, typically calculated by Monte Carlo simulations.[2,47] The studies presented above clearly indicate that the accuracy of the CE model as characterized by the CV score exhibits considerable uncertainty with respect to the training data selection when the training set size is small, i.e. several hundred, and it is therefore crucial to check how such uncertainty affects statistical thermodynamic properties calculated from the CE model. In this part, we investigate how the order-disorder transition behavior is affected by the size and selection of training data, also by conducting a series of parallel Monte Carlo simulations using CE models trained by different training sets. For a given cluster set and training set size, 16 least square fittings are performed with different data of the same size and the resultant CE models are then used in the subsequent Monte Carlo simulation of the carbon vacancy distribution in α-MoC$_{1-x}$ with $x = 1/3$. Taking

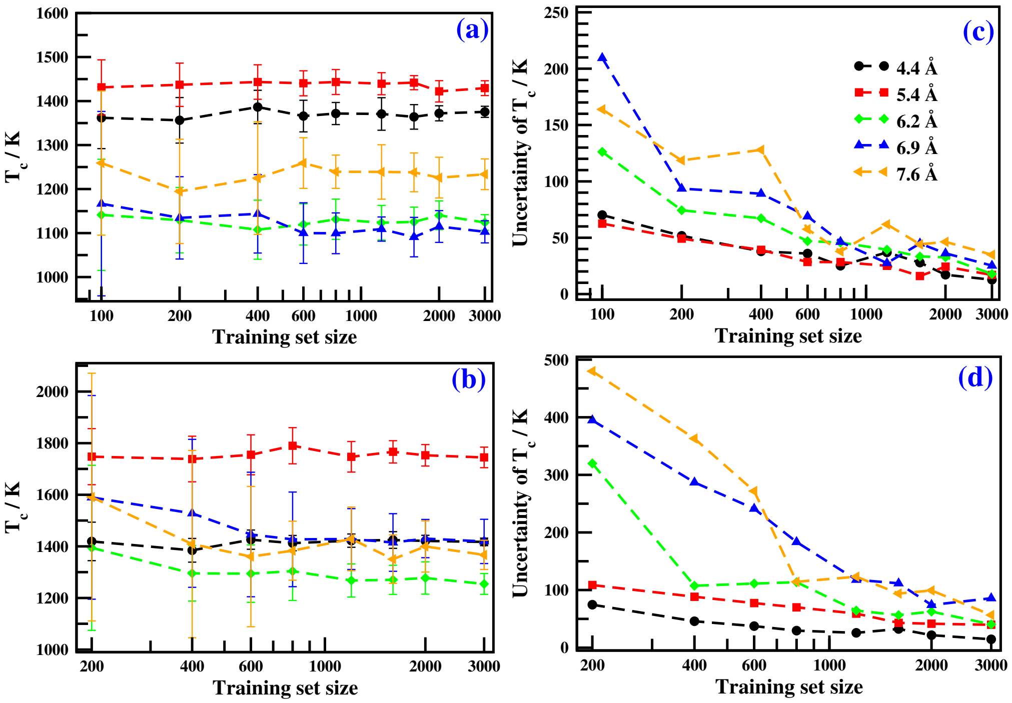

Figure 5: The $T_\mathrm{c}$convergence test to the training data size for three-body model series and four-body model series respectively. The reduction tendency of $T_\mathrm{c}$standard error to the training data size. The standard errors are evaluated by 16 parallel simulations with CE models trained with different data of the same size.

the T6.2 model as an example, the CCFs of smallest two-body clusters are presented as a function of temperature in Fig. 4. CCFs of all two-body clusters obtained from different CE models are presented in Fig.S3-Fig.S12 in Supplementary Materials. In general the results from different models exhibit a similar trend. At low temperature, the CCFs deviate strongly from those of the totally random state, corresponding to an ordered phase, and as the temperature increases, they gradually get close to those of the fully random state, indicating a typical second-order phase transition. It is noteworthy that even well above the transition temperature, the CCFs of some two-body clusters still exhibit significant deviation from those of the fully disordered state, indicating the existence of short-range order. What is noteworthy is that there exists considerable uncertainty in calculated CCFs caused by random selection of training data, especially when N_{train} is small (i.e. a few hundred), and the impact of training data selection becomes negligible as the training set size increases to 3000.

In order to check the convergence of calculated CCFs quantitatively, we extract the critical temperature T_{c} for the order-disorder transition from each simulation by fitting the CCF of the 2NN cluster with the cubic spline, and determining T_{c} as the temperature at which the curvature of the CCF curve changes the sign. The mean value and uncertainty of T_{c} for given D_{cut} and training data size are then calculated, as shown in Fig. 5. In the results for the three-body model series, presented in Fig.5(a) and (c), the mean of T_{c} converges rather quickly as a function of N_{train}, especially for simple CE models ($D_{\mathrm{cut}} = 4.4$ and 5.4 Å). In contrast, the uncertainty in calculated T_{c} converges much more slowly as N_{train} increases, and is significantly larger for more complex CE models. To be more specific, the T_{c} uncertainty from the T4.4 and T5.4 models is already smaller than 50 K when $N_{\mathrm{train}} \geq 200$, but that from more complex models (T6.2, T6.9 and T7.6) requires $N_{\mathrm{train}} \simeq 2000$ to achieve a similar accuracy. The results for the four-body model series, shown in Fig. 5(b) and (d), exhibit similar features, but the uncertainty in T_{c} is almost twice larger due to significantly increased model complexity. Even with $N_{\mathrm{train}} = 3000$, T_{c} from the F7.6 model still has a uncertainty of about 100 K.

It is interesting to make a comparison between the convergence behaviors of the T_{c} uncertainty and the CV score with respect to the training data size. For the three-body model series, although the CV scores can be regarded as converged with about 200 training data, there is still significant uncertainty of T_{c} over 50 K for simple T4.4 and T5.4 and over 100K for complex T6.2, T6.9 and T7.6 models (Fig. 5(c). The four-body cluster model series shows similar behavior. For example, the F6.2 model trained with 800 data, well converged in terms of the CV score, still gives an T_{c} uncertainty of over 100 K.

We close this section by discussing the effects of the model complexity, characterized by D_{cut}, on the calculated T_{c}, using the results obtained with $N_{\mathrm{train}} = 3000$, in which the uncertainty of T_{c} is relatively small. As shown in Fig.5(a) and (c), T_{c} does not show any obvious convergence tendency as the model complexity (D_{cut}) increases. Within the three-body model series, T_{c} from T6.2 differs from that of T7.6 by about 100 K, even though the difference in the CV from the two models is less than 0.1 meV/atom. The same is also true for the four-body model series. The T_{c} values from three-body and four-body series with the same D_{cut} also differ significantly, e.g. 150 K for D_{cut}=6.2Å. It is therefore clear that the convergence of the CE model with respect to cluster selection as characterized by the CV score does not necessarily lead to robust statistical thermodynamic properties.

4 Conclusion

To summarize, in this work, using carbon-defective α-MoC$_{1-x}$ as a typical substitutionally disordered system, we have systematically studied the convergence of the CV score in building CE models and the resultant thermodynamic properties, taking the order-disorder phase transition temperature T_{c} as a representative, with respect to the training set size and random selection of training data. Aided by the deep neural network-based machine learning force field technique, a large training data pool containing more than ten thousand structures with different carbon vacancy configurations has been efficiently constructed with the accuracy of first-principle calculation, and is used for the subsequent CE model building and convergence test. The main findings of this work can be summarized as the following points: 1) The mean value of the CV score converges quite rapidly with increasing training set size (N_{train}), but the uncertainty due to random selection of training data converges much more slowly, and becomes negligible only when N_{train} is as large as about several thousand, which is about one order of magnitude larger than the training set size typically used in previous CE model building. 2) The calculated order-disorder transition temperature T_{c} exhibits significant uncertainty with respect to random selection of training data, especially when the CE model is complex (i.e. with a large D_{cut}) and N_{train} is small (i.e. a few hundred), and the uncertainty decreases significantly to be less than 100 K when N_{train} reaches several thousand. It is therefore clear that the convergence of the CV score alone can not guarantee robust statistical thermodynamic modeling results. 3) With a large training set, although the CV score converges well with respect to model complexity, the calculated T_{c} does not show a clear convergence with respect to D_{cut}, which calls for further methodological development in the CE-based framework to achieve more accurate prediction of thermodynamic properties. The results presented in this work are obtained by using the least square fitting technique to build the CE models. We have also tested the more sophisticated compressive sensing technique implemented in the least absolute shrinkage and selection operator (LASSO) algorithm[20] to build the CE models, and we obtained similar results regarding the effects of training set size and selection. It should be emphasized that the findings summarized above are achieved thanks to the availability of the large training data pool generated with the aid of the machine learning force field well trained and validated based on DFT calculation. One can expect that machine-learning techniques will play increasingly more important roles in theoretical study of substitutionally disordered materials, as clearly demonstrated recent works in the literature (see Ref. 59 and references therein).

Acknowledgement

The author thanks Dr. Han Wang, Dr. Linfeng Zhang and Mr. Yuzhi Zhang for the help on the use of DeePMD-kit and DP-GEN. This work is partly supported by the National Natural Science

Foundation of China (Projects No. 21873005 and 21911530231) and High-performance Computing Platform of Peking University.

Supporting Information Available

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/xxxx.

The model deviation test for all the locally relaxed structures of the candidate configurations, Figure S1-S2. All the CCFs data in the Monte Carlo simulation in this work are also included, Figure S3-S12.

References

(1) Sanchez, J. M.; Ducastelle, F.; Gratias, D. Generalized cluster description of multicomponent systems. *Physica A* **1984**, *128*, 334–350.

(2) Zunger, A. In *First-principles Statistical mechanics of semiconductor alloys and intermetallic compounds*; Turchi, P. E. A., Gonis, A., Eds.; Plenum Press, 1994; pp 361 – 418.

(3) Fontaine, D. D. In *Solid State Physics*; Ehrenreich, H., Turnbull, D., Eds.; Academic Press, 1994; Vol. 47; pp 33–176.

(4) Ruban, A. V.; Abrikosov, I. A. Configurational thermodynamics of alloys from first principles: effective cluster interactions. *Rep. Prog. Phys.* **2008**, *71*, 046501.

(5) van de Walle, A. Methods for First-Principles Alloy Thermodynamics. *JOM* **2013**, *65*, 1523–1532.

(6) van de Walle, A.; Ceder, G. Automating first-principles phase diagram calculations. *J. Phase Equilib.* **2002**, *23*, 348–359.

(7) Zarkevich, N. A.; Johnson, D. D. Reliable First-Principles Alloy Thermodynamics via Truncated Cluster Expansions. *Phys. Rev. Lett.* **2004**, *92*, 255702.

(8) Hart, G. L. W.; Blum, V.; Walorski, M. J.; Zunger, A. Evolutionary approach for determining first-principles hamiltonians. *Nat. Mater.* **2005**, *4*, 391–394.

(9) Magri, R.; Froyen, S.; Zunger, A. Electronic structure and density of states of the random Al0.5Ga0.5As, GaAs0.5P0.5, and Ga0.5In0.5As semiconductor alloys. *Phys. Rev. B* **1991**, *44*, 7947–7964.

(10) Xu, X.; Jiang, H. Cluster expansion based configurationalaveraging approach to bandgapsof semiconductor alloys. *J. Chem. Phys.* **2019**, *150*, 034102.

(11) Van der Ven, A.; Ceder, G.; Asta, M.; Tepesch, P. D. First-principles theory of ionic diffusion with nondilute carriers. *Phys. Rev. B* **2001**, *64*, 184307.

(12) van de Walle, A. A complete representation of structure–property relationships in crystals. *Nat. Mater.* **2008**, *7*, 455–458.

(13) Zhou, F.; Maxisch, T.; Ceder, G. Configurational Electronic Entropy and the Phase Diagram of Mixed-Valence Oxides:The Case of LixFePO4. *Phys. Rev. Lett.* **2006**, *97*, 155704.

(14) Connolly, J. W. D.; Williams, A. R. Density-functional theory applied to phase transformations in transition-metal alloys. *Phys. Rev. B* **1983**, *27*, 5169–5172.

(15) Blum, V.; Hart, G. L. W.; Walorski, M. J.; Zunger, A. Using genetic algorithms to map first-principles results to model Hamiltonians: Application to the generalized Ising model for alloys. *Phys. Rev. B* **2005**, *72*, 165113.

(16) Drautz, R.; Díaz-Ortiz, A. Obtaining cluster expansion coefficients in ab initio thermodynamics of multicomponent lattice-gas systems. *Phys. Rev. B* **2006**, *73*, 224207.

(17) Seko, A.; Koyama, Y.; Tanaka, I. Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations. *Phys. Rev. B* **2009**, *80*, 165122.

(18) Mueller, T.; Ceder, G. Bayesian approach to cluster expansions. *Phys. Rev. B* **2009**, *80*, 024103.

(19) Arnold, B.; Ortiz, A. D.; Hart, G. L.; Dosch, H. Structure-property maps and optimal inversion in configurational thermodynamics. *Phys. Rev. B* **2010**, *81*, 094116.

(20) Nelson, L. J.; Hart, G. L.; Zhou, F.; Ozoliņš, V., et al. Compressive sensing as a paradigm for building physics models. *Phys. Rev. B* **2013**, *87*, 035125.

(21) Nelson, L. J.; Ozoliņš, V.; Reese, C. S.; Zhou, F.; Hart, G. L. Cluster expansion made easy with Bayesian compressive sensing. *Phys. Rev. B* **2013**, *88*, 155105.

(22) Kristensen, J.; Zabaras, N. J. Bayesian uncertainty quantification in the evaluation of alloy properties with the cluster expansion method. *Comput. Phys. Commun* **2014**, *185*, 2885–2892.

(23) Seko, A.; Tanaka, I. Cluster expansion of multicomponent ionic systems with controlled accuracy: importance of long-range interactions in heterovalent ionic systems. *J. Phys.: Condens. Matter* **2014**, *26*, 115403.

(24) Herder, L. M.; Bray, J. M.; Schneider, W. F. Comparison of cluster expansion fitting algorithms for interactions at surfaces. *Surf. Sci.* **2015**, *640*, 104–111.

(25) Aldegunde, M.; Zabaras, N.; Kristensen, J. Quantifying uncertainties in first-principles alloythermodynamics using cluster expansions. *J. Comput. Phys.* **2016**, *323*, 17 – 44.

(26) Nguyen, A. H.; Rosenbrock, C. W.; Reese, C. S.; Hart, G. L. Robustness of the cluster expansion: Assessing the roles of relaxation and numerical error. *Phys. Rev. B* **2017**, *96*, 014107.

(27) Leong, Z.; Tan, T. L. Robust cluster expansion of multicomponent systems using structured sparsity. *Phys. Rev. B* **2019**, *100*, 134108.

(28) Kleiven, D.; Akola, J.; Peterson, A. A.; Vegge, T.; Chang, J. H. Training sets based on uncertainty estimates in the cluster-expansion method. *Journal of Physics: Energy* **2021**, *3*, 034012.

(29) Garbulsky, G. D.; Ceder, G. Linear-programming method for obtaining effective cluster interactions in alloys from total-energy calculations: Applications to the fcc Pd-V system. *Phys. Rev. B* **1995**, *51*, 67 – 72.

(30) Seko, A.; Tanaka, I. Grouping of structures for cluster expansion of multicomponent systems with controlled accuracy. *Phys. Rev. B* **2011**, *83*, 224111.

(31) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phy.* **2016**, *145*, 170901.

(32) Chan, H.; Narayanan, B.; Cherukara, M. J.; Sen, F. G.; Sasikumar, K.; Gray, S. K.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S. Machine Learning Classical Interatomic Potentials for Molecular Dynamics from First-Principles Training Data. *J. Phys. Chem. C* **2019**, *123*, 6941–6957.

(33) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390.

(34) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.

(35) Zhang, L.; Han, J.; Wang, H.; Saidi, W. A.; Car, R., et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *arXiv preprint arXiv:1805.09003* **2018**,

(36) Zhang, L.; Lin, D.-Y.; Wang, H.; Car, R.; E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **2019**, *3*.

(37) Lin, L.; Zhou, W.; Gao, R.; Yao, S.; Zhang, X.; Xu, W.; Zheng, S.; Jiang, Z.; Yu, Q.; Li, Y.-W.; Shi, C.; Wen, X.-D.; Ma, D. Low-temperature hydrogen production from water and methanol using Pt/α-MoC catalysts. *Nature* **2017**, *544*, 80–83.

(38) Yao, S. et al. Atomic-layered Au clusters on α-MoC as catalysts for the low-temperature water-gas shift reaction. *Science (New York, N.Y.)* **2017**, *357*, 389–393.

(39) Lin, L.; Yao, S.; Gao, R.; Liang, X.; Yu, Q.; Deng, Y.; Liu, J.; Peng, M.; Jiang, Z.; Li, S.; Li, Y.-W.; Wen, X.-D.; Zhou, W.; Ma, D. A highly CO-tolerant atomically dispersed Pt catalyst for chemoselective hydrogenation. *Nat. Nanotechnol* **2019**, *14*, 354–361.

(40) Zhang, X.; Zhang, M.; Deng, Y.; Xu, M.; Artiglia, L.; Wen, W.; Gao, R.; Chen, B.; Yao, S.; Zhang, X., et al. A stable low-temperature H 2-production catalyst by crowding Pt on α-MoC. *Nature* **2021**, *589*, 396–401.

(41) Rudy, E.; Windisch, S.; Stosick, A. J.; Hoffman, J. R. CONSTITUTION OF BINARY MOLYBDENUM–CARBON ALLOYS. *Trans. Met. Soc. AIME* **1967**, *239*, 1247.

(42) Lee, J. S.; Volpe, L.; Ribeiro, F.; Boudart, M. Molybdenum carbide catalysts: II. Topotactic synthesis of unsupported powders. *J. Catal.* **1988**, *112*, 44–53.

(43) Sanchez, J. M. Cluster expansions and the configurational energy of alloys. *Phys. Rev. B* **1993**, *48*, 14013–14015.

(44) Sanchez, J. M. Cluster expansion and the configurational theory of alloys. *Phys. Rev. B* **2010**, *81*, 224202.

(45) van de Walle, A. Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit. *CALPHAD* **2009**, *33*, 266 – 278.

(46) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed.; Springer: New York, 2009.

(47) van de Walle, A.; Asta, M. Self-driven lattice-model Monte Carlo simulations of alloy thermodynamic properties and phase diagrams. *Modell. Simul. Mater. Sci. Eng.* **2002**, *10*, 521.

(48) Wei, S.-H.; Ferreira, L. G.; Bernard, J. E.; Zunger, A. Electronic properties of random alloys: Special quasirandom structures. *Phys. Rev. B* **1990**, *42*, 9622–9649.

(49) Wang, H.; Zhang, L.; Han, J.; Weinan, E. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun* **2018**, *228*, 178–184.

(50) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(51) Zhang, Y.; Wang, H.; Chen, W.; Zeng, J.; Zhang, L.; Wang, H.; Weinan, E. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput. Phys. Commun* **2020**, *253*, 107206.

(52) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **1996**, *6*, 15–50.

(53) Kresse, G.; Furthmüller, J. Efficient iterative schemes for \textit{ab initio} total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, *54*, 11169.

(54) Blöchl, Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, 17953–17979.

(55) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758–1775.

(56) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. **1996**, *77*, 3865.

(57) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(58) van de Walle, A.; Asta, M.; Ceder, G. The Alloy Theoretic Automated Toolkit: A User Guide. *Calphad* **2002**, *26*, 539–553.

(59) Hart, G. L. W.; Mueller, T.; Toher, C.; Curtarolo, S. Machine learning for alloys. *Nature Rev. Mater.* **2021**, *6*, 730 – 755.