Strategy to consider element distribution when constructing training datasets for developing machine learning potentials of alloys based on a Monte-Carlo-like method

Zhipeng Zhang,[1] Liuqing Chen,[1] Junyi Guo,[1] Xianyin Duan,[2,*] Bin Shan,[3,†] and Xianbao Duan[1,2,§]

[1]*Hubei Key Laboratory of Plasma Chemistry and Advanced Materials and School of Materials Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, Hubei, China*

[2]*Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan 430081, Hubei, China*

[3]*State Key Laboratory of Material Processing and Die and Mould Technology and School of Materials Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China*

(Received 5 August 2022; accepted 20 September 2022; published 30 September 2022)

Molecular dynamics simulations can explore the characteristics and evolution of microstructures in alloys outside of experiments, with reliability and accuracy guaranteed by the interatomic potentials employed. Machine learning potential (MLP) is widely used for its accuracy close to first-principles calculations. When developing an MLP, the construction of the training dataset is crucial, determining the accuracy and generalization of the MLP. In this work, a Monte-Carlo-like (MCL) strategy is proposed to construct training datasets for developing MLPs of alloys, which is characterized by the efficient consideration of element distributions in alloys. As an example, a training dataset for the equimolar NbTiZrHf alloy is constructed based on the MCL strategy, and the corresponding MLP is developed subsequently. By comparing with two traditional strategies, it is found that the training dataset constructed based on the MCL strategy has greater dispersion, and the corresponding MLP has better prediction performance. In addition, a hybrid molecular statics and Monte Carlo simulation with the MCL-based MLP is performed to optimize the element distribution of the equimolar NbTiZrHf alloy, and segregation and short-range ordered structures are observed in the final configuration, which is consistent with the experimental results reported in the literature. The MCL strategy proposed in this work can provide a fast solution for considering the element distribution when constructing training datasets for developing MLPs of alloys.

DOI: 10.1103/PhysRevB.106.094107

I. INTRODUCTION

In recent years, high-entropy alloys (HEAs) or multicomponent alloys have been rapidly developed [1–6]. Different from traditional alloys, which are mainly composed of one or two principal elements and a trace of auxiliary elements, HEAs generally have more than four or five principal elements [1,2]. Due to the cocktail effect brought about by the mixing of various elements, HEAs often exhibit outstanding mechanical, physical, chemical, and biological properties, such as high thermal stability, high corrosion resistance, etc., and thus have potential applications in aerospace, nuclear energy, and other fields [3–6]. In early studies on HEAs, it was generally believed that the elements in HEAs were completely randomly distributed in the matrix, thus forming corresponding random solid solutions. However, a growing number of experimental and theoretical research have shown that this is not the case. In other words, the elements in HEAs are not completely random and disordered, but form segregation or short-range ordered structures (SROs) in some localized regions, which have important effects on many properties of HEAs [7–9]. Due to the compositional complexity of HEAs [10], it is still very challenging to investigate SROs at the atomic scale based on experiments. As an important complement to the experimental study, molecular dynamics (MD) simulations can provide insight into the dynamic evolution process of SROs and further explore their impact mechanism on various properties [11–19].

When performing MD simulations on a specific material system, the first and very crucial step is to construct the corresponding interatomic potential, which is used to accurately describe the interactions between atoms in the system. For metal or alloy systems including HEAs, the potentials used earlier are mostly empirical, such as embedded atom method and modified EAM [14,19]. Such potentials are usually constructed by first proposing some descriptive formulas with parameters based on the physical model of atomic interactions, and then determining the relevant parameters by fitting some properties of a specific material obtained from experiments or first-principles calculations. Recently, researchers have attempted to introduce machine learning theory into the development of interatomic potentials, resulting in a new class of potentials quite different from empirical potentials, namely machine learning potentials (MLPs) [20]. Typical MLPs are based on the framework of the neural network,

*xyduan@wust.edu.cn
†bshan@mail.hust.edu.cn
§xianbao.d@gmail.com

developed by inputting a large amount of data obtained from first-principles calculations to train the weight coefficients. Although MLPs are quite different from empirical potentials in form, they are essentially the same, which is to map the information related to the atomic configuration, such as atomic coordinates, species, etc., into potential energy or atomic forces. In contrast, the accuracy of MLPs is generally higher than that of empirical potentials, approaching that of first-principles calculations. As a trade-off, the computational cost of MLPs is also typically greater than that of empirical potentials, but significantly lower than that of first-principles calculations. As computing power increases, the requirement for accuracy outweighs the savings for computational cost, resulting in the growing application of MLPs in MD simulations [16,17].

Although MLPs have different theoretical frameworks, the development process is similar. The first step in developing an MLP for a specific material system is to construct a training dataset where each sample corresponds to a unique atomic configuration [20]. By performing first-principles calculations on each atomic configuration, the corresponding total potential energy and atomic forces can be obtained. The coordinates and species directly determined by the configuration, as well as the calculated potential energy and atomic forces, together constitute a training sample. Therefore, the basis for constructing a training dataset is to construct an atomic configuration dataset, which is essentially sampling in the corresponding configuration space. For pure metals, the difference between samples in the configuration space is merely atomic coordinates, while for alloys, the difference also includes the element distribution. Obviously, for HEAs composed of multiple elements, the configuration space is huge, so it is impossible to construct a configuration dataset based on exhaustion. Instead, the configuration space can only be limitedly sampled based on a certain strategy. Depending on the purpose for which the MLP is used, the sampling strategy will also vary. The most common strategy for constructing the configuration dataset is to perform different modes of deformation operations, such as stretching, shearing, etc., or MD simulations at different temperatures on completely random solid-solution models and extract some configurations from the corresponding trajectories [21,22]. It should be pointed out that deformation operations or MD simulations at lower temperatures can only change the relative positions of atoms in the configuration, but hardly change the element distribution. Therefore, the MLP developed using the training dataset constructed from random element distribution models (named as the Random strategy) is difficult to describe segregation or SROs that may appear in alloys. Another improved strategy is to generate so-called special quasirandom structures (SQS), referred to as the SQS strategy, which optimizes the shape of the lattice jointly with the occupation of the atomic sites, thus ensuring that the configuration space searched is exhaustive and not biased by a prespecified lattice shape [23,24]. Nevertheless, the diversity of atomic configurations constructed based on the SQS strategy is still limited, and the element distribution characteristics corresponding to the configuration space cannot be fully exploited, leading to the possibility of failing to reproduce segregation or SROs under special conditions.

In this work, we propose a Monte-Carlo-like strategy (called the MCL strategy) to construct training datasets for developing MLPs. In the implementation of the MCL strategy, we first define an average probability parameter (APP) that characterizes the distribution of different element pairs, so that an atomic configuration with a specific element distribution can be described by a set of APPs. Then, we discretize each set of APPs within its value range, which is equivalent to the uniform sampling of the configuration space. Finally, we introduce the Monte-Carlo-like method to target each set of APPs and construct the corresponding configuration. To verify the effectiveness of the MCL strategy, we construct the training dataset of equimolar NbTiZrHf HEA as an example, and further develop the corresponding MLP. Meanwhile, we also construct the training datasets based on the Random strategy and the SQS strategy and develop corresponding MLPs. By comparison, it is found that the training dataset constructed based on the MCL strategy has greater dispersion, while the corresponding MLP has better prediction performance. Further, we perform a hybrid molecular statics (MS) and Monte Carlo (MC) simulation using the MCL-based MLP to optimize the element distribution of the equimolar NbTiZrHf HEA. By comparing with the experimental results reported in the literature, it is found that the MLP can accurately reproduce the segregation and SROs present in NbTiZrHf HEA, again verifying the reliability of the MCL strategy.

II. METHODOLOGY

A. Construction of the training datasets

As mentioned above, the first step in developing an MLP for a specific alloy system is to construct a training dataset based on atomic configurations. In this study, we selected the quaternary equimolar NbTiZrHf HEA as an example, and adopted different strategies, including the Random strategy, the SQS strategy, and the MCL strategy, to construct training datasets and develop corresponding MLPs, respectively. The following is an introduction to the MCL strategy proposed in the present work.

To quantitatively describe the element distribution of HEAs and thus provide an indicator for constructing the training dataset, a simple parameter, named as the *average probability parameter*, was proposed and applied to construct an atomic configuration with a certain element distribution. Considering the APP between element m and element n in the HEA (denoted as $\bar{P}_{mn}$), its definition is as follows:

$$\bar{P}_{mn} = \left\langle \frac{Z_{mn}}{Z_m} \right\rangle, \tag{1}$$

where Z_m is the total number of all-type atoms around a particular m-type atom within a specified cutoff distance, and Z_{mn} is the number of n-type atoms within the same cutoff distance. The notation $\langle \ldots \rangle$ means to take the average of the value within it over all m-type atoms in the alloy. In the present work, the cutoff distance was taken between the first and the second nearest neighbors. As a result, Z_m can be determined directly by the crystal structure of the alloy, which is equal to the number of first-nearest neighbors. From the above definition, it can be seen that the value of $\bar{P}_{mn}$ is between 0 and 1. The closer $\bar{P}_{mn}$ is to 1, the more likely n-type atoms will appear

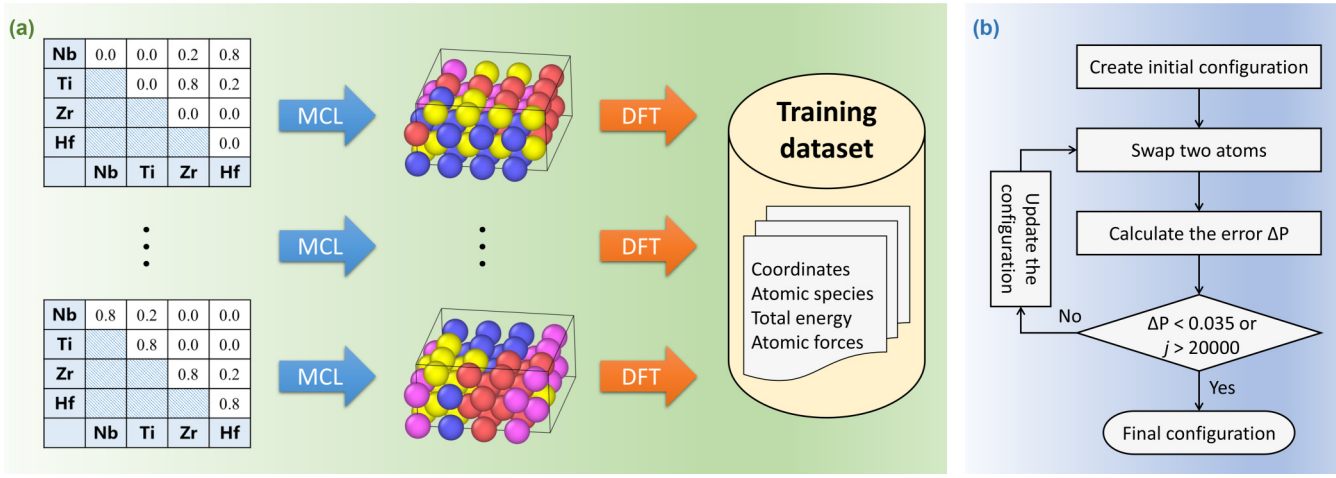

FIG. 1. (a) Schematic diagram of the construction of the training dataset used for developing the MLP of NbTiZrHf HEA based on the MCL strategy. (b) Flowchart for generating an atomic configuration targeting a specified APP set using the MCL method.

around m-type atoms, resulting in a more ordered arrangement between m-type atoms and n-type atoms, whereas, if $\bar{P}_{mn}$ is close to 0, it means that the m-type and n-type atoms tend to be separated from each other. For a given HEA, each pair of elements, including elements of the same type and different types, corresponds to such an APP. As far as the quaternary NbTiZrHf HEA is concerned, there are ten pairs of elements, of which four pairs are composed of the same element (i.e., Hf-Hf, Nb-Nb, Ti-Ti, and Zr-Zr) and six pairs are composed of different elements (i.e., Hf-Nb, Hf-Ti, Hf-Zr, Nb-Ti, Nb-Zr, and Ti-Zr). Therefore, for the NbTiZrHf HEA with a specific atomic configuration, it corresponds to a set of ten APPs. For the convenience of description, *a set of APPs* corresponding to a specific atomic configuration is abbreviated as *an APP set*. It should be pointed out that the APPs in an APP set are not completely independent, but have some inherent constraints. Obviously, for any m-type element in the alloy, the sum of all APPs associated with the m-type element should be 1.

With the definition of APP, it can be seen that the APP set is a quantitative description of the atomic configuration with a specific element distribution. Therefore, if we want to construct a configuration dataset with high dispersion, that is, each configuration in the dataset has essentially different element distribution as much as possible, we only need to construct configurations corresponding to different APP sets. If these APP sets are highly dispersed, then of course the corresponding configurations are dispersed. With that come two problems to be solved: one is how to define a series of APP sets with high dispersion, and the other is how to generate a corresponding atomic configuration according to a specified APP set. In the present work, the first problem was solved by discretizing the APP set in its valid range and the second was solved by proposing an MCL method. As a note, when mentioning the *MCL method* in this paper, we refer to the process of constructing a configuration targeting a specific APP set from an initial configuration, while the *MCL strategy* refers to a set of methods for constructing a training dataset with the MCL method as the core. Figure 1(a) shows the schematic diagram of the construction of the training dataset used for developing the MLP of NbTiZrHf HEA based on the MCL strategy, explained below.

Firstly, the APP set is discretized within its valid range. For each APP in the set, its theoretical range is from 0 to 1. However, considering the actual element distribution in HEA, the APP cannot reach 1. For the NbTiZrHf HEA, assuming the APP between Hf and Nb is 1, it means that the first-nearest neighbors of Hf atoms are all Nb atoms, and the first-nearest neighbors of Nb atoms are all Hf atoms; that is, these two types of atoms are completely separated from the other two, which is unrealistic in an alloy. Therefore, we took a value of 0.8 as the upper bound. Another thing to consider is the discrete interval. Obviously, the smaller the discrete interval, the more APP sets will be generated, and the greater the subsequent computational cost will be. For example, if the discrete interval is 0.2, there will be about 1000 APP sets, and if 0.1 is taken, there will be nearly 30 000 APP sets. Considering the subsequent computational cost, the interval of 0.2 was taken, resulting in a total of 986 APP sets.

Secondly, the MCL method is used to generate the atomic configuration targeting a specified APP set. Figure 1(b) shows the corresponding flowchart. Specifically, we first built a random single-phase solid-solution model of equimolar NbTiZrHf HEA as the initial configuration. According to previous research [25–27], we set the lattice structure to body-center cubic (bcc) and created a $4a \times 4a \times 2a$ (a is the lattice constant) supercell with a total of 64 atoms (verification of the supercell size will be simply discussed in Sec. III A). The lattice constant a was defined as the value corresponding to the lowest energy of the potential energy per atom versus the lattice constant curve of the random solid-solution NbTiZrHf HEA model obtained by performing density-functional theory (DFT) calculations. Then, two atoms of different types in the configuration were swapped, resulting in a new configuration. Next, the APP set corresponding to the new configuration was calculated and then the error between the APP set and the target APP set was computed. The error, expressed as ΔP,

is defined as follows:

$$\Delta P = \sqrt{\frac{1}{10}\sum_{i=1}^{10}\left(\bar{P}_i - \bar{P}_i^{\text{target}}\right)^2}, \tag{2}$$

where i denotes the ith pair of elements in the alloy, $\bar{P}_i$ is the calculated ith APP of the configuration, and $\bar{P}_i^{\text{target}}$ is the target ith APP. In Eq. (2), the number 10 represents the ten pairs of elements in the quaternary NbTiZrHf HEA. For alloys with N elements, the value should be replaced by $N(N+1)/2$. The next step was to decide whether to continue the process. Two different criteria were considered: one is based on the error ΔP and the other is based on the total number of iterations. If the error ΔP is less than 0.035 or the number of iterations is greater than 20 000, the process will stop; otherwise, it will continue. Here, the values used in the two criteria are empirical, which can be determined by testing the convergence of the error ΔP. In this work, the number of iterations can fully guarantee the error ΔP of the final configuration converges to a constant value. As the process continues, the MCL method was employed to update the configuration. Assuming that the current iteration step is j and the error of the corresponding configuration is ΔP_j, if the error ΔP_j is smaller than that of the previous configuration ΔP_{j-1}, the current configuration is directly accepted for the next iteration. Conversely, if ΔP_j is greater than ΔP_{j-1}, the current configuration is accepted with the following probability:

$$p_j = e^{\frac{-|\Delta P_j - \Delta P_{j-1}|}{kT}}, \tag{3}$$

where k is the Boltzmann constant and T is the temperature. In the present work, T was set to 300 K. If, unfortunately, the current configuration is rejected, the previous configuration will be used again. After the configuration has been updated, the process will continue. Observing the above process, it can be found that the MCL method used in the process is very similar to the traditional MC method [28], with the main difference being that one criterion is changed from the potential energy to the error of the APP set. For each process using the MCL method, an atomic configuration will be generated, targeting an APP set. Therefore, the process will be performed many times according to the total number of the discretized APP sets, resulting in the same number of configurations.

Finally, by performing DFT calculations on each configuration, the training dataset can be constructed from the atomic configuration dataset obtained in the previous step, as shown in Fig. 1(a). As mentioned above, each sample in the training dataset is composed of atomic coordinates, species, total potential energy, and atomic forces. In terms of computational details, Vienna *Ab initio* Simulation Package (VASP) [29] was adopted to perform all DFT calculations in the present work. Exchange and correlation were treated at the Perdew-Burke-Ernzerhof functional level [30]. The plane-wave energy cutoff was set at 400 eV. Brillouin-zone sampling was performed using the Monkhorst-Pack scheme [31] with a $2 \times 2 \times 4\,k-$point mesh. The global break condition of energy for the electronic self-consistency loop was set to be 1×10^{-8} eV and the maximum number of electronic self-consistency step was set to be 60. The results show that, for all structures calculated using DFT in this work, the energy and force precisions for each structure can reach about 1×10^{-7} eV and 1×10^{-5} eV/Å, respectively.

To compare the training dataset generated based on the MCL strategy with traditional strategies, two more training datasets were constructed based on the Random strategy and the SQS strategy, each including 1000 samples. For the training dataset constructed based on the Random strategy, the element distribution of each configuration was completely randomly generated. For the training dataset constructed based on the SQS strategy, the samples were generated by performing the SQS process on 50 initial configurations with random element distribution, and sampling 20 configurations for each process. The lattice structures, lattice parameters, and details of DFT calculations for generating the three training datasets were all the same, with the only difference being the element distribution.

The dispersion of the training dataset is crucial for developing an MLP with good prediction performance. To learn about the dispersion of the training datasets constructed based on three different strategies in the present work, two indicators were employed to evaluate them: one is the potential energy per atom and the other the well-known Warren-Cowley parameter (WCP) [32,33]. For the convenience of readers, the WCP of the m-n pair in the alloy is given as follows:

$$\text{WCP} = 1 - \frac{Z_{mn}}{\omega_n Z_m}, \tag{4}$$

where ω_n is the atomic fraction of n-type atoms in the alloy, and Z_{mn} and Z_m are the same as in Eq. (1). According to the definition, if the WCP is close to 0, the m-type atoms appear randomly around the n-type atoms. If the WCP is less than 0, the m-type atoms are more likely to appear around the n-type atoms, and if the WCP is greater than 0, the m-type atoms are with less probability to appear around the n-type atoms.

B. Development of the machine learning potentials

After the construction of the training dataset for a specified HEA, the corresponding MLP can be developed by training a specific machine learning model. So far, many researchers have developed a series of models for MLPs, and many have provided the corresponding codes [34–36]. In contrast, DeePMD [36–39] is one of the relatively mature MLP models: on the one hand, it provides a package called DEEPMD-KIT [37] that can be conveniently used to develop MLPs from training datasets, and on the other hand, it is compatible with the widely used MD software package LAMMPS [40]. Since the birth of DeePMD, extensive MD studies have been carried out based on DeePMD, ranging from simple water molecules [36] to complex HEAs [21,22,41]. In the present work, the software package DEEPMD-KIT was adopted to construct the MLP of the quaternary NbTiZrHf HEA, and LAMMPS was employed to perform all MS or MC simulations with the developed MLPs.

When developing an MLP using DEEPMD-KIT, some parameters need to be set. In the present work, the relevant parameters were set as follows: (1) the embedded network was set to a three-layer residual network with the nodes in each layer being 25, 50, and 100, respectively; (2) the fitting

network was set to a three-layer fully connected network with 240 nodes in each layer; (3) the cutoff distance was set to 8.0 Å with the smooth operation starting from 2.0 Å; (4) the start value and the limit value of energy error for the prefactor in the loss function were set to 1000 and 1, respectively; those of force error were set to 0.02 and 2, respectively; and those of virial error were set to 0 because no virial data were used in the present work; and (5) a total of 100 000 epochs were performed during the training. For the above three training datasets, the corresponding MLPs were developed with the same parameter settings.

C. Construction of the testing datasets

To evaluate the reliability, accuracy and generalization of the developed MLPs, testing datasets need to be constructed. To avoid spuriously high prediction accuracy, the testing dataset should meet at least the following two requirements: (1) it should be independent of the training datasets; and (2) the samples in it are well dispersed. In the present work, we propose a similar strategy to MCL to construct the testing dataset. The main difference is the evolution direction of the Monte-Carlo-like process: for the MCL strategy of generating training samples, the evolution direction is to reduce the error of the APP set, so as to ensure that the APP set of the other configuration is closer to the corresponding target APP set, while for the method of generating testing samples, the evolution direction is to increase the error, thereby ensuring that the other configuration is more different from the initial one. Therefore, it can be said that the process of generating testing samples is based on a reverse MCL strategy, so we call it the rMCL strategy for short. The process of constructing a testing dataset based on the rMCL strategy can be described as follows briefly.

Firstly, several initial atomic configurations of single-phase solid-solution NbTiZrHf HEA with random element distribution were generated. The lattice points for these atomic configurations were the same as those used in the training dataset. Secondly, the process based on the rMCL strategy was performed on each initial configuration, resulting in a corresponding evolution trajectory of configuration, which contains many other configurations. By extracting configurations from each trajectory, the atomic configurations for constructing the testing dataset could be obtained. The generation process can be described as (1) calculate the APP set of the initial configuration; (2) swap two atoms of different types in the initial configuration arbitrarily to generate another configuration; (3) calculate the error of the APP set between the other configuration and the initial one; (4) decide whether to continue the process: only if the number of iterations is greater than a specified number, the process will stop; and (5) update the configuration using the rMCL method as the process continues: if the error ΔP_j is *larger* than ΔP_{j-1}, the current configuration is accepted for the next iteration; otherwise, the current configuration is accepted with the probability given by Eq. (3). After the configuration has been updated, the process will continue. By comparing the above process with that of MCL, it can be observed clearly that the main difference is the method used for saving the other configuration: for MCL, the method is to save configurations with smaller errors from the target, while for rMCL, the method is to save those with larger errors from the initial one. It should be pointed out that although the error of the APP set between the configuration in the trajectory and the initial one increases with the number of iterations, there may be an inherent correlation between them, which might affect the dispersion performance of the generated testing dataset. Therefore, it is recommended not to generate a very long trajectory from one initial configuration and then extract all new configurations from the trajectory to construct the testing dataset, but to generate several relatively short trajectories from different initial configurations and then extract some configurations from each trajectory. In the present work, four initial configurations were used to generate four trajectories, and 50 configurations were extracted from each trajectory to constitute four testing datasets, labeled as testing 1, 2, 3, and 4, respectively. Finally, DFT calculations with the same settings as those used to generate the training datasets were performed on each configuration to obtain the corresponding total potential energy and atomic forces to construct the testing datasets.

III. RESULTS AND DISCUSSION

A. Evaluation of the training datasets

To evaluate the effectiveness of the MCL strategy, the errors of the APP set between the generated atomic configurations and corresponding targets are plotted in Fig. 2(a) for all samples in the training dataset. It can be seen clearly that most configurations have errors between the stopping criteria 0.035 and 0.1, with only a few having errors slightly greater than 0.1. It is worth mentioning that to explore whether the supercell size used in the present work has an effect on the error of the generated configuration, we selected 40 APP sets as targets and used a larger $4a \times 4a \times 4a$ supercell to generate corresponding configurations. The inset in Fig. 1(a) shows the comparison of the errors using two different supercells. As can be seen, the errors are basically the same, demonstrating that the supercell size used in the present work is appropriate. To reduce the subsequent computational cost, a smaller supercell size of $4a \times 4a \times 2a$ was adopted.

To gain further insight into the distribution of APP error over its value domain, Fig. 2(b) presents the comparison of APP between the actual values and corresponding targets for all configurations, taking the Nb-Hf pair as an example. It can be observed that the actual APP values are basically consistent with corresponding targets in trend. In detail, when the value is between 0.1 and 0.7, the actual APP values agree well with the corresponding targets, while when the value is less than 0.1 or greater than 0.7, the actual values have relatively large deviations from the targets. Such a result is reasonable because when the APP takes a value close to the boundaries of the value range, the element distributions of corresponding configurations are in an extreme state, such as severe segregation, which is unlikely to appear in statistical theory or in practice. Therefore, it can be concluded from Fig. 2 that the MCL strategy used in the present work can efficiently generate atomic configurations with small errors from the target APP sets.

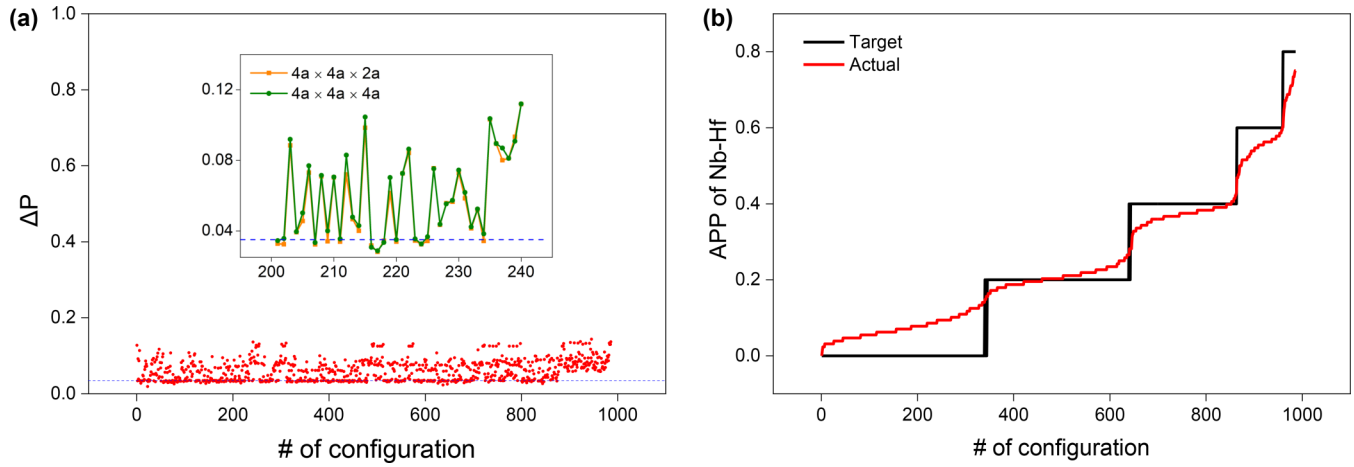

FIG. 2. (a) Errors of the APP set between the atomic configurations in the training dataset generated based on the MCL strategy and corresponding targets. The inset shows a comparison of the errors for 40 configurations generated using two different supercells. (b) Comparison of APP of Nb-Hf between the actual values of the generated configurations and corresponding targets.

To learn about the dispersion of the training dataset constructed based on three different strategies, the WCPs of different element pairs for each atomic configuration in the training datasets constructed based on the MCL strategy, the Random strategy, and the SQS strategy are computed and shown in Figs. 3(a)–3(c), respectively. As seen in Fig. 3(a), the WCPs for the training dataset constructed based on the MCL strategy vary from −1.6 to 1.0 with a variation interval

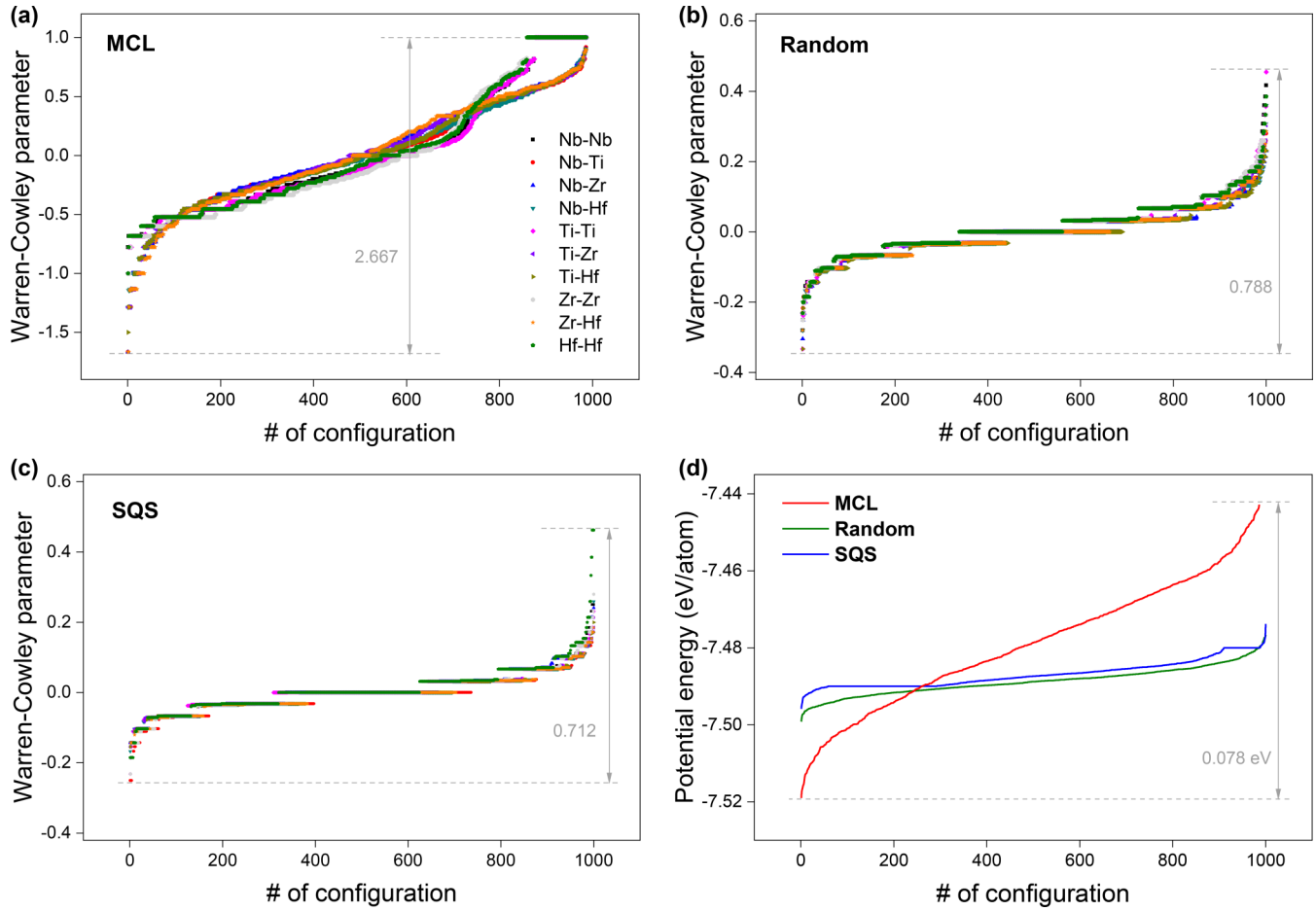

FIG. 3. WCPs of different element pairs for each atomic configuration in the training datasets constructed based on (a) the MCL strategy, (b) the Random strategy, and (c) the SQS strategy. (d) Comparison of potential energy per atom for each atomic configuration in the training datasets constructed based on different strategies.

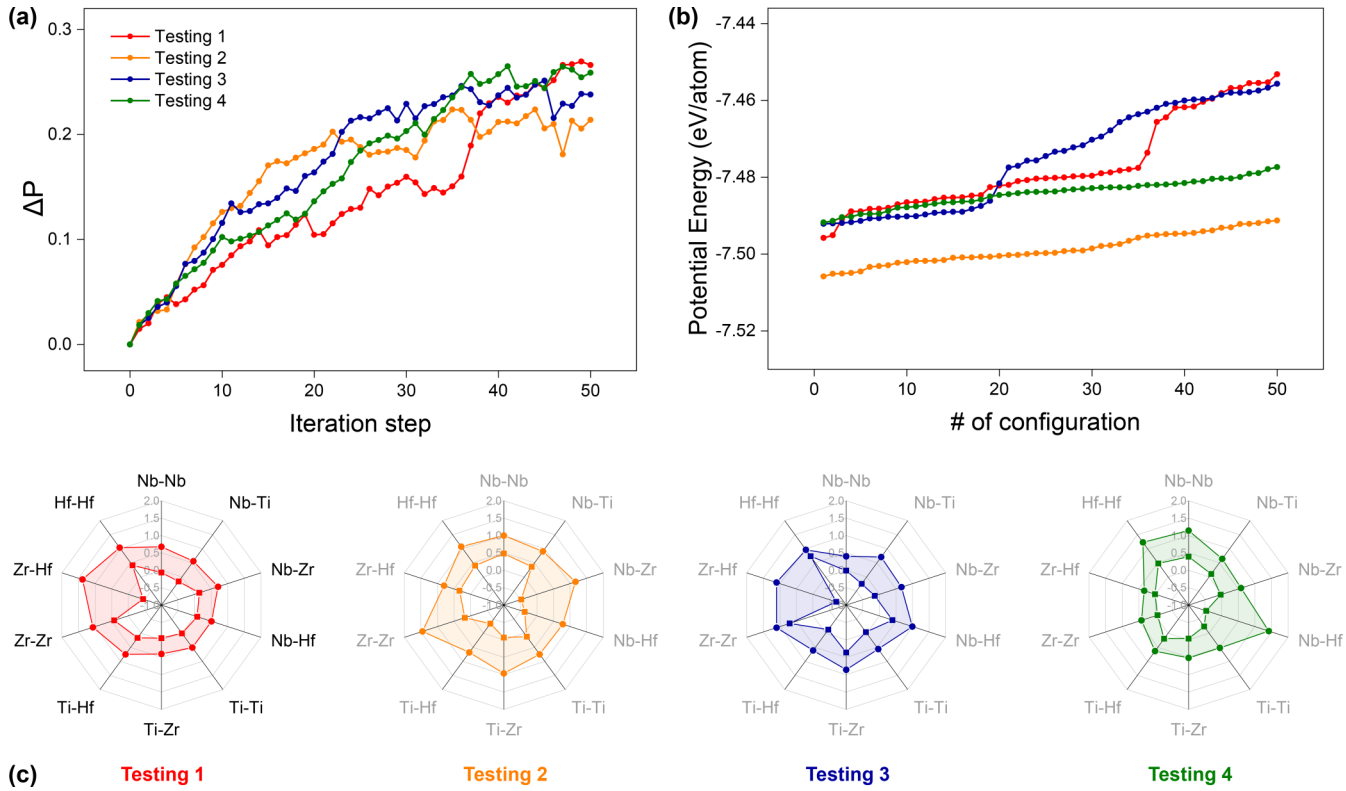

FIG. 4. (a) APP set error, (b) potential energy per atom, and (c) WCPs of different element pairs for each configuration in the four involved testing datasets constructed based on the rMCL strategy.

of 2.667. Comparing Figs. 3(b) and 3(c), the training datasets constructed based on the Random strategy and the SQS strategy have relatively small variation intervals of WCP, 0.788 and 0.712, respectively. Therefore, it can be concluded that the training dataset constructed based on the MCL strategy has a higher dispersion of element distribution, which makes the corresponding MLP more likely to capture the segregation or SROs that may exist in the configuration. It should be noted that WCP is only an indicator describing the element distribution of an atomic configuration and has no direct relationship with the potential energy. In other words, a training dataset with high dispersion of WCP cannot guarantee the high dispersion of potential energy. Therefore, it is also necessary to understand the dispersion of potential energy separately. Figure 3(d) shows the comparison of potential energy per atom for each atomic configuration in the training datasets constructed based on different strategies. It can be observed that the maximum potential-energy difference is of the training dataset constructed based on the MCL strategy and reaches 0.078 eV per atom, which is significantly larger than those of the other two datasets, indicating that even thermodynamically, the training dataset constructed based on the MCL strategy exhibits better dispersion. In conclusion, the training datasets constructed based on the MCL strategy have better performance on the dispersion of samples.

B. Evaluation of the testing datasets

Next, the four testing datasets constructed based on the rMCL strategy were evaluated. Figure 4(a) shows the curves of the APP set errors between the generated configurations based on the rMCL strategy and corresponding initial configurations as a function of iteration steps for the four involved testing datasets. As can be seen clearly, the APP set errors of the generated configurations show an overall upward trend with slight fluctuations, which is consistent with the aim of the rMCL strategy employed. In other words, as the number of iteration steps increases, the generated configuration has an increasingly large difference in element distribution from the initial one. In addition, it can also be found that during the process of generating the configurations, although there are some differences in the APP set error between these four testing datasets, they are close in order of magnitude. To further explore the differences, we characterized the four testing datasets from the two aspects of potential energy per atom and WCP, and the results are shown in Figs. 4(b) and 4(c), respectively. What should be noted, for the convenience of comparison, is that the abscissa of Fig. 4(b) is set as the serial number of the configuration, which is arranged in ascending order according to the potential energy corresponding to the configuration in each testing dataset, regardless of the generation order. From Fig. 4(b), the potential energies of the configurations in Testing 1, 3, and 4 are close and generally high, while those for testing 2 are much lower, indicating that testing 2 is fundamentally different from the other three testing datasets. Comparing testing 1, 3, and 4, there is a sudden increase in potential energy for testing 1 and 3, but not for testing 4, with the result that almost half of the configurations in testing 1 and 3 have relatively low energies close to those in testing 4, while the others have higher energies.

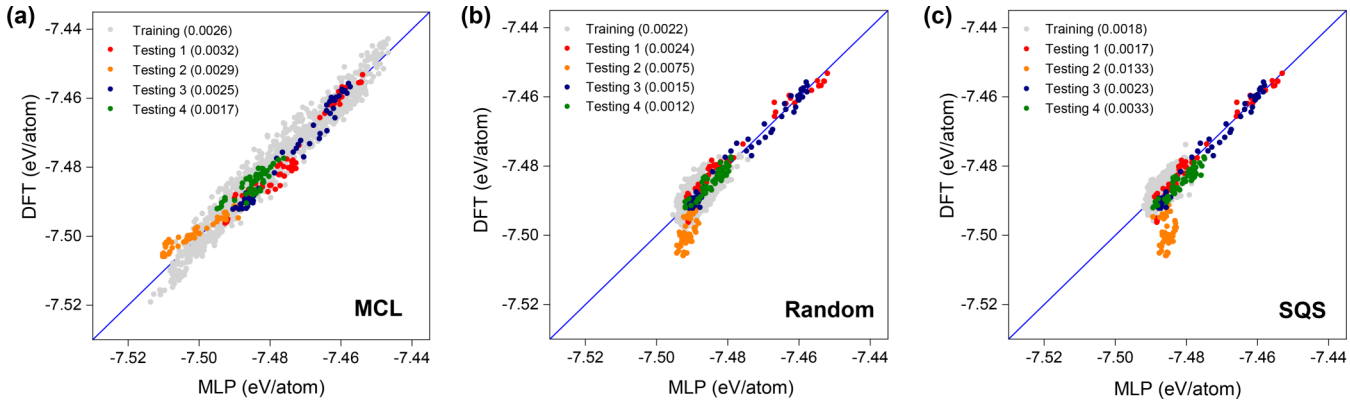

FIG. 5. Comparison of the potential energy per atom for each configuration in the training and testing datasets calculated by DFT and MS with MLPs developed using training datasets constructed based on different strategies: (a) MCL, (b) Random, and (c) SQS. Numbers in the legend indicate the MAEs of corresponding datasets.

Figure 4(c) shows the WCP ranges of different element pairs for all configurations in each testing dataset, where the inner and outer boundaries represent the corresponding minimum and maximum values, respectively. It can be observed that the patterns of testing 1 and 3 are basically the same, which is consistent with the potential-energy results presented in Fig. 4(b). In contrast, testing 2 and 4 are quite different from them in pattern. Therefore, according to the results of various aspects, it can be inferred that testing 1 and 3 should be close, while testing 2 and 4 should be relatively unique. In conclusion, the rMCL strategy can be used to generate configurations that have a large difference from the initial one in element distribution, and the four testing datasets constructed in the present work have more or fewer differences from each other.

C. Evaluation of the machine learning potential

After constructing the three training datasets based on different strategies, namely MCL, Random, and SQS, the corresponding MLPs for NbTiZrHf HEA were developed using DeepMD as the framework, respectively. To test the accuracy and prediction performance of these MLPs and further evaluate these strategies used for generating the training datasets, the three training datasets and the four testing datasets constructed based on the rMCL strategy were employed. Specifically, the potential energies for each configuration in these datasets were calculated using both MS simulations with the three developed MLPs and DFT calculations. All results are plotted in Fig. 5, where Figs. 5(a)–5(c) correspond to the MCL strategy, the Random strategy, and the SQS strategy, respectively. To facilitate quantitative comparisons, the mean absolute error (MAE) relative to the DFT results was calculated for each dataset and presented after the corresponding legend in Fig. 5. The analysis of Fig. 5 is given below.

First, from the relative positions of the samples corresponding to the training dataset and the testing datasets in each subfigure of Fig. 5, it can be found that almost all testing samples are within the range of the training samples constructed based on the MCL strategy. In contrast, for the training samples constructed based on the Random strategy and the SQS strategy, almost all samples in testing 4, a part of samples in testing 1 and 3, and a few samples in testing 2 are within the range of corresponding training samples. Therefore, it can be concluded again that the training dataset constructed based on the MCL strategy has better dispersion than those based on the Random strategy and the SQS strategy. Second, it can be seen from Fig. 5 that the training samples (gray dots) constructed based on different strategies basically lie around the 45 ° line (blue line) representing the good agreement between the DFT and MS results. Quantitatively, the MAE of each training dataset is about 0.002 eV per atom, which is a very low value, indicating that the MLP developed based on each training dataset has good reproduction accuracy. It can also be found that the MAE of the training dataset constructed based on the MCL strategy is slightly larger than those based on the Random strategy and the SQS strategy, mainly because the training dataset constructed based on the MCL strategy has a wider distribution of potential energy [as seen from Fig. 3(d)], which may reduce the corresponding convergence accuracy to some extent. Fourth, from Fig. 5(a), the MAEs of the four testing datasets are all very close to that of the training dataset, demonstrating that the MLP developed based on the MCL strategy has good prediction performance and can accurately calculate the potential energy of configurations not in the training dataset. Fifth, the same four testing datasets have different levels of MAE when using MLPs developed based on the Random strategy and the SQS strategy, as shown in Figs. 5(b) and 5(c). Specifically, the MAEs of testing 1, 3, and 4 are relatively small, while that of testing 2 is relatively large, almost 5 to 10 times larger than the others, which can also be seen in Figs. 5(b) and 5(c), where the samples of testing 2 (orange dots) deviate severely from the 45 ° line. For testing 4 (green dots), the samples are basically within the ranges of the training datasets, so it is to be expected that the corresponding potential energies can be accurately calculated. For testing 1 (red dots) and testing 3 (navy blue dots), there are quite a few samples with higher potential energies that are outside the corresponding training datasets, but their potential energies are still calculated accurately, reflecting the prediction ability of corresponding MLPs. For testing 2, the samples are basically with lower potential energies than the training datasets, and their potential energies cannot be calculated accurately compared to the DFT results. From these results, it appears that the MLPs constructed based on the Random strategy and the SQS

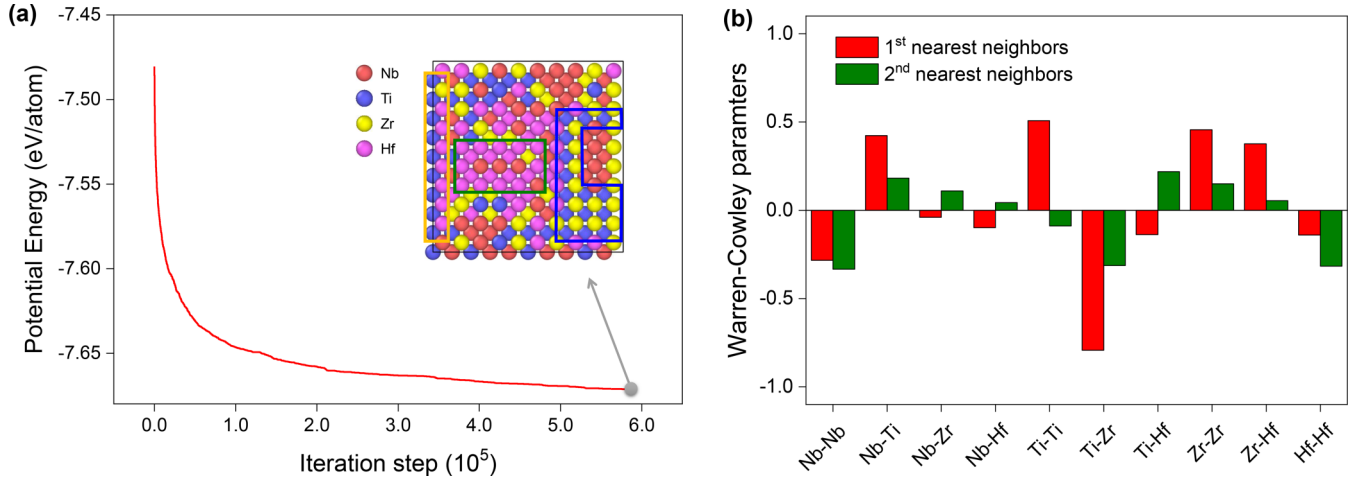

FIG. 6. (a) Evolution of the potential energy per atom with the iteration step for NbTiZrHf HEA. The inset shows the snapshot of the configuration at the end of the MS/MC simulation. (b) WCPs of different element pairs for the final configuration of NbTiZrHf HEA, where the red and green bars correspond to the WCPs considering the first- and second-nearest neighbors, respectively.

strategy method have better predictive ability when computing configurations with higher potential energies than the training samples, and less ideal for configurations with lower energies. In summary, the MLP developed based on the MCL strategy exhibits better predictive ability than those developed based on the Random strategy and the SQS strategy, mainly due to the wider energy coverage of the training dataset constructed based on the MCL strategy.

D. Application of the machine learning potential

To further verify the reliability of the developed MLP based on the MCL strategy, we performed a hybrid MS/MC simulation to optimize the element distribution of equimolar NbTiZrHf HEA. Specifically, a $10a \times 10a \times 10a$ supercell of NbTiZrHf HEA with randomly distributed elements was first built as the initial configuration, where the lattice constant a was equal to that of the configuration in the training dataset. Then, the hybrid MS/MC simulation was performed on the initial configuration, which can be briefly described as (1) calculate the potential energy per atom of the initial configuration using MS simulation with the developed MCL-based MLP, denoted as E_i; (2) swap two atoms of different types in the initial configuration to generate another configuration; (3) calculate the potential energy per atom of the new configuration, denoted as E_{i+1}; (4) decide whether to continue the process: if the energy drop $E_i - E_{i+1}$ is less than 1×10^{-5} eV per atom, the process will stop; otherwise, the process will continue; and (5) update the configuration using the classical MC strategy: if the potential energy E_{i+1} is less than E_i, the current configuration is accepted for the next iteration; otherwise, the current configuration is accepted with a certain probability. Finally, a total of about 600 000 MC steps were performed.

Figure 6(a) plots the potential energy per atom of NbTiZrHf HEA as a function of iteration steps during the MS/MC process. As can be seen from the figure, the potential energy drops sharply in the first tens of thousands of steps and then decreases slowly until it approaches an equilibrium state, resulting in an energetically favorable configuration. To gain an intuitive understanding of the resulting element distribution, we visualized the final configuration according to element type, as shown in the inset of Fig. 6(a). It can be observed clearly that some elements in the configuration tend to form segregation, such as Hf-Hf (pink), while some element pairs tend to form SROs, such as Ti-Zr (blue-yellow). To quantitatively understand the aggregation tendency of different element pairs in the final configuration, we computed the corresponding WCPs of different element pairs. Since the value of WCP can be affected by the cutoff distance used, we set two different cutoff distances, 3 and 4 Å, which correspond to the WCPs considering the first- and second-nearest neighbors, respectively, and the results are displayed in Fig. 6(b). From the WCPs for the first-nearest neighbors (red bars), the same element pairs of Nb-Nb and Hf-Hf and different element pairs of Ti-Zr and Ti-Hf all have negative WCPs, suggesting that they tend to aggregate together and appear in each other's first-nearest neighbors. In contrast, the element pairs Ti-Ti, Zr-Zr, Nb-Ti, and Zr-Hf have positive WCPs, demonstrating that they tend to separate from each other and have less chance of appearing in each other's first-nearest neighbors. By comparing with the WCPs for the second-nearest neighbors (green bars), a decrease in the absolute value of the negative WCP of the Ti-Zr pair can be observed, indicating that Ti atoms and Zr atoms have less probability of appearing in each other's second-nearest neighbors, while the WCP of the Ti-Ti pair changes from positive to negative, and that of the Zr-Zr pair decreases significantly, demonstrating that both Ti atoms and Zr atoms tend to appear in their own second-nearest neighbors. Combining these analyses, it can be inferred that the Ti-Zr pair has a strong tendency to form ordered arrangements, which can also be observed from the inset of Fig. 6(a), where Ti atoms and Zr atoms form SROs with a B2 structure, as shown in the navy blue wireframe. A similar result appears for the Ti-Hf pair, as displayed by the orange wireframe. For the Hf-Hf pair, the corresponding WCP for the second-nearest neighbors becomes more negative, indicating that Hf atoms are likely to appear in their own second-nearest neighbors.

Therefore, Hf atoms tend to form segregation, as shown by the green wireframe. A similar result can be seen for the Nb-Nb pair. In fact, some of these results obtained by performing the hybrid MS/MC simulation with the developed MCL-based MLP can be confirmed experimentally. Bu *et al.* [9] used *in situ* transmission electron microscopy to observe the element distribution in the as-cast NbTiZrHf HEA and found both Ti-rich clusters enriched with Zr atoms and Hf-rich clusters, which is in good agreement with the results predicted in the present work.

In summary, the developed MCL-based MLP has good accuracy and prediction performance, so the MCL strategy proposed in the present work is effective to construct training datasets used for developing MLPs of HEAs. It should be noted that the MCL strategy here only considers the element distribution of atomic configuration when constructing the training dataset, and does not consider the corresponding atomic positions, which is also very important when developing practical MLPs, especially for research on mechanical properties. Therefore, when constructing an MLP for a specific HEA, a suggested scheme is to use our proposed MCL strategy to account for its element distribution, and use different modes of deformation operation, such as stretching, shearing, etc., or *ab initio* MD simulations, to account for the displacement of atoms.

IV. SUMMARY

In this work, we proposed an MCL strategy to construct training datasets for developing MLPs of alloys, especially for HEAs. Specifically, we first defined a parameter called APP to describe the element distribution of an atomic configuration. As a result, an atomic configuration with a specific element distribution can be represented by a set of APPs. Then, we discretized the APP set within its value range to achieve uniform sampling of the element distribution space. After that, we targeted each discretized APP set to construct configurations with specific element distributions using an MCL method. As an example, we constructed the training dataset of equimolar NbTiZrHf HEA, and further developed the corresponding MLP based on the framework of DeePMD. To illustrate the characteristics of the MCL strategy, we also constructed two more training datasets based on the traditional Random strategy and SQS strategy and developed the corresponding MLPs. To test these developed MLPs, we also proposed a strategy called rMCL, which can be regarded as a reverse version of the MCL strategy, to construct four testing datasets of NbTiZrHf HEA. We then evaluated the three training datasets using two metrics, WCP and potential energy per atom, and found that the training dataset constructed based on the MCL strategy had greater dispersion. Four testing datasets were also evaluated similarly, validating the effectiveness of the rMCL strategy. Further, we employed the four testing datasets to evaluate the accuracy and prediction performance of the MLPs developed based on different strategies and found that the MCL-based MLP had the best prediction performance, which can be attributed to the high dispersion of the corresponding training datasets. Finally, we investigated the optimized elemental distribution of NbTiZrHf HEA by performing a hybrid MS/MD simulation with the MCL-based MLP. Through visualization and WCP analysis of the final configuration, we found that the Ti-Zr pair tends to form SROs with a B2 structure, while the Hf-Hf pair tends to form segregation, which is consistent with the results observed in the experiments. In conclusion, our proposed MCL strategy can be used to consider the element distribution when constructing training datasets for developing MLPs. Besides, our proposed APP can be used to describe the element distribution of atomic configurations for alloys, especially for HEAs, and our proposed rMCL strategy can be used to construct testing datasets with high dispersion.

ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China (Grant No. 52273237) and the Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering (Grant No. MTMEOF2020B01). Thanks to the High-Performance Computing Center of Wuhan University of Science and Technology for providing computing resources and technical support.

[1] M. C. Gao, J. Yeh, P. K. Liaw, and Y. Zhang, *High-Entropy Alloys Fundamentals and Applications* (Springer International Publishing, New York City, 2016).

[2] C. Niu, First principles studies of NiFeCrCoMn high entropy alloys, Doctoral dissertation, North Carolina State University, 2015.

[3] X. Wang and F. Kong, Recent development in high-entropy alloys and other high-entropy materials, J. Aeronaut. Mater. **39**, 1 (2019).

[4] O. N. Senkov, D. B. Miracle, K. J. Chaput, and J. P. Couzinie, Development and exploration of refractory high entropy alloys—A review, J. Mater. Res. **33**, 3092 (2018).

[5] X. Yan and Y. Zhang, Functional properties and promising applications of high entropy alloys, Scr. Mater. **187**, 188 (2020).

[6] C. Han, Q. Fang, Y. Shi, S. B. Tor, C. Chua, and K. Zhou, Recent advances on high-entropy alloys for 3D printing, Adv. Mater. **32**, 1903855 (2020).

[7] M. Tian, C. Wu, Y. Liu, H. Peng, J. Wang, and X. Su, Phase stability and microhardness of $CoCrFeMn_xNi_{2-x}$ high entropy alloys, J. Alloys Compd. **811**, 152025 (2019).

[8] D. Karlsson, A. Marshal, F. Johansson, M. Schuisky, M. Sahlberg, J. M. Schneider, and U. Jansson, Elemental segregation in an AlCoCrFeNi high-entropy alloy-A comparison between selective laser melting and induction melting, J. Alloys Compd. **784**, 195 (2019).

[9] Y. Bu, Y. Wu, Z. Lei, X. Yuan, H. Wu, X. Feng, J. Liu, J. Ding, Y. Lu, H. Wang *et al.*, Local chemical fluctuation mediated ductility in body-centered-cubic HEAs, Mater. Today **46**, 28 (2021).

[10] S. R. Spurgeon, C. Ophus, L. Jones, A. Petford-Long, S. V. Kalinin, M. J. Olszta, R. E. Dunin-Borkowski, N. Salmon, K. Hattar, W. D. Yang *et al.*, Towards data-driven next-generation transmission electron microscopy, Nat. Mater. **20**, 274 (2021).

[11] W. Jian, Z. Xie, S. Xu, Y. Su, X. Yao, and I. J. Beyerlein, Effects of lattice distortion and chemical short-range order on the mechanisms of deformation in medium entropy alloy CoCrNi, Acta Mater. **199**, 352 (2020).

[12] W. Sekkal, R. Besson, and A. Legris, Atomic scale modeling of structural phase transformations in AlCrFeMnMo high-entropy alloys during thermal treatments, J. Alloys Compd. **876**, 160201 (2021).

[13] S. Chen, Z. H. Aitken, S. Pattamatta, Z. Wu, Z. Yu, D. J. Srolovitz, P. K. Liaw, and Y. Zhang, Simultaneously enhancing the ultimate strength and ductility of high-entropy alloys via short-range ordering, Nat. Commun. **12**, 1 (2021).

[14] X. Huang, L. Liu, W. Liao, J. Huang, H. Sun, and C. Yu, Characterization of nucleation behavior in temperature-induced BCC-to-HCP phase transformation for high entropy alloy, Acta Metall. Sin-Engl. **34**, 1546 (2021).

[15] E. Antillon, C. Woodward, S. I. Rao, B. Akdim, and T. A. Parthasarathy, Chemical short range order strengthening in a model FCC high entropy alloy, Acta Mater. **190**, 29 (2020).

[16] H. Zheng, L. T. W. Fey, X. Li, Y. Hu, L. Qi, C. Chen, S. Xu, I. J. Beyerlein, and S. P. Ong, Multi-scale investigation of chemical short-range order and dislocation glide in the MoNbTi and TaNbTi refractory multi-principal element alloys, arXiv:2203.03767 (2022).

[17] S. Yin, Y. Zuo, A. Abu-Odeh, H. Zheng, X. Li, J. Ding, S. P. Ong, M. Asta, and R. O. Ritchie, Atomistic simulations of dislocation mobility in refractory high-entropy alloys and the effect of chemical short-range order, Nat. Commun. **12**, 1 (2021).

[18] Q. Li, H. Sheng, and E. Ma, Strengthening in multi-principal element alloys with local-chemical-order roughened dislocation pathways, Nat. Commun. **10**, 1 (2019).

[19] X. Huang, L. Liu, X. Duan, W. Liao, J. Huang, H. Sun, and C. Yu, Atomistic simulation of chemical short-range order in HfNbTaZr high entropy alloy based on a newly-developed interatomic potential, Mater. Des. **202**, 109560 (2021).

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

[21] F. Z. Dai, B. Wen, Y. Sun, H. Xiang, and Y. Zhou, Theoretical prediction on thermal and mechanical properties of high entropy $(Zr_{0.2}Hf_{0.2}Ti_{0.2}Nb_{0.2}Ta_{0.2})C$ by deep learning potential, J. Mater. Sci. Technol. **43**, 168 (2020).

[22] F. Z. Dai, and Y. Sun, B. Wen, H. Xiang, and Y. Zhou, Temperature dependent thermal and elastic properties of high entropy $(Ti_{0.2}Zr_{0.2}Hf_{0.2}Nb_{0.2}Ta_{0.2})B_2$: Molecular dynamics simulation by deep learning potential, J. Mater. Sci. Technol. **72**, 8 (2021).

[23] X. Liu, J. Zhang, M. Eisenbach, and Y. Wang, Machine learning modeling of high entropy alloy: The role of short-range order, arXiv:1906.02889 (2019).

[24] A. van de Walle, P. Tiwary, M. de Jong, D. L. Olmsted, M. Asta, A. Dick, D. Shin, Y. Wang, L. Q. Chen, and Z. K. Liu, Efficient stochastic generation of special quasirandom structures, Calphad **42**, 13 (2013).

[25] X. Duan, Z. Zhang, H. He, Z. Liu, X. Duan, and B. Shan, A systematic investigation on quaternary NbTiZr-based refractory high entropy alloys using empirical parameters and first principles calculations, Model. Simul. Mater. Sci. **29**, 075002 (2021).

[26] C. H. Tu, Y. C. Lai, S. K. Wu, and Y. H. Lin, The effects of annealing on severely cold-rolled equiatomic HfNbTiZr high entropy alloy, Mater. Lett. **303**, 130526 (2021).

[27] Y. D. Wu, Y. H. Cai, T. Wang, J. J. Si, J. Zhu, Y. D. Wang, and X. D. Hui, A refractory $Hf_{25}Nb_{25}Ti_{25}Zr_{25}$ HEA with excellent structural stability and tensile properties, Mater. Lett. **130**, 277 (2014).

[28] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method* (John Wiley & Sons, Hoboken, 2016).

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

[30] J. P. Perdew, K. Burke, and Y. Wang, Generalized gradient approximation for the exchange-correlation hole of a many-electron system, Phys. Rev. B. **54**, 16533 (1996).

[31] H. J. Monkhorst and J. D. Pack, Special points for Brillouin-zone integrations, Phys. Rev. B. **13**, 5188 (1976).

[32] B. E. Warren, *X-ray Diffraction* (Dover publications, Mineola, 1990).

[33] J. M. Cowley, An approximate theory of order in alloys, Phys. Rev. **77**, 669 (1950).

[34] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, J. Comput. Phys. **285**, 316 (2015).

[35] N. Artrith and A. Urban, An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO2, Comput. Mater. Sci. **114**, 135 (2016).

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

[37] H. Wang, L. Zhang, J. Han, and W. E., DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics, Comput. Phys. Commun. **228**, 178 (2018).

[38] L. Zhang, J. Han, H. Wang, W. Saidi, and R. Car, End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems, Adv. Neural Inf. Process. Syst. **31**, 4441 (2018).

[39] J. Han, L. Zhang, and R. Car, Deep potential: A general representation of a many-body potential energy surface, Commun. Comput. Phys. **23**, 629 (2018).

[40] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, J. Comput. Phys. **117**, 1 (1995).

[41] I. A. Balyakin and A. A. Rempel, Machine learning interatomic potential for molten TiZrHfNb, in *The VII International Young Researchers' Conference Physics, Technology, Innovations (PTI-2020)*, edited by V. A. Volkovich, I. V. Kashin, A. A. Smirnov, and E. D. Narkhov, AIP Conf. Proc. No. 2313 (AIP, New York, 2020), p. 030037.