

Chemistry

Special Topic: AI for Chemistry

Constructing machine learning potential for metal nanoparticles of varying sizes via basin-hopping Monte Carlo and active learning

Fu-Qiang Gong¹, Ke Xiong¹ & Jun Cheng^{1,2,3,*}¹State Key Laboratory of Physical Chemistry of Solid Surface, iChEM, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China;²Laboratory of AI for Electrochemistry (AI4EC), Tan Kah Kee Innovation Laboratory (IKKEM), Xiamen 361005, China;³Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China*Corresponding author (email: chengjun@xmu.edu.cn)

Received 23 December 2023; Revised 18 March 2024; Accepted 19 March 2024; Published online 20 March 2024

Abstract: Nanoparticles, distinguished by their unique chemical and physical properties, have emerged as focal points within the realm of materials science. Traditional theoretical approaches for atomic simulations mainly include empirical force field and *ab initio* simulations, with the former offering efficiency but limited reliability, and the latter providing accuracy but restricted to systems of relatively small sizes. Herein, we propose a systematic strategy and automated workflow designed for collecting a diverse types of atomic local environments within a training dataset. This includes small nanoclusters, nanoparticles, as well as surface and bulk systems with periodic boundary conditions. The objective is to construct a machine learning potential tailored for pure metal nanoparticle simulations of varying sizes. Through rigorous validation, we have shown that our trained machine learning potential is capable of effectively driving molecular dynamics simulations of nanoparticles across a wide temperature range, especially within the nanoscale regime. Remarkably, this is achieved while preserving the accuracy typically associated with *ab initio* methods.

Keywords: condensed matter physics, nanoparticles, machine learning potential, workflow

INTRODUCTION

Nanoparticles, composed of a few hundred to millions of atoms, exhibit chemical and physical properties distinct from their bulk counterparts due to the surface effect and quantum size effect [1,2]. The unique properties of nanoparticles make them promising candidates for applications across various research domains, including but not limited to catalysis [3], biology [4], and medicine [5]. Therefore, nanoparticles have been at the forefront of the rapidly advancing field of materials science over the last few decades.

Currently, numerous studies on nanoparticles have been carried out experimentally [6,7] as well as theoretically [8,9]. For experimental studies, significant efforts are directed towards the atomic-level characterization and precise synthesis of nanoparticles [10,11], which aim to design and obtain nanoparticles with enhanced performance tailored for specific applications. As for theoretical investigations, given that the

majority of physical and chemical properties of nanoparticles are linked to their configurations, previous studies have concentrated on unraveling the relationship between the structure and property of nanomaterials using global search methods [12, 13]. Nevertheless, it has been discovered that the structure of nanoparticles is highly flexible [14, 15], so that it is insufficient to study the properties of nanoparticles solely based on the most stable structure. Consequently, extensive efforts have been dedicated to developing molecular dynamics (MD)-based methods aimed at studying the dynamic structures and properties of nanoparticles. This includes the investigation of structural evolution and corresponding properties [16–18], wherein an accurate description of the potential energy surface (PES) of nanoparticles is required.

In the computation of the energy for a given structure, two primary theoretical methodologies are commonly employed: empirical force fields [19] and first-principles methods [20]. Regarding the former approach, despite its efficiency, empirical force field methods are prone to instability and may yield inaccurate results when predicting the energy-structure relationship [21]. On the contrary, first-principles methods, like density functional theory (DFT), provide an accurate approach for calculating properties at an *ab initio* level, particularly suitable for models comprising several dozen to a hundred atoms [22]. However, it is important to note that the computational cost associated with DFT scales cubically with the size of the simulation systems [23]. Machine learning potential (MLP) methods have emerged as a promising solution to this challenge, enabling simulations at an *ab initio* level [24, 25]. For instance, Chen *et al.* [26] proposed a universal machine learning framework based on graph convolutional neural networks, successfully predicting and screening various kinds of high-performance alloy electrocatalysts, which were experimentally verified, demonstrating the promising applications of MLPs in the field of materials research. As for cluster studies, several studies [27, 28] have employed MLP for structural searches to identify the global minima of clusters. These studies demonstrate outstanding performance when compared to results obtained through DFT. However, it is noteworthy that these studies have predominantly concentrated on the ground states of clusters, with the systems under investigation still limited to the range of several dozen atoms—a scale affordable to DFT calculations. As mentioned previously, there is a growing interest in the dynamic nature of clusters. When considering nanoparticles, which are most commonly employed in real-world applications [29], the number of atoms can be easily over thousands, exceeding the practical limits of *ab initio* simulations. Thus, a key challenge arises: how to construct an MLP for the simulations of nanoparticles with varying sizes, particularly in the nanometer scale. This MLP should not only be applicable to local minima of the PES but also accurately predict all kinds of isomers at finite temperature.

In this study, we have combined the basin-hopping Monte Carlo algorithm [30], deep learning methods [31], MD simulations, and DFT calculations. This integration forms an active learning workflow designed to construct a single MLP that is universally applicable to *ab initio* molecular dynamics (AIMD) simulations of nanoparticles, irrespective of their sizes. We consider a wide range of cluster types, including nanoclusters, nanoparticles, surface models with low Miller indices, and bulk models with different stacking patterns. Based on a systematic and automatic data collection, we have successfully constructed an MLP capable of simulating nanoparticles of varying sizes. Rigorous validation procedures further guarantee that our MLP yields accuracy comparable to that of *ab initio* calculations.

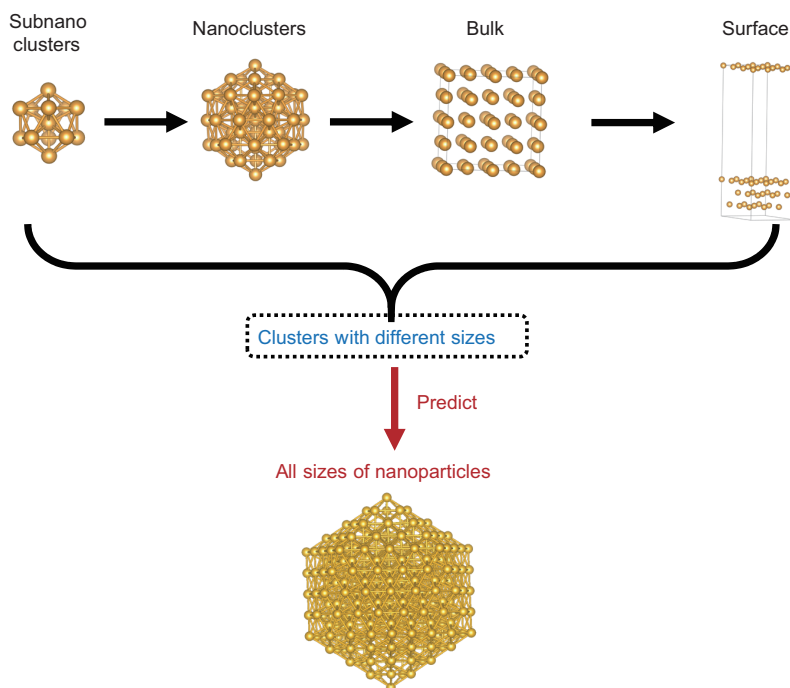


Figure 1 Schematic plot of the strategy to construct the machine learning potential for nanoparticles with varying sizes.

MATERIALS AND METHODS

Strategy

To construct an MLP for metal nanoparticles, it is essential to include diverse atomic local environments in the training dataset. This is due to the intrinsic nature of the machine learning training process, which involves the fitting of atomic local environments to atomic energy [32]. When the cluster size gets down to the subnanometer regime, all atoms can be regarded as surface atoms with low coordination numbers. Conversely, as the cluster size increases into the nanometer scale, the atoms in nanoparticles can be decomposed into two main categories, namely bulk atoms and surface atoms with relatively higher coordination numbers compared to the small clusters. To ensure the applicability of the trained MLP to nanoparticles of varying sizes, the training dataset should contain the atomic local environment ranging from the minimally coordinated limit to the maximally coordinated limit. However, because of the cubic scaling relationship between the computational cost of first-principle calculations and the sizes of the simulated particles [23], labeling clusters with sizes up to 100 atoms on an *ab initio* level presents an exceedingly challenging task. Due to the excellent transferability ensured by the relatively large dataset containing as many local structures as possible, MLP exhibit excellent performance for the accurate prediction of interpolated data points, even when they are absent from the training dataset [21]. Therefore, a natural and straightforward approach for the construction of an MLP for metal nanoparticles involves the collection of labeled data of various sizes of (sub-)nanoclusters, nanoparticles, surface and bulk models with periodic boundary conditions, as depicted in Figure 1.

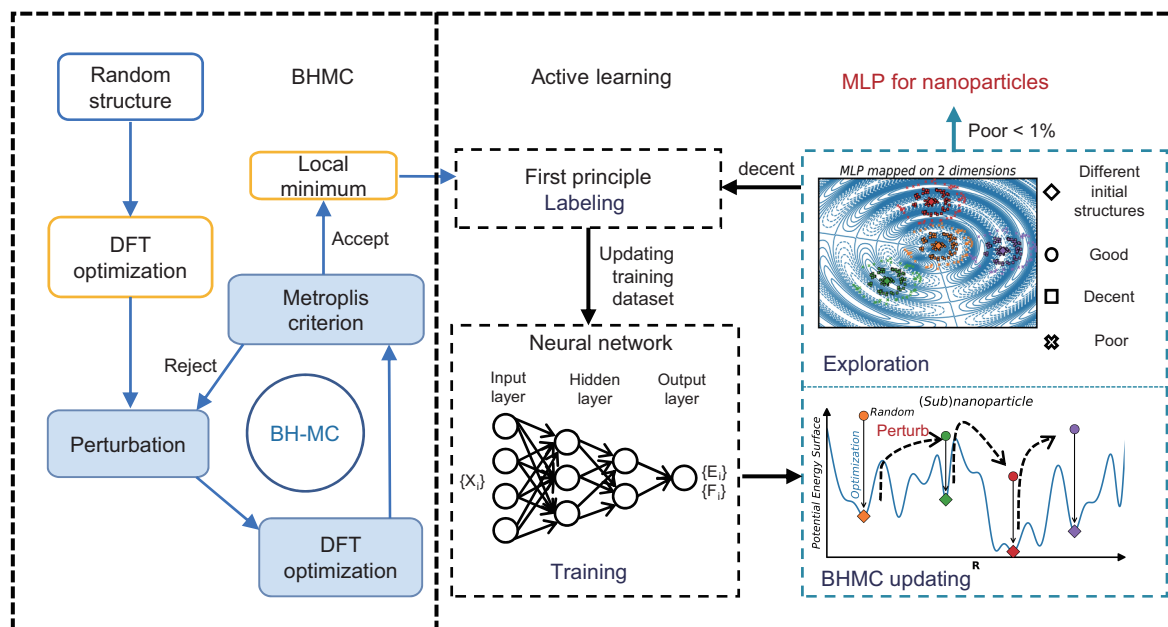


Figure 2 The diagram of the automated workflow framework, which has two main components, namely, the basin hopping Monte Carlo (BHMC) and active learning.

Workflow

The key to constructing an MLP lies in the collection of a high-quality dataset. In this context, we have designed an automatic workflow aimed at the iterative generation of a training dataset with the accuracy comparable to DFT for metal nanoclusters and nanoparticles. This workflow contains two main modules, namely, the basin hopping Monte Carlo (BHMC) and active learning, as illustrated in Figure 2.

Basin hopping Monte Carlo (BHMC)

The initialization of the workflow aims to collect a small amount of labeled dataset to start the active learning module. To ensure effective exploration of the configuration space for clusters, numerous minimum structures, which show very different atomic arrangements shown in the snapshots of Figure 3, are systematically generated via BHMC algorithm. This strategy ensures that the exploration within the active learning module is initiated from various local minima on the PES. The main procedure is shown in the following.

(1) The BHMC starts with a random structure, which is generated through the atomic simulation environment (ASE) package [33]. In this step, the atomic coordinates of the cluster are randomly assigned, subject to the constraint that the distance between any two atoms within the cluster falls within specified boundaries, thereby preventing excessively short or long interatomic distances.

(2) The geometry optimization method is applied on the initially generated random structure via DFT calculations. Subsequently, the energy and atomic coordinates of the optimized structure are recorded as E_i and X_i , respectively.

(3) The previously optimized structure is subjected to perturbation by adding random numerical values to the atomic coordinates.

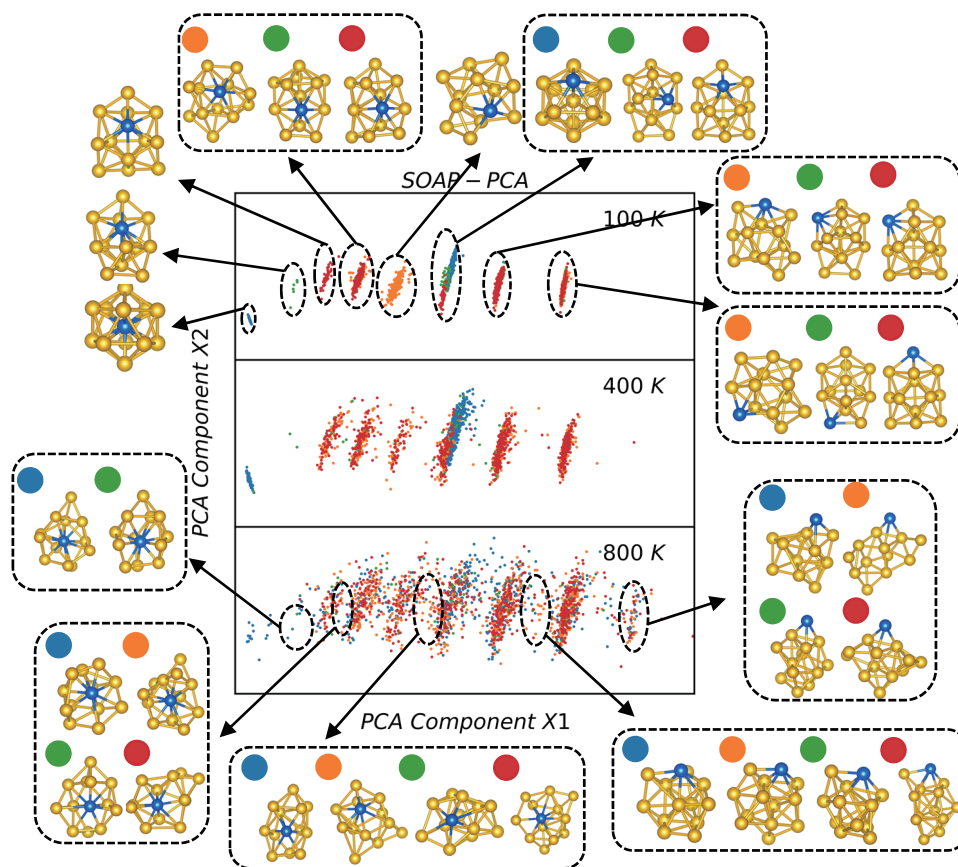


Figure 3 Unsupervised machine learning of the structures sampled from Cu_{13} clusters at different temperatures with different initial configurations (upper panel: 100 K; middle panel: 400 K; bottom panel: 800 K), where different colors denote different initial structures. A structural similarity map of local environments is obtained by PCA. The blue and yellow balls in the snapshots indicate the copper atoms that are used to calculate the SOAP descriptors and other copper atoms, respectively.

(4) The geometry optimization method is applied to the perturbed structure using DFT calculations. The resulting energy and coordinates of the optimized structure are saved as E_j and X_j , respectively.

(5) The metropolis criterion is employed to determine whether the perturbed structure is accepted ($\exp^{-(E_j-E_i)/k_B T} > \text{random}(0, 1)$) or rejected.

According to the criterion in step (5), if the energy of the new structure (E_j) is lower than that of the old one (E_i), the new structure is saved. Even if the energy is higher, there is still a possibility of accepting the new structure. This approach leads to the generation of a canonical ensemble sampling on the PES, with a specific focus on the energetically lowest-lying isomers [34]. It should be noted that the trajectories of geometry optimizations will be saved as the initial training dataset.

Active learning

The objective of the active learning module is to progressively expand the training dataset until the constructed MLP demonstrates predictive capabilities for the studied nanoparticles. To achieve this, the concurrent learning workflow, initially developed by Zhang *et al.* [35], is employed and adapted for the expansion of the training dataset, which has found widespread application across diverse system types [36, 37]. The

main components of the active learning module include four steps: training, BHMC updating, exploration, and labeling.

(1) Training: In this step, all the DFT calculated structures, along with their corresponding energies and atomic forces, are collected as a training dataset to train four MLPs. Here, the deep potential smooth edition (DP-SE) model [38] implemented in the DeepMD-kit package [31], is employed, which decomposes the total energy (E_{tot}) of a structure into atomic contributions (E_i) ($E_{\text{tot}} = \sum_i E_i$). The atomic forces (F_i) are derived through the gradient of the total energy with respect to atomic coordinates (R_i) ($F_i = \nabla_{R_i} E_{\text{tot}}$). To map a structure to its corresponding energy, DP-SE employs two sets of neural networks, specifically, the embedding network and fitting network. The former serves to map the Cartesian coordinates of the structure to atomic local environments, commonly referred to as descriptors, thereby preserving translational, rotational, and permutational symmetry. Meanwhile, the latter is responsible for fitting the relationship between atomic energy and the associated descriptors. It is worth emphasizing that four MLPs are trained using the same training dataset but with different random seeds. This approach ensures that all the MLPs demonstrate effective predictive performance for the training dataset; however, variations may emerge when it is applied to new structures.

(2) BHMC updating: In this step, the initial structures for exploration will be updated through BHMC algorithm, allowing the sampled configuration space to be further extended. Notably, instead of using DFT calculations for geometry optimization, the trained MLPs can be applied to accelerate this process.

(3) Exploration: In this step, structures that have been saved from BHMC simulations serve as initial points to sample the configuration space. Among the four MLPs, one is employed to drive the MD simulations, while the remaining three are utilized for the computation of energies and atomic forces for the structures sampled during the MD simulations. Due to the limited training dataset in the initial iterations, the MLPs face challenges in providing precise predictions for all the structures obtained from MD simulations. To assess the MLPs' performance on the sampled structures, the maximum deviation in atomic forces (f_{max}) among the model ensembles is employed as a performance metric:

$$f_{\text{max}} = \max_i \sqrt{\langle |f_{w,i}(R_i) - \langle f_{w,i}(R_i) \rangle|^2 \rangle}, \quad (1)$$

where $f_{w,i}$ indicates the predicted force on i -th atom by the MLP with parameter w , and $\langle \dots \rangle$ represents the average value predicted by the four MLPs. Here, the user-defined parameters, denoted as $f_{\text{max,low}}$ and $f_{\text{max,high}}$, are used to categorize structures into three classifications, namely good, decent, and poor. Structures for which f_{max} is less than $f_{\text{max,low}}$ (good) are considered accurately predicted by the MLPs. Conversely, structures with f_{max} exceeding $f_{\text{max,high}}$ (poor) may exhibit significant deviations from the PES, possibly rendering them physically unrealistic. Structures with the corresponding f_{max} falling within the range between $f_{\text{max,low}}$ and $f_{\text{max,high}}$ (decent) are subject to further labeling in subsequent steps.

(4) Labeling: At this stage, structures categorized as “decent” following the exploration step will undergo DFT calculations. The resulting energy values and atomic forces, along with their associated configurations, will be added to the training dataset, thereby contributing to the enhancement of the performance of the MLPs.

Based on such a workflow, the studied systems gradually include nanoclusters, nanoparticles, as well as surface and bulk models. When larger particles are assigned as the initial structures for exploration, it is observed that the maximum force deviations across all sampled configurations fall below the threshold ($f_{\text{max,low}}$),

as shown in Figures S1–S3. This marks the successful construction of an MLP for metal (sub)nanoparticles.

Computational detail

Training setup

During the training step, the DP-SE method [38], implemented in DeepMD-kit [31], is used to train MLPs, with the cutoff and smooth cutoff of the local environment set to 8.0 and 0.5, respectively. The size of the embedding neural network and fitting neural network is set to {25,50,100} and {240,240,240}, respectively. In each iteration, the MLP is trained for 400,000 gradient steps with an exponentially decaying learning rate spanning from 0.001 to 3.5×10^{-8} .

MD setup

In the exploration step, several MD simulations are conducted to explore the configuration space of systems utilizing the LAMMPS simulation software [39]. In the case of cluster systems (M_x , $x=10-55$), these simulations are carried out in the canonical ensemble (NVT) ensemble, covering the temperatures ranging from 50 to 1500 K (50, 100, 200, 300, 500, 700, 900, 1200, and 1500 K). For bulk systems (face-centered cubic (FCC), body-centered cubic (BCC), and hexagonal close packed (HCP) crystals), the simulations are conducted in the isothermal-isobaric ensemble (NPT) ensemble, with temperature values spanning from 50 to 3000 K (50, 100, 200, 300, 500, 700, 900, 1200, 1500, 2000, 2500, and 3000 K) and pressures ranging from 1 to 50,000 atm (1 atm= 1.01325×10^5 Pa) (1, 10, 100, 1000, 5000, 10,000, 20,000, and 50,000 atm). Meanwhile, for surface systems (100, 110, 111 surface of FCC, BCC, and HCP crystals), the simulations are performed in the NVT ensemble, with the temperature range from 50 to 3000 K (50, 100, 200, 300, 500, 700, 900, 1200, 1500, 2000, 2500, and 3000 K).

DFT setup

The aforementioned DFT calculations are conducted using the CP2K/QUICKSTEP package [40,41], employing the Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional [42] along with the Grimme D3 dispersion correction [43] to describe interatomic interactions. The core electrons are described by Goedecker-Teter-Hutter (GTH) type pseudopotentials [44], while valence electrons are expanded using a Gaussian-type triple basis with two sets of polarization functions (TZV2P) [45]. For each structure, self-consistent calculations are terminated when the charge density residuals are less than 1.0×10^{-6} atomic units.

RESULTS AND DISCUSSION

By combining the BHMC algorithm with an active learning workflow, we have successfully constructed MLPs for several kinds of pure metal (sub)nanoparticles, including Au, Ag, Cu, Pt, Pd, Ni, Ru, Rh, Si, and Al. These MLPs can be utilized to conduct simulations of nanoparticles of different sizes across a wide

temperature range with *ab initio* accuracy. Subsequently, we will utilize the copper potential as an illustrative example to present our results.

To begin, we will elucidate the rationale behind the utilization of the BHMC algorithm as the initialization method for the active learning workflow. This strategy can be attributed to two primary considerations. Firstly, in the exploration step of each iteration, initial structures for MD simulations need to be assigned. In the case of bulk and surface systems, a variety of methods are available for acquiring the corresponding structures, including databases like the Materials Project [46] or through experimental characterization. While for cluster systems, databases containing relevant structures are relatively scarce, with the majority of available data being derived through geometry optimization employing empirical force field calculations, such as The Cambridge Cluster Database [47]. Given the relatively small size of cluster systems, acquiring corresponding structures through current characterization techniques remains a challenging task. Therefore, we employed the BHMC algorithm combined with DFT calculations to obtain the cluster structures. Secondly, the structural configurations derived through this approach may exhibit significant diversity, suggesting that in the exploration step, sampling can be initiated from different points of the PES. In Figure 3, we use a Cu_{13} cluster as an example to illustrate the aforementioned demonstration. Here, we have used the smooth-overlap-of-atomic-positions (SOAP) descriptor [48] to represent the local environment. Subsequently, by utilizing the unsupervised machine learning method, specifically principal component analysis (PCA), we have projected the cluster's structural features onto two dimensions, enabling the direct observation of structural similarities. As revealed by the SOAP-PCA analysis, the local environments of the four Cu_{13} structures generated through the BHMC algorithm exhibit notable dissimilarities, resulting in their clear categorization into several distinct regions. However, it should be noted that for different initial structures, points of different colors may overlap. This is because we are reducing the dimensionality based on the atomic local environments rather than the entire structure. As shown in the dashed circles in Figure 3, atomic local environments exhibit similarity across different structures. Notably, at lower temperatures, the blue balls represent the clustered atomic local environments. Moving from left to right, the first dashed circle reveals copper atoms located at the centroid of a hexagon, while the copper atoms of the second one are at the centroid of a pentagon. Furthermore, the third one indicates copper atoms situated at an edge site, coordinated with five surrounding Cu atoms, whereas the fourth one denotes copper atoms located at another edge site, coordinated with four neighboring Cu atoms. It is evident that, by employing diverse initial structures to explore the PES, we have further included various kinds of atomic local environments into our training dataset. Moreover, as the temperature is raised, the atomic local environments among these regions become increasingly interconnected. As shown in the snapshots under elevated temperature conditions in Figure 3, we observe structural distortions in the atomic local environments. For instance, in the third dashed circle at a higher temperature (800 K), the points lie between those in the fourth and sixth circles observed at a lower temperature (100 K), which correspond to the centroids of the hexagon and pentagon, respectively. Notably, the copper atom within these points is located at the center of a distorted hexagon. Therefore, by initiating simulations with different structures and under different temperature conditions, our exploration can cover different regions of the PES, thus enhancing the efficiency of the exploration.

By collecting a training dataset comprising 26,754 frames of cluster systems, along with 391 frames of bulk and 7297 frames of surface systems, we have constructed an MLP for nanoparticles of varying sizes. To ensure the accuracy of the constructed MLPs with DFT results, a comparative analysis of energies and

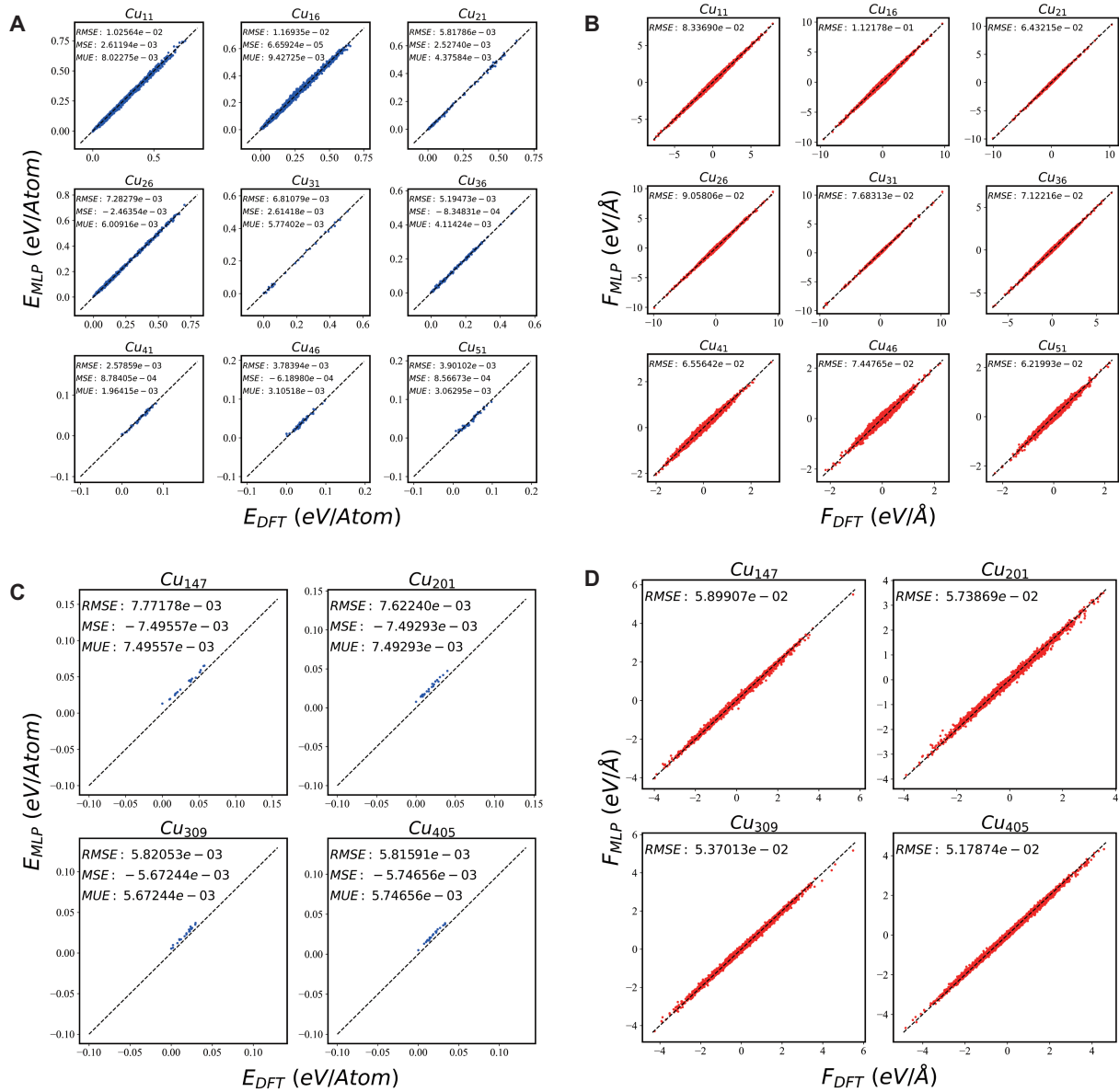


Figure 4 (A) Comparison of the DFT energies (E_{DFT}) and the MLP energies (E_{MLP}) of the different cluster systems in the training dataset. (B) Comparison of the DFT forces (F_{DFT}) and the MLP forces (F_{MLP}) of the different cluster systems in the training dataset. (C) Comparison of E_{DFT} and E_{MLP} of the Cu₁₄₇, Cu₂₀₁, Cu₃₀₉, Cu₄₀₅ sampled from MD simulations. (D) Comparison of F_{DFT} and F_{MLP} of the Cu₁₄₇, Cu₂₀₁, Cu₃₀₉, Cu₄₀₅ sampled from MD simulations. RMSE, mean-signed-error (MSE), and mean-unsigned-error (MUE) are shown in the figure.

forces computed by both methods has been conducted. The results of this comparison are presented in Figure 4A and B and detail values are provided in Table S1. The root mean squared errors (RMSEs) of energies approximate the order of 10^{-3} eV/atom, and the RMSEs for forces are around 10^{-2} eV/Å across all the trained systems, suggesting that the constructed MLP can accurately predict the energies and atomic forces of the corresponding structures within the training dataset. Additionally, we further tested the performance of the trained MLP at different temperatures, as illustrated in Figure S4. The simulations at various temperature conditions achieved first-principles accuracy, even under extreme temperatures such as 5000 K, which were

not considered during the exploration. Despite this, the trained MLP still demonstrated high accuracy. For those larger nanoparticles that are completely absent from our dataset, the transferability of the MLP has also been evaluated, specifically Cu₁₄₇, Cu₂₀₁, Cu₃₀₉, and Cu₄₀₅, as illustrated in Figure 4C and D and detail values are provided in Table S2. The RMSEs of energies are observed to be around 6×10^{-3} eV/atom, and the RMSEs for forces are around 5×10^{-2} eV/Å for all the validation systems, which are not included in our training dataset. Therefore, the constructed MLP can be employed to conduct the simulation of nanoparticles with *ab initio* accuracy. However, it should be noted that the energies predicted by the constructed MLPs exhibit a small but constant bias compared with the energies calculated by DFT. This discrepancy may originate from the nature of energy prediction in MLP, where the predicted energies are derived by summing the local atomic energies to determine the overall structural energy. When such a model is absent from the training dataset, the neural network cannot effectively learn to account for it, resulting in a constant bias in energy predictions. Additionally, due to the workflow where the maximum force deviation is used as a performance metric of MLP, the weights of forces are higher during the training process, resulting in less accurate predictions of absolute energies. However, this does not affect the accuracy of MLP, as the relative energy and forces predicted for the structures remain accurate. This impact can be further decreased through increasing the weights of energies during the training process, potential refinement or fine-tuning tailored to specific systems [49, 50]. Fortunately, in most cases, it is the relative energies we care about, rather than the absolute energy values. For a specific nanoparticle system, such a constant bias can be effectively canceled out when calculating the relative energies. More importantly, the atomic forces, which are the first-order derivatives of energy with respect to atomic coordinates, as predicted by the MLP, are almost aligned with the corresponding DFT results, thus the shape of the PES predicted by MLP remains in good agreement with that derived from DFT calculations.

The proposed methodology, wherein a dataset including small clusters, surface, and bulk systems is collected to train an MLP suitable for nanoparticle simulations of varying sizes, appears to yield successful outcomes. However, a more pivotal inquiry is to the underlying rationale behind its effectiveness. In the case of large nanoparticles, all atoms can be divided into two main classes, namely core and surface atoms. Core atoms, characterized by coordinative saturation (coordination number, CN=12), can be effectively represented through the construction of bulk systems with periodic boundary conditions. Conversely, surface atoms exhibit a diverse local environment, including atoms located at low miller index surfaces, as well as corner and edge atoms, among others. These different kinds of surface atoms are associated with varying CNs. Therefore, we further conduct a comparative analysis of the probability density of CNs (P(CN)) (see method in Supplementary information) for all structures within the training dataset and those derived from MD trajectories of nanoparticles, as shown in Figure 5A. The analysis reveals that atoms with lower CNs within the training dataset predominantly originate from cluster models, whereas atoms with higher CNs are primarily sourced from bulk and surface models. Notably, the P(CN) for nanoparticle systems exhibits two distinctive peaks, both of which are completely covered in our training dataset. Additionally, we have employed unsupervised machine learning techniques to evaluate the degree of similarity in atomic local environments between structures within the training dataset and those sampled from MD simulations of nanoparticles, as depicted in Figure 5B. A similar analysis has been conducted for Cu₁₈₆₅, Cu₂₅₃₁, and Cu₃₅₃₅, as shown in Figures S1–S3, to demonstrate the applicability of our MLP for larger nanoparticles. It is evident that atomic local environments within the training dataset cover those in MD trajectories completely,

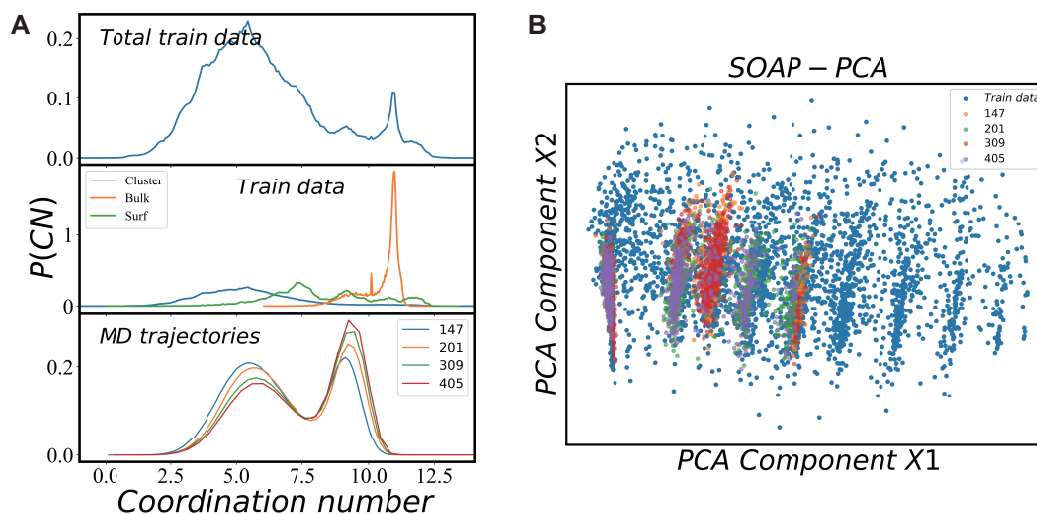


Figure 5 (A) The P(CN) of all the training dataset (upper panel), the corresponding system types (middle panel), and validation systems (bottom panel). (B) Unsupervised machine learning of the structures in the training dataset and the structures sampled from MD simulations of Cu₁₄₇, Cu₂₀₁, Cu₃₀₉, and Cu₄₀₅. A structural similarity map of local environments is obtained by PCA.

indicating the local environments of nanoparticles have already been included in the training dataset, thereby enabling the trained MLP to be transferable for clusters that are not subjected to exploration.

To demonstrate the effectiveness of the trained MLP in simulating certain physical phenomena, the melting curve of Cu₁₁₀₃ is calculated utilizing the trained MLP, as depicted in Figure S5. It reveals that the total energy of such a nanoparticle shows linear change with temperature in both low- and high-temperature regions, while exhibiting a pronounced increase within the intermediate temperature range, indicating quasi-first-order phase transition [51]. Furthermore, we conducted additional analysis of atomic dynamics employing the Lindemann index analysis [52]. Our findings illustrate that with an increasing temperature, surface atoms tend to melt at lower temperatures, a phenomenon widely observed in numerous studies, including experimental investigations [53–56].

CONCLUSIONS

As the application of machine learning in materials research becomes increasingly widespread, particularly with the introduction of large-scale models [57], the collection of diverse and high-quality datasets gains growing significance. When it comes to nanomaterials, overcoming the limitations of traditional first-principles computational simulations, where the number of simulated atoms often does not exceed one hundred, and the simulation time is typically limited to one hundred picoseconds, is an inevitable trend for future theoretical investigations. In this context, we have taken the advantages of MLPs for the fitting of atomic local environments. By collecting a dataset with various kinds of atomic local environments, we have successfully constructed a potential applicable to simulations of metal nanoparticles of different sizes, all while ensuring *ab initio* accuracy. We believe that the strategy we have proposed is not only applicable to the development of potentials for pure metal nanoparticles but also holds promise for more complex systems, such as alloys, oxides, and catalytic systems, which are also the focus for our future research endeavors.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements

F.-Q.G. gratefully acknowledges Xiamen University and iChEM for a Ph.D. studentship, and the help from DP modeling community.

Funding

This work was supported by the National Science Fund for Distinguished Young Scholars (22225302), the National Natural Science Foundation of China (92161113, 21991151, 21991150 and 22021001), the Fundamental Research Funds for the Central Universities (20720220008, 20720220009 and 20720220010), the Laboratory of AI for Electrochemistry (AI4EC), and IKKEM (RD2023100101 and RD2022070501).

Author contributions

J.C. designed and supervised the project. F.-Q.G. carried out the calculation. F.-Q.G. and J.C. analyzed the results. All authors discussed the results and contributed to manuscript preparation.

Conflict of interest

The authors declare no conflict of interest.

Supplementary information

The supporting information is available online at <https://doi.org/10.1360/nso/20230088>. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

- 1 Jena P, Sun Q. Super atomic clusters: Design rules and potential for building blocks of materials. *Chem Rev* 2018; **118**: 5755–5870.
- 2 Nair AS, Pathak B. Computational strategies to address the catalytic activity of nanoclusters. *WIREs Comput Mol Sci* 2021; **11**: e1508.
- 3 Heiz U, Landman U. *Nanocatalysis*. Berlin, Heidelberg: Springer, 2007.
- 4 Saptarshi SR, Duschl A, Lopata AL. Interaction of nanoparticles with proteins: Relation to bio-reactivity of the nanoparticle. *J Nanobiotechnol* 2013; **11**: 26.
- 5 Salata OV. Applications of nanoparticles in biology and medicine. *J Nanobiotechnol* 2004; **2**: 3.
- 6 Zhuang ZH, Chen W. Application of atomically precise metal nanoclusters in electrocatalysis. *J Electrochem*, 2021; **27**: 125.
- 7 Fichthorn KA, Yan T. Shapes and shape transformations of solution-phase metal particles in the sub-nanometer to nanometer size range: Progress and challenges. *J Phys Chem C* 2021; **125**: 3668–3679.
- 8 Coquet R, Howard KL, Willock DJ. Theory and simulation in heterogeneous gold catalysis. *Chem Soc Rev* 2008; **37**: 2046–2076.
- 9 Zhang J, Glezakou V. Global optimization of chemical cluster structures: Methods, applications, and challenges. *Int J Quantum Chem* 2021; **121**: e26553.
- 10 Shnoudeh AJ, Hamad I, Abdo RW, *et al.* Synthesis, characterization, and applications of metal nanoparticles. *Biomater Bionanotechnol* 2019: 527–612.

- 11 Wei JQ, Chen XD, Li SZ. Electrochemical syntheses of nanomaterials and small molecules for electrolytic hydrogen production. *J Electrochem* 2022; **28**: 2214012.
- 12 Doye JPK, Wales DJ. Global minima for transition metal clusters described by Sutton-Chen potentials. *New J Chem* 1998; **22**: 733–744.
- 13 Wei GF, Liu ZP. Subnano Pt particles from a first-principles stochastic surface walking global search. *J Chem Theor Comput* 2016; **12**: 4698–4706.
- 14 Vargas A, Santarossa G, Iannuzzi M, *et al.* Fluxionality of gold nanoparticles investigated by Born-Oppenheimer molecular dynamics. *Phys Rev B* 2009; **80**: 195421.
- 15 Avanesian T, Dai S, Kale MJ, *et al.* Quantitative and atomic-scale view of CO-induced Pt nanoparticle surface reconstruction at saturation coverage via DFT calculations coupled with *in situ* TEM and IR. *J Am Chem Soc* 2017; **139**: 4551–4558.
- 16 Pavan L, Rossi K, Baletto F. Metallic nanoparticles meet metadynamics. *J Chem Phys* 2015; **143**: 184304.
- 17 Sun JJ, Cheng J. Solid-to-liquid phase transitions of sub-nanometer clusters enhance chemical transformation. *Nat Commun* 2019; **10**: 5400.
- 18 Gong FQ, Guo YX, Fan QY, *et al.* Dynamic catalysis of sub-nanometer metal clusters in oxygen dissociation. *Next Nanotechnol* 2023; **1**: 100002.
- 19 Lloyd LD, Johnston RL. Theoretical analysis of 17–19-atom metal clusters using many-body potentials. *J Chem Soc Dalton Trans* 2000; 307–316.
- 20 Lee MS, Chacko S, Kanhere DG. First-principles investigation of finite-temperature behavior in small sodium clusters. *J Chem Phys* 2005; **123**: 164310.
- 21 Jiang W, Zhang Y, Zhang L, *et al.* Accurate deep potential model for the Al-Cu-Mg alloy in the full concentration space. *Chin Phys B* 2021; **30**: 050706.
- 22 Guedes-Sobrinho D, Wang W, Hamilton IP, *et al.* (Meta-)stability and core-shell dynamics of gold nanoclusters at finite temperature. *J Phys Chem Lett* 2019; **10**: 685–692.
- 23 Fonseca Guerra C, Snijders JG, te Velde G, *et al.* Towards an order-N DFT method. *Theor Chem Accounts* 1998; **99**: 391–403.
- 24 Behler J. Perspective: Machine learning potentials for atomistic simulations. *J Chem Phys* 2016; **145**: 170901.
- 25 Behler J, Csányi G. Machine learning potentials for extended systems: A perspective. *Eur Phys J B* 2021; **94**: 142.
- 26 Chen L, Tian Y, Hu X, *et al.* A universal machine learning framework for electrocatalyst innovation: A case study of discovering alloys for hydrogen evolution reaction. *Adv Funct Mater* 2022; **32**: 2208418.
- 27 Tuo P, Ye XB, Pan BC. A machine learning based deep potential for seeking the low-lying candidates of Al clusters. *J Chem Phys* 2020; **152**: 114105.
- 28 Wang X, Wang H, Luo Q, *et al.* Structural and electrocatalytic properties of copper clusters: A study via deep learning and first principles. *J Chem Phys* 2022; **157**: 074304.
- 29 Stark WJ, Stoessel PR, Wohlleben W, *et al.* Industrial applications of nanoparticles. *Chem Soc Rev* 2015; **44**: 5793–5805.
- 30 Paz-Borbón LO, López-Martínez A, Garzón IL, *et al.* 2D-3D structural transition in sub-nanometer Pt_N clusters supported on CeO₂ (111). *Phys Chem Chem Phys* 2017; **19**: 17845–17855.
- 31 Zeng J, Zhang D, Lu D, *et al.* DeePMD-kit v2: A software package for deep potential models. *J Chem Phys* 2023; **159**: 054801.
- 32 Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys Rev Lett* 2007; **98**: 146401.
- 33 Hjorth LA, Jørgen MJ, Blomqvist J, *et al.* The atomic simulation environment — A Python library for working with atoms. *J Phys-Condens Matter* 2017; **29**: 273002.
- 34 Gehrke R, Reuter K. Assessing the efficiency of first-principles basin-hopping sampling. *Phys Rev B* 2009; **79**: 085412.

- 35 Zhang Y, Wang H, Chen W, *et al.* DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput Phys Commun* 2020; **253**: 107206.
- 36 Huang J, Zhang L, Wang H, *et al.* Deep potential generation scheme and simulation protocol for the $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ -type superionic conductors. *J Chem Phys* 2021; **154**: 094703.
- 37 Zhuang YB, Bi RH, Cheng J. Resolving the odd-even oscillation of water dissociation at rutile $\text{TiO}_2(110)$ -water interface by machine learning accelerated molecular dynamics. *J Chem Phys* 2022; **157**: 164701.
- 38 Zhang L, Han J, Wang H, *et al.* End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. In: *Proceedings of the 32nd Conference on Neural Information Processing Systems*. Montreal, 2018.
- 39 Thompson AP, Aktulga HM, Berger R, *et al.* LAMMPS-A flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput Phys Commun* 2022; **271**: 108171.
- 40 Hutter J, Iannuzzi M, Schiffmann F, *et al.* CP2K: Atomistic simulations of condensed matter systems. *WIREs Comput Mol Sci* 2014; **4**: 15–25.
- 41 Vande VJ, Krack M, Mohamed F, *et al.* Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comput Phys Commun* 2005; **167**: 103–128.
- 42 Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. *Phys Rev Lett* 1996; **77**: 3865–3868.
- 43 Grimme S, Antony J, Ehrlich S, *et al.* A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* 2010; **132**: 154104.
- 44 Hartwigsen C, Goedecker S, Hutter J. Relativistic separable dual-space Gaussian pseudopotentials from H to Rn. *Phys Rev B* 1998; **58**: 3641–3662.
- 45 Vande VJ, Hutter J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *J Chem Phys* 2007; **127**: 114105.
- 46 Jain A, Ong SP, Hautier G, *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater* 2013; **1**: 011002.
- 47 Wales DJ, Doye JPK, Dullweber A, *et al.* The cambridge cluster database. <https://www-wales.ch.cam.ac.uk/CCD.html>.
- 48 Bartók AP, Kondor R, Csányi G. On representing chemical environments. *Phys Rev B* 2013; **87**: 184115.
- 49 Zhang L, Wang H, Car R, *et al.* Phase diagram of a deep potential water model. *Phys Rev Lett* 2021; **126**: 236001.
- 50 Pinheiro M, Ge F, Ferré N, *et al.* Choosing the right molecular machine learning potential. *Chem Sci* 2021; **12**: 14396–14413.
- 51 Valsson O, Parrinello M. Thermodynamical description of a quasi-first-order phase transition from the well-tempered ensemble. *J Chem Theor Comput* 2013; **9**: 5267–5276.
- 52 Hansen K. *Statistical Physics of Nanoparticles in the Gas Phase*. Vol. 2. Cham: Springer, 2013.
- 53 Zhao SJ, Wang SQ, Cheng DY, *et al.* Three distinctive melting mechanisms in isolated nanoparticles. *J Phys Chem B* 2001; **105**: 12857–12860.
- 54 Schmidt M, Haberland H. Phase transitions in clusters. *Comptes Rendus Physique* 2002; **3**: 327–340.
- 55 Zhang X, Li B, Liu HX, *et al.* Atomic simulation of melting and surface segregation of ternary Fe-Ni-Cr nanoparticles. *Appl Surf Sci* 2019; **465**: 871–879.
- 56 Foster DM, Pavludis T, Kioseoglou J, *et al.* Atomic-resolution imaging of surface and core melting in individual size-selected Au nanoclusters on carbon. *Nat Commun* 2019; **10**: 2583.
- 57 Zhang D, Bi H, Dai FZ, *et al.* DPA-1: Pretraining of attention-based deep potential model for molecular simulation. arXiv: 2208.08236.