Ab Initio Molecular Dynamics Simulation of Zinc metalloproteins with Enhanced Self-Organizing Incremental High Dimensional Neural Network

Mingyuan Xu[1], Tong Zhu[1,2]* and John Z. H. Zhang[1,2,3,4]*

State Key Lab of Precision Spectroscopy, Shanghai Engineering Research Center of Molecular Therapeutics & New Drug Development, Shanghai Key Laboratory of Green Chemistry & Chemical Process, School of Chemistry and Molecular Engineering, East China Normal University, Shanghai 200062, China

[2]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China

[3]Department of Chemistry, New York University, NY, NY 10003, USA

[4]Collaborative Innovation Center of Extreme Optics, Shanxi University, Taiyuan, Shanxi 030006, China

*Correspondence should be addressed to: tzhu@lps.ecnu.edu.cn or john.zhang@nyu.edu

Abstract

Artificial neural network provides the possibility to develop molecular potentials with both the efficiency of the classical molecular mechanics and the accuracy of the quantum chemical methods. In this work, we developed ab initio based neural network potential (NN/MM-RESP-MBG) to perform molecular dynamics study for metalloproteins. The interaction energy, atomic forces, and atomic charges of metal binding group in NN/MM-RESP-MBG are described by a neural network potential trained with energies and forces generated from density functional calculations. Here, we used our recently proposed E-SOI-HDNN model to achieve the automatic construction of reference dataset of metalloproteins and the active learning of neural network potential functions. The predicted energies and atomic forces from the NN potential show excellent agreement with the quantum chemistry calculations. Using this approach, we can perform long time AIMD simulations and structure refinement MD simulation for metalloproteins. In 1 ns AIMD simulation of four common coordination mode of zinc-containing metalloproteins, the statistical average structure is in good agreement with statistic value of PDB Bank database. The neural network approach used in this study can be applied to construct potentials to metalloproteinase catalysis, ligand binding and other important biochemical processes and its salient features can shed light on the development of more accurate molecular potentials for metal ions in other biomacromolecule system.

Introduction

Zinc ions play an important role in enzyme catalysis, signal transduction and the structural stability of proteins. There are more and more evidences that zinc-containing metalloproteins are associated with many human diseases, such as cancer, rheumatism and dementia. The study of theoretical modeling methods for zinc-containing metalloproteins is of significance for the development of related disease target drugs through computer-aided drug design.

For known Zn^{2+} containing metalloproteins, zinc binding site is mostly located at the interface between the protein and the cell fluid. Since the outermost electron layer of zinc ions is full-shell, the coordination mode of zinc ions is very flexible. In aqueous solutions, Zn^{2+} and water molecules form an octahedral six-coordinated structure; while in proteins, zinc has different coordination modes, the most common of which is the nearly regular tetrahedral coordination mode. This difference in coordination mode mainly determined by the interaction between Zn^{2+} and coordination molecules.

In traditional classical force field, the structure of the zinc binding group will be destroyed seriously in long time MD simulations due to the polarization effect and charge transfer effect between zinc ion and coordinated atoms are not properly considered. Even the total structure of metalloprotein will be distorted by the strong electrostatic interaction between +2 e charge of zinc ion and the protein environment. In a recent study of Friedman et al, 2 ns MD simulation of a zinc-containing metalloprotein (PDB ID: 2WCB) was performed with CHARMM 27 force field. The coordination mode of Zn^{2+} changed from a tetrahedral structure composed of three imidazole rings and a carboxyl group in the crystal structure to a six-coordinate structure that two water molecules squeezed into the metal binding group. In the study of MMP3 protein by Zhu et at., the AMBER 99SB force field also failed to maintain the Zn^{2+} coordination structure correctly. And the Friedman calculated the interaction energy between Zn^{2+} and its coordinated ligands in several representative complexes using both QM and MM methods respectively. As a result, it was found that the relative order of the QM interaction energies could not be calculated by ignoring or using only the classical force field method. Although the QM method performs well in generality and accuracy, it is limited by the computational cost. Recently, although hybrid QM/MM method[1], linear-scaling and/or fragmentation QM methods[2-5] offer the possibilities to treat large molecular systems, the computational cost is still an obstacle for long time AIMD simulation.

In the past two decades, several force fields for zinc ions were introduced, such as the SIBFA model of Gresh et al[6, 7]., the CTPOL model of Lim et al.[8, 9], the SLEF model of Wu et al., the AMOEBA model of Ren et al.[10, 11], the 12–6–4 LJ-type non-bonded model of Li and Merz[12], the ABEEM of Yang et al.[13], the Drude oscillator model of Roux et al.[14], a new CT model of Rick et al.[15], and the QPCT model for zinc in our previous work.[16] In these force fields, the polarization and charge transfer effects were included to some degrees, so the calculated

results were clearly improved.[17] However, this improvement is not always guaranteed and largely limited by the form of the potential function and the quality of parameters.

Fortunately, machine learning methods provide the possibility to develop molecular potentials with both the efficiency of the MM method and the accuracy of the QM method. Müller et al. performed an assessment and validation study of many machine learning methods, including the kernel ridge regression (KRR), support vector regression (SVR) and multilayer neural networks (NN) in the prediction of molecular atomization energies[18]. Among many machine learning methods, the artificial NNs offer an interesting approach for the construction of fully polarizable, non-rigid and reactive high dimensional potential energy surfaces starting from a set of QM data. They constitute a very flexible and unbiased class of mathematical functions, which in principle is able to approximate any real-valued function to arbitrary accuracy. Since Behler and Parrinello proposed the high-dimensional neural network (HDNN) scheme[19-23], many different neural network models have been proposed to learn the high dimensional potential energy surface of water, small organic molecules and metal materials. For example, the GDML and DTNN of Müller et al.[24-26], the kCON model of Hammer et al., and the Deep Potential model of Wang and co-workers[27]. Yang et al. also proposed a novel NN force field for water system based on an electrostatically embedded two-body expansion scheme.[28] Thanks to open source packages like DeepMD-kit[29] and TensorMol[30,31], training a neural network potential for specific molecular systems is generally straightforward at present.

Recently, we proposed a neural network potential model (NN/MM-RESP) which can describe the interactions between zinc ion and water accurately and reproduce Zn^{2+} hydration structure well in MD simulations. In addition, our recently proposed E-SOI-HDNN model can be used to achieve automatic construction of datasets and self-validation of model prediction results. Here, combined with NN/MM-RESP method and enhanced self-organizing incremental high dimensional neural network model (E-SOI-HDNN), we developed NN/MM-RESP-MBG to solve the difficulty of metalloprotein theoretical modelling. Its calculation efficiency is close to the molecular force field and the accuracy is close to QM calculations. A set of one nanosecond ab initio MD simulation were performed for each of the four common coordination modes of zinc-containing metalloproteins. In the ab initio MD simulations, the coordination pattern of metalloproteins were well maintained and the results show great agreement with the PDB Bank statistical values and crystal structures. All these results indicating that NN/MM-RESP-MBG can correctly describe the interaction between Zn^{2+} and proteins. In addition, the method can be easily extended to deal with different biomacromolecule systems containing other metal ions.

This paper is organizing as follows. In theory and method, the basic principles of NN/MM-RESP-MBG method and the automatic construction method of the neural network model for metalloproteins are briefly introduced. In the results and discussion, ab initio kinetics

simulations were performed on the four common coordination modes of zinc-containing metalloproteins, the reliability of the trajectory and the distribution of coordinated bond and angle were analyzed. Finally, brief conclusions and outlooks are given in the last section.

Theory and method

A. NN/MM-RESP-MBG methods

In NN/MM-RESP-MBG, if the distance between any atom of a residue that forms a coordination bond with metal ions is less than 2.8 Å or shorter, the side chain or main chain which contains the coordinated atom of this particular residue is treated as a member of the metal binding group. For example, there are three cysteine and one histidine residues that coordinate to the zinc ion in a CCCH type zinc finger (PDB ID: 2L30) as shown in Figure 1. All the atoms shown by ball-and-stick model within the dotted circle will be included in the metal binding group (MBG). According to the statistical value in PDB Bank, the cutoff distance of 2.8 Å was chosen because it covers metal-ligand bond distance in most common metalloproteins and is able to deal with most abnormal bond length structures in MD simulations. In addition, the hydration atoms are added to saturated the metal binding group at the position of broken bonds.

Figure 1. Division of metal binding groups in a CCCH type zinc finger protein which PDB ID
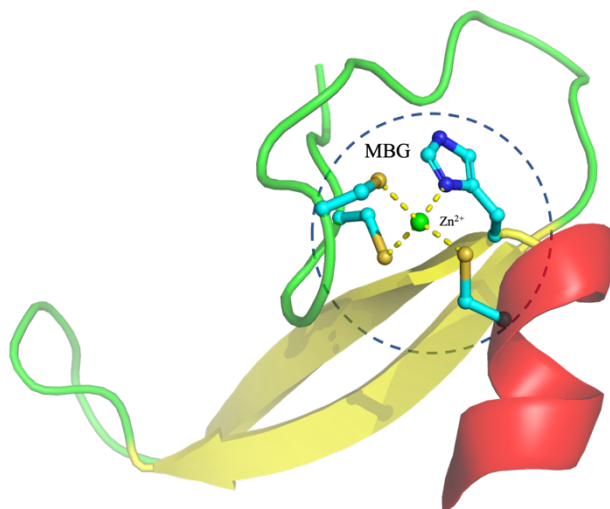
is 2L30. The part shown by the ball-and-stick model within the dotted circle is defined as metal binding group (MBG).

In NN/MM-RESP-MBG, a strategy similar to the QM/MM method is used for the total energy calculation. The QM energy and atomic forces of the entire MBG region will be predicted by E-SOI-HDNN model, while the rest of the system is described by the classical force field. The interaction between MBG group and the rest parts is calculated with the Coulomb and Lennard-Jones potential (mechanical embedding) in AMBER FF14SB. But the

atomic charge parameter is replaced by the RESP charge predicted by the E-SOI-HDNN model. Then the total energy of the system can be expressed as follows.

$$E_{total} = E_{MBG}^{E-SOI-HDNN} + E_{MM} + \sum_{i \in MBG} \sum_{j \notin MBG} (E_{i,j}^{ele} + E_{i,j}^{vdw}) \qquad (1)$$

B. E-SOI-HDNN model for metalloproteins

In order to achieve self-validation of neural network prediction results, rapid construction of datasets focused on metalloproteins, enhanced self-organizing incremental high-dimensional neural network (E-SOI-HDNN) model was used to construct the potential energy surface of MBG group. Its structure is as shown in Figure 2.
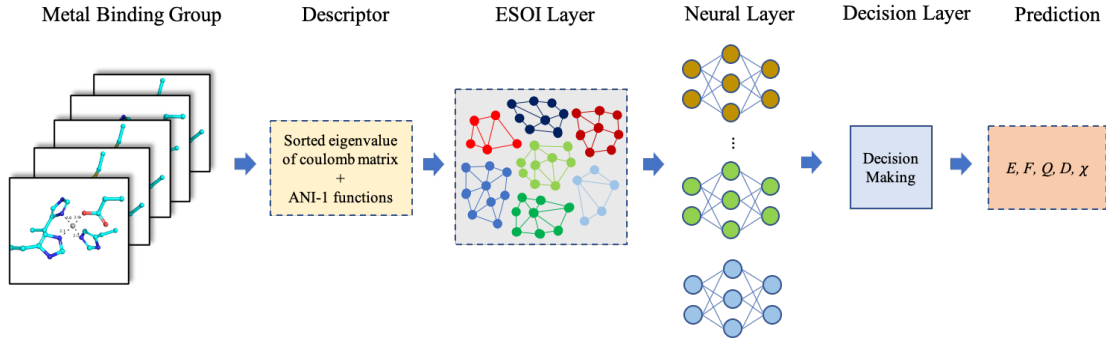
Figure 2. The structure of enhanced self-organizing incremental high dimensional neural network.

In E-SOI-HDNN model, each MBG structure is represented by two set of descriptors: regularized sorted eigen spectrum of the coulomb matrix and ANI-1 symmetry functions. The calculation of coulomb matrix is as shown in Eq (2).

$$C_{ij} = \begin{cases} 0.5Z_i^{2.4} & \forall i = j \\ \frac{Z_i Z_j}{|R_i - R_j|} & \forall i \neq j \text{ and } i \notin virtual \ atoms \\ 0 & \forall i \neq j \text{ and } i \in virtual \ atoms \end{cases} \qquad (2)$$

The regularized sorted eigen spectrum of coulomb matrix act as stimuluses to make adjustments to E-SOI layer. And the ANI-1 symmetry functions S_α of Isayev and co-workers are used as descriptor to fit energy, atomic forces and RESP charge in training process of neural layer. S_α consists the radial and angular part as shown in Eq (3) and Eq (4).

$$S_\alpha(radial) = \sum_{j \neq i} e^{-\eta(R_{ij} - R_s)^2} f_c(R_{ij}) \qquad (3)$$

$$S_\alpha(angular) = 2^{1-\zeta} \sum_{j \neq i, j \neq k} \left(1 + cos(\theta_{ijk} - \theta_s)\right)^\zeta \times e^{-\eta\left(\frac{R_{ij} + R_{ik}}{2} - R_s\right)^2} f_c(R_{ij}) f_c(R_{ik}) \qquad (4)$$

According to the intrinsic spectral similarity of MBG structure, E-SOI layer will finds the representative structures from the reference dataset, estimate the density of similar structures

and make cluster for datasets. The neural layer consists of a set of meta-networks, each of which corresponds to a sub-cluster in the E-SOI layer network. In this work, in order to deal with the interaction between the MBG and MM parts, we use modified Behler-Parrinello type high-dimensional neural network (HDNN) as the meta-network in neural layer, which considers the environmental charge and Van der Waals correction to predict the energy, atomic forces and RESP charge of the MBG region.

After the MBG structure R_t in the dataset is inputted, E-SOI layer will determine the m nearest subclasses according to the Euclidean distance between the corresponding eigen spectrum ε_t and the weight vector W of the corresponding node in the network. Then structure R_t will be added into training set of the corresponding meta-networks. In this way, each meta-network will be trained independently with its own reference dataset. The decision layer receives the predictions from neural layer and calculate the average of m meta-networks' predictions as the final prediction. An error indicator χ_t is defined to quantitatively describe the estimation of neural network errors for a configuration of R_t as shown in Eq (5).

$$\chi_t = \max\left\|F_{net_i,j}(R_t) - \langle F_{net_i,j}(R_t)\rangle\right\| \tag{5}$$

where $F_{net_i,j}(R_t)$ is the atomic force predictions of meta-network net_i for jth atom in structure R_t, and $\langle F_{net_i,j}(R_t)\rangle$ is the average force predictions of corresponding meta-networks for atom j.

According to χ_t, the probability P_{known} of a given structure as a known structure can be estimated by error indicator χ_t as shown in Eq (6).

$$P_{known} \sim 1 - \int_0^{\chi_t} 2N\left(0, \sum_{net_i \in M} w_{net_i}^2\right) d\chi_{t,j} \tag{6}$$

Then we can roughly define the known structure, questionable structure and unknown structure as shown in Eq (7)

$$R_t = \begin{cases} known & if\ 0 < \chi_t \leq \delta \\ questionable & if\ \delta < \chi_t \leq 2\delta \\ unknown & if\ \chi_t > 2\delta \end{cases} \tag{7}$$

Here, δ is the standard deviation of folded normal distribution of $\chi_{t,j}$ and can be calculated by Eq (8).

$$\delta = \sqrt{\sum_{net_i \in M} w_{net_i}^2} \tag{8}$$

And the final predictions of E-SOI-HDNN model are the average of an ensemble of meta-networks M as shown as follows.

$$E_{E-SOI-HDNN}(R_t) = \langle E_{net_i}(R_t)\rangle_{net_i \in M} \tag{9}$$

$$F_{E-SOI-HDNN,j}(R_t) = -\nabla_j E_{E-SOI-HDNN}(R_t) = \langle F_{net_i,j}(R_t)\rangle_{net_i \in M} \qquad (10)$$

$$Q_{E-SOI-HDNN,j}(R_t) = \langle Q_{net_i,j}(R_t)\rangle_{net_i \in M} \qquad (11)$$

C. Automated construction of datasets for metal binding group

Combined with E-SOI-HDNN model, we construct the dataset for four common coordination modes (CCCC, CCCH, CCHH, HHHD) of zinc-containing metalloproteins through active learning and MD simulation sampling. The main workflow is as shown in Figure 3.
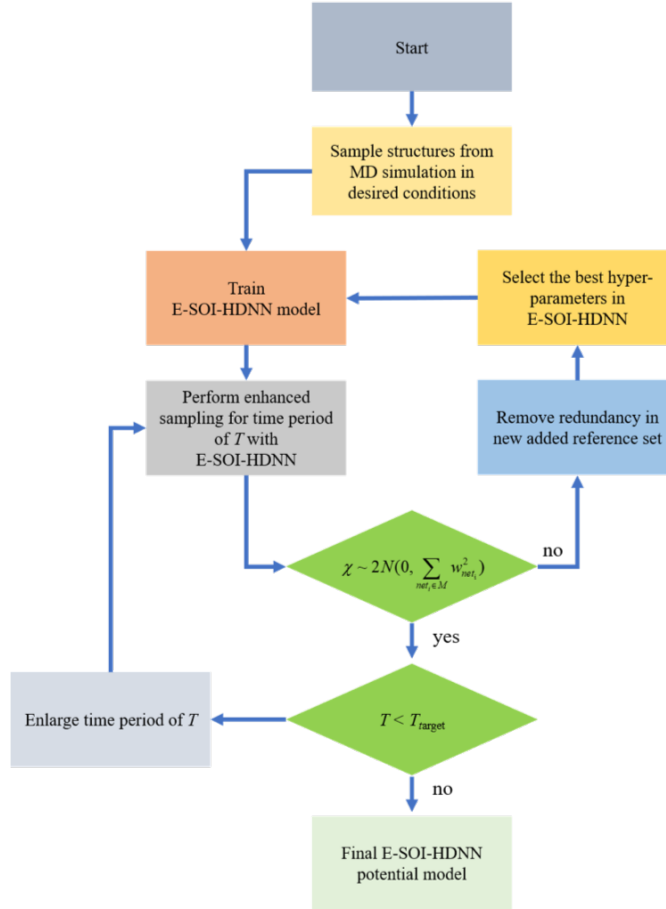
Figure 3. Automatic construction process of E-SOI-HDNN potential function model for metalloproteins.

Here, we illustrate the main workflow with an example of CCCH type zinc finger protein (PDB ID: 2L30).

(1) Perform a 100 fs QM/MM simulation with a time step of 1 fs. Take out the MBG structure and perform QM calculations as the initial dataset.

(2) Train the E-SOI-HDNN model. First, all the datasets are used for the training of E-SOI layer. After the training of E-SOI layer, the datasets are assigned to each meta-network

according to the clustering result. In each round of active learning process, the genetic algorithm is used to randomly generate structural parameters of each meta-network based on the best candidate structural parameters. The entire reference set is divided into training set and test set according to 9:1. After the training set is allocated by the E-SOI layer, the dataset of each meta-network is divided into a meta-training set and a meta-testing set according to 8:2 and the training of meta-network is carried out independently.

(3) It should be emphasized that the subclasses that account for less than 5% of the total number of training sets are merged into the nearest class to avoid the excessive number of clusters at the E-SOI layer to reduce the performance of each meta-network. And the training of the meta network continues until the prediction error of the atomic forces on each meta training set is less than $2\ kcal/(mol \cdot Å)$.

(4) Use the E-SOI-HDNN model to perform two sets of NN/MM-RESP-MBG dynamics simulations. One of them perform simulated annealing between 300 K and 400 K, which is mainly responsible for expanding the diversity of datasets. Another normal MD simulation is performed at 300K to determine whether the active learning process is iteratively completed. At each step of the simulation, the reliability of structural prediction is judged according to Eq (7). All questionable and unknown structures are put into the screening queue. In each round of active learning, the MD simulations are restarted from the initial structures.

(5) Judge the reliability of the trajectory. In each round of active learning, the error indicators χ_t in each MD step are tracked. When the average error indicator of 200 consecutive steps $\langle\chi_t\rangle_{200} > 2\delta$, the E-SOI-HDNN model is considered to be insufficiently sampled in the current phase space and the MD simulation is automatically stopped. Until the simulation time can reach 1 nanosecond at 300 K and all the error indicator $\chi_t < 2\delta$, the active learning process is completed. Then go to step (8).

(6) Eliminate abnormal and redundant structures in the screening queue. The structure in the screening queue. Here, a structure is regarded as an abnormal structure and be discarded directly if there is a distance between two atoms is less than 0.6 Å or greater than 15 Å. when the number of structures in the screening queue exceeds 1000, E-SOI layer will determine the types of each structure. Then we randomly select 1000 structures of "internal", "edge" and "noise" types add into reference dataset according to 1:3:1.

(7) Generate the best candidate hyperparameters of E-SOI-HDNN model with genetic algorithm. Go back to step (2).

(8) Train the final E-SOI-HDNN model with the iteratively completed datasets and best candidate hyperparameters.

D. Computational Details

In this work, all the QM calculations are performed with Gaussian 16 at M062X/SDD level. The classical molecular force field used in NN/MM-RESP-MBG MD simulations is AMBER FF14SB. To better match the Amber FF14SB force field, the ESP data was calculated at the HF/6-31G* level, then RESP charges were fitted with the RESP module in the Amber 18 package. In the training of the E-SOI layer, the maximum age of node connection age_{max} is set to 10 times and every 500 times input are defined as a learning cycle. In the training of the neural layer, the initial structural parameters of the meta-networks are [200,200,200]. All the NN/MM-RESP-MBG MD simulations and the training of E-SOI-HDNN model are carried out through ESOI-CHEM package.

Here, we construct the E-SOI-HDNN model for four common coordination modes of zinc-containing modes of zinc-containing metalloproteins. According to the number of various types of amino acids in the metal binding group, four types of MBG can be abbreviated as CCCC, CCCH, CCHH and HHHO where C represents cysteine, H represents histidine and O is aspartic or glutamate residues. We selected 4 different representative proteins for 4 coordination modes which PDB ID were 1ZIN (CCCC), 2L30 (CCCH), 1AAY (CCHH) and 1HFS (HHHO). Among 4 zinc-containing metalloproteins, the crystal structures of 1ZIN and 1AAY are in good agreement with the statistics of PDB Bank, while the bond length or angle distribution of crystal structure of 1HFS and 2L30 clearly deviate from the statistic distribution. Before the MD simulations, we optimized the protein crystal structure with AMBER FF14SB in a 25 Å cubic water box under periodic conditions. Then a 500 ps heating simulation, 5ns relaxation and 10 ns production simulation were performed to relax the protein structure. During the MD simulations, structural constrains were added to the metal binding group to prevent the coordination area from being destroyed. After the pretreatment discussed above, the optimized proteins were placed in a water ball with a radius of 25 Å as the initial structure of NN/MM-RESP-MBG MD simulations.

Result and discussion

**A. Performance of E-SOI-HDNN model**

The performance of E-SOI-HDNN model on the training set and test set of the four common coordination modes of zinc-containing metalloproteins is as shown in Table 1. It can be seen that the E-SOI-HDNN model can predict the energy, atomic forces and RESP charge accurately. On the test set, the root means square error (RMSE) of energy of CCHH type is the largest, only 1.78 kcal/mol and all the RMSE of atomic forces in four coordination modes are

less than 1.8 $kcal/(mol \cdot Å)$. It indicates that the E-SOI-HDNN can be applied to AIMD simulations.

Table 1. The performance of E-SOI-HDNN model on training set and test set of four common coordination modes of zinc-containing metalloproteins.

MBG	PDB ID	Number of Classes in E-SOI Layer	Training Set/Test Set			
			Structure Number	RMSE of E (kcal/mol)	RMSE of **F** (kcal/(mol·Å))	RMSE of Q (e)
CCCC	1ZIN	7	11900/1200	1.43/1.29	1.53/1.43	0.04/0.05
CCCH	2L30	12	28156/3200	1.38/1.34	1.68/1.75	0.03/0.04
CCHH	1AAY	14	45328/5100	1.78/1.64	1.41/1.52	0.02/0.03
HHHO	1HFS	11	27100/3000	1.30/1.26	1.63/1.72	0.04/0.03

Then, 1 ns AIMD simulation on the four main modes of metalloproteins were performed with NN/MM-RESP-MBG method respectively. In the dynamic simulations, the reliability is judged in real time whether the simulation result is reliable with Eq (7). The distribution of the error indicator χ_t of four trajectories are as shown in Figure 4. In order to facilitate analysis and representation, the RMSE of the E-SOI-HDNN model on the test set is used to replace the RMSE of each meta-network to calculate the δ in Eq (8). As can be seen in Figure 4, the maximum of error indicator in all four sets of trajectories is 4.68 $kcal/(mol \cdot Å)$ and the error indicator of all the structures are within the range of (0, 2δ), which means that there are not any unknown structures in simulations and the trajectory results are reliable.
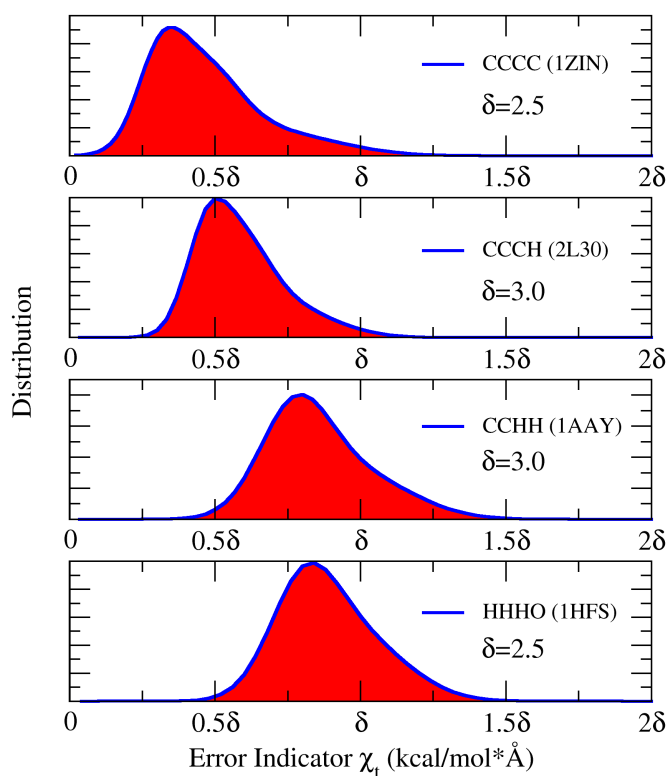
Figure 4. The distribution of error indicator χ_t of four common coordination modes of zinc-containing metalloproteins.

B. Charge distribution of Zn^{2+} and coordinated atoms

In the classical force field such as AMBER FF14SB, the charge of Zn^{2+} is fixed at +2 e. There are strong electrostatic interactions between Zn^{2+} and protein environment. Zn^{2+} will attract other polar molecules like water into the coordination area and destroy the normal four-coordinated structure. This problem has been partially improved in the advanced force field considering polarization and charge transfer effects. In NN/MM-RESP-MBG, the short-range polarization and charge transfer effects between Zn^{2+} and coordination atoms are fully considered by E-SOI-HDNN model. And the RESP charge of metal binding group are used to treat the electrostatic interaction between MBG and MM area. In the 1 ns AIMD simulations, it was found that the charge distribution of Zn^{2+} is obviously different under different coordination modes. Its distribution is as shown in Figure 5 and the average resp charge of coordinated atoms are shown in Table 2.

Table 2. The RESP charge of Zn^{2+} and coordinated atoms in four different coordination mode of MBG in 1 ns AIMD simulations.

1ZIN (CCCC)	Coordinated Atoms	Zn^{2+}	S^{γ}@C5	S^{γ}@C8	S^{γ}@C25	S^{γ}@C28
	E-SOI-HDNN RESP charge	1.25	-0.85	-0.87	-0.86	-0.84
	Amber charge	2	-0.88	-0.88	-0.88	-0.88
2L30 (CCCH)	Coordinated Atoms	Zn^{2+}	S^{γ}@C5	S^{γ}@C8	N^{δ}@H37	S^{γ}@C40
	E-SOI-HDNN RESP Charge	0.98	-0.87	-0.95	-0.41	-0.81
	Amber Charge	2	-0.88	-0.88	-0.57	-0.88
1AAY (CCHH)	Coordinated Atoms	Zn^{2+}	S^{γ}@C5	S^{γ}@C10	N^{ε}@H23	N^{ε}@H27
	E-SOI-HDNN RESP charge	0.83	-0.82	-0.81	-0.37	-0.42
	Amber charge	2	-0.88	-0.88	-0.57	-0.57
1HFS (HHHO)	Coordinated Atoms	Zn^{2+}	N^{ε}@H64	$O^{\delta 1}/O^{\delta 2}$@H66	N^{ε}@H78	N^{δ}@H92
	E-SOI-HDNN RESP Charge	0.98	-0.51	-0.64 / -0.87	-0.43	-0.38
	Amber Charge	2	-0.57	-0.88	-0.57	-0.57

In the CCCC coordination mode, Zn^{2+} forms coordination bonds with four cysteine residues. Compared with other coordination modes, the average RESP charge of Zn^{2+} is the highest, which is +1.25 e and significantly lower than +2 e in the classical force field. As shown in Table 2, the average RESP charge of S atom is almost the same as the AMBER charge. It indicates that the defects of AMBER force field are mainly coming from the unreasonable charge of Zn^{2+}. In CCCH coordination mode, Zn^{2+} forms coordination bonds with three cysteines and one histidine in the crystal structure of 2L30. The average RESP charge of Zn^{2+} is +0.98 e, which shows that the charge transfer effect between Zn^{2+} and the protein is stronger than that of CCCC mode. It is clear that the average RESP charge of coordinated N atom on histidine is smaller than AMBER charge while the S atom is basically same as that in CCCC mode. It indicates that the charge transfer effect between Zn^{2+} and coordinated N atoms are more obvious in NN/MM-RESP-MBG. In CCHH coordination mode, the average RESP charge of Zn^{2+} is +0.83 e which is 0.15 e less than that in CCCH mode. Compared with the Amber FF14SB force field, the RESP charge of coordinated N^{ε} atom in histidine increased by about 0.15 e. It can be seen that Amber has defects in handling the charge transfer interaction between Zn^{2+} and N atoms. Finally, we analyze the coordination mode of HHHO, Zn^{2+} forms coordination bonds with an aspartic residue and three histidine residues in the crystal structure of 1HFS. Generally, when carboxyl group coordinated with Zn^{2+}, there are two coordination cases of dioxygen coordination and mono oxygen coordination, which mainly depends on the interaction between the coordinated carboxyl and the protein environment. In the crystal structure of 1HFS, the

coordination mode between carboxyl group and zinc ion is mono oxygen coordination due to the hydrogen bond between $O^{\delta 1}$ of ASP66 and a hydrogen atom of TYR68. The average RESP charge of Zn^{2+} is +1.12 e. It is worth noting that in Amber FF14SB, the charge of the two oxygen atoms on the carboxyl group is both +0.88 e due to the incorrect consideration of polarization and charge transfer effects. It overestimates the interaction between carboxyl group and Zn^{2+} and make it difficult to maintain a mono oxygen coordination mode. In NN/MM-RESP-MBG, the charge of the $O^{\delta 2}$ atom coordinated to Zn^{2+} on the carboxyl group is -0.87 e, while the average RESP charge of the non-coordinated $O^{\delta 1}$ atom is only -0.64 e. In the 1ns AIMD simulation, due to the polarization and charge transfer effect are well considered by neural network potential model, the carboxyl single coordination structure is well maintained.

C.   AIMD simulation with NN/MM-RESP-MBG

In this work, we performed 1ns AIMD simulations for four common coordination mode of zinc-containing metalloproteins. First, we verified the accuracy of NN/MM-RESP-MBG in 1ZIN (CCCC type) and 1AAY (CCHH type). Then we refined the crystal structure of 2L30 and 1HFS with NN/MM-RESP-MBG MD simulations.
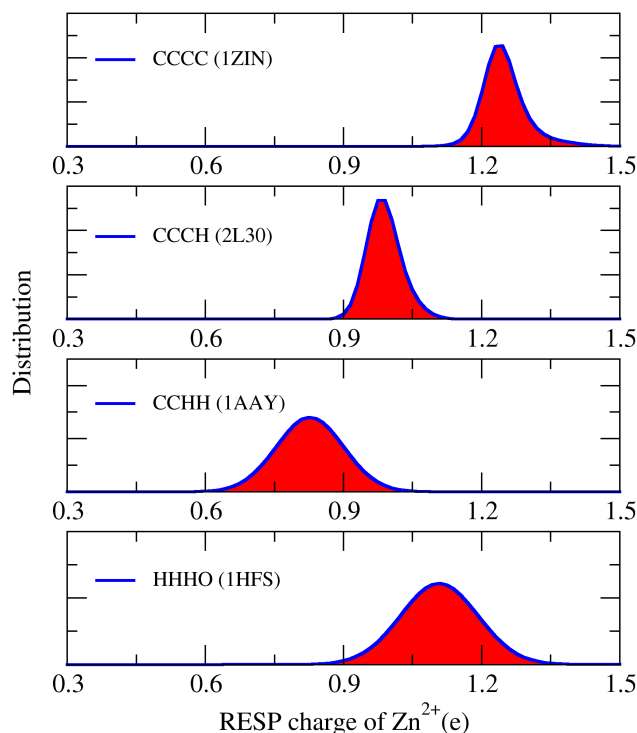
Figure 5. RESP Charge distribution of Zn^{2+} under four different coordination modes.

C.1   Structure of metal binding group in AIMD simulation

Firstly, we analyzed the structure of the Zn^{2+} coordination region of the protein 1ZIN in 1ns AIMD simulation. The distribution of the bond length and angle between Zn^{2+} and coordinated atoms are as shown in Figure 6 and Figure 7. And the comparation between statistic average in MD simulation and PDB Bank statistical value and crystal structure is as shown in Table 3. It can be seen that the statistical average values of all the coordinate bond lengths and angles are within the error range given by PDB Bank, indicating the accuracy of the results. In the AIMD simulation, the statistical average of the bond length of the Zn-S bond is 2.4 Å, which is about 0.05 Å longer than the statistical value of the PDB Bank. There are two possible reasons. On the one hand, the statistical values of all PDB Banks given in this article values of all PDB Banks are statistically averaged in all different coordination modes, which are not specific to the CCCC coordination mode. On the other hand, the theory level may have an impact on the accuracy of AIMD simulations. Here, the E-SOI-HDNN model is trained at the M062X/SDD level where the accuracy and efficiency for the structure optimization of zinc-containing Cambridge small molecule database are well balanced. The performance of this theory level has not been thoroughly compared for metalloproteins. Based on the above considerations, we believe that the current results are sufficiently accurate at the current level. In addition, the result angle distribution is also in good agreement with PDB Bank statistics. The statistical average values in the simulation are all within the error distribution range of PDB Bank statistics.

Table 3. Comparation between the coordination bond and angle statistical average value in NN/MM-RESP-MBG MD simulation and PDB Bank statistical values and crystal structure of 1ZIN. The unit of bond length is Å, and the unit of angle is °.

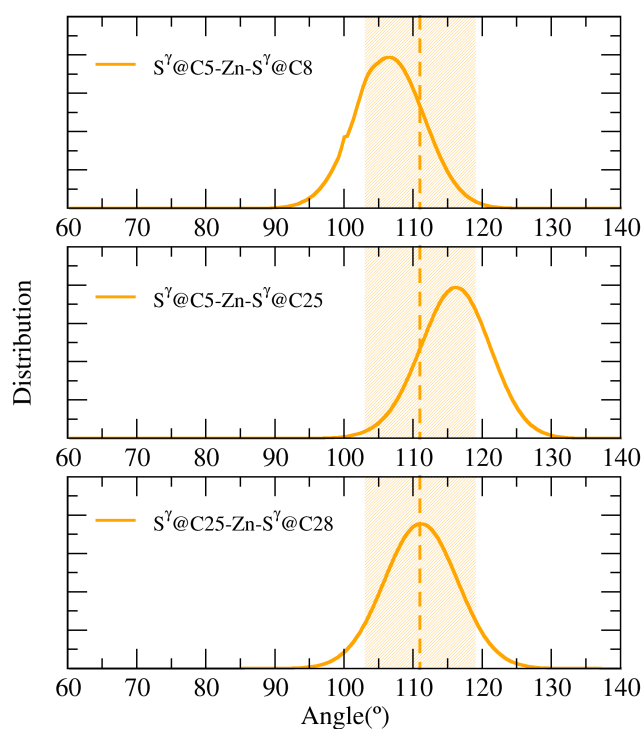| Zinc-ligand geometry	PDB survey	NN/MM-RESP-Metal	X-ray
Zn-S$^\gamma$@C5	2.35 ± 0.09	2.41	2.33
Zn-S$^\gamma$@C8	2.35 ± 0.09	2.39	2.3
Zn-S$^\gamma$@C25	2.35 ± 0.09	2.4	2.32
Zn-S$^\gamma$@C28	2.35 ± 0.09	2.41	2.33
∠S$^\gamma$@C5-Zn-S$^\gamma$@C8	111 ± 8	107	114
∠S$^\gamma$@C5-Zn-S$^\gamma$@C25	111 ± 8	116	106
∠S$^\gamma$@C25-Zn-S$^\gamma$@C28	111 ± 8	111	112

Figure 6. The distribution of coordination bond length in 1 ns NN/MM-RESP-MBG MD simulation of CCCC mode, where the dotted line represents the average value of PDB Bank statistics and the orange area represents the statistic error distribution given by PDB Bank database for Zn-S bond length.
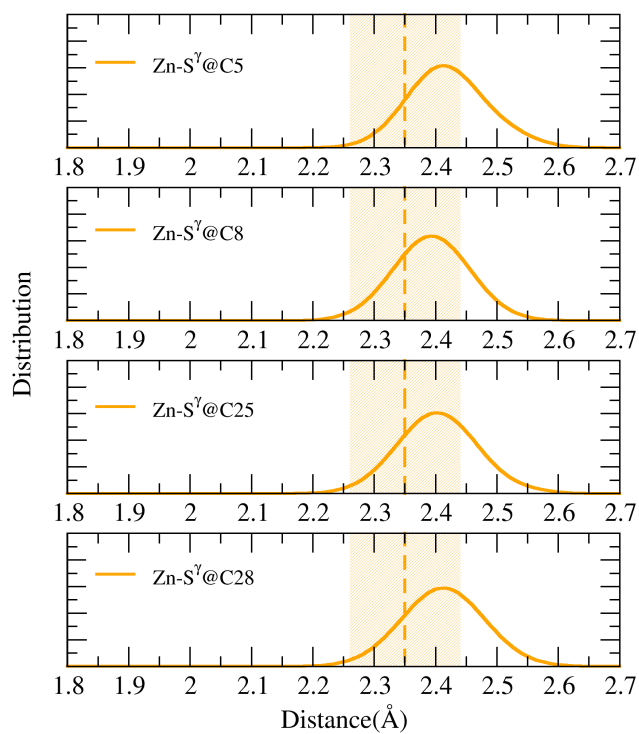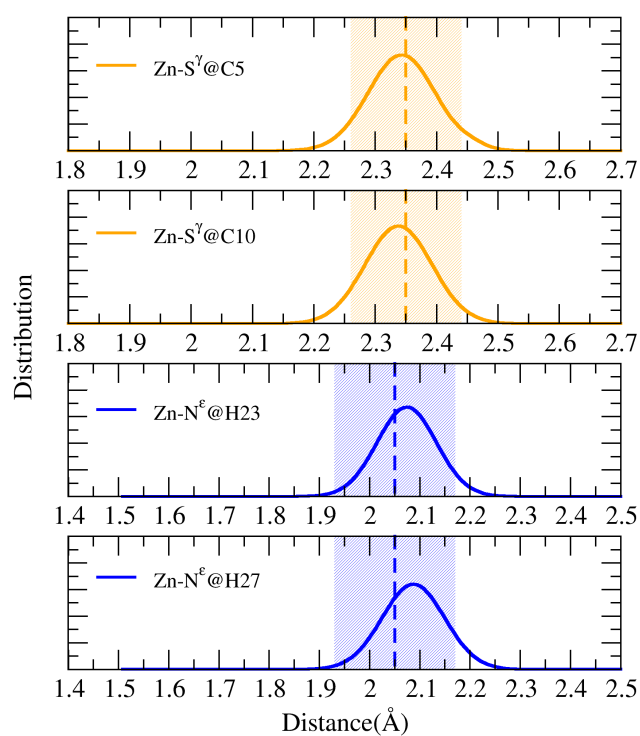
Figure 7. The distribution of coordination angle in 1 ns NN/MM-RESP-MBG MD simulation of CCCC mode, where the dotted line represents the average value of PDB Bank statistics and the orange area represents the statistic error distribution given by PDB Bank database for S-Zn-S angle.

In the coordination mode of CCHH, the distribution of the bond length and angle between Zn^{2+} and coordinated atoms in AIMD simulation of 1AAY are as shown in Figure 8 and Figure 9. And the comparison of the statistical average value in the simulation with PDB Bank statistical value and crystal structure is as shown in Table 4. The average bond length of the Zn-S bond and Zn-N are 2.35 Å and 2.08 Å respectively, both of them are in good agreement with the statistical value of the PDB Bank. But for the coordination angle of N-Zn-N, there is an obvious deviation from the statistical value and crystal structure in the PDB Bank as shown in Figure 9. In order to check the problem, we found relevant QM/MM simulation results in recent work through literature research. The statistical average value of N-Zn-N in NN/MM-RESP-MBG MD simulation is 97°. In recent 50ps QM/MM simulation of 1AAY, the statistical average of N-Zn-N is 99°. Two results are in good agreement with each other. It indicates the deviation of coordination angle of N-Zn-N does not come from the inaccurate prediction of the E-SOI-HDNN model. In the following work, we will continue to verify whether the angular distribution of N-Zn-N is same on more CCHH type zinc-containing metalloproteins.

Table 4. Comparison between the coordination bond and angle statistical average value in NN/MM-RESP-MBG MD simulation, PDB Bank statistical values and crystal structure of 1AAY. The unit of bond length is Å, and the unit of angle is °.

Zinc-ligand geometry	PDB survey	NN/MM-RESP-Metal	X-ray	QM/MM (50 ps)
Zn-Sγ@C5	2.35 ± 0.09	2.35	2.29	2.32 ± 0.06
Zn-Sγ@C10	2.35 ± 0.09	2.34	2.29	2.34 ± 0.07
Zn-Nε@H23	2.05 ± 0.12	2.07	2.04	2.12 ± 0.07
Zn-Nε@H27	2.05 ± 0.12	2.09	2.04	2.13 ± 0.07
∠Nε@H23-Zn-Nε@H27	107 ± 8	97	105	99 ± 7
∠Nε@H23-Zn-Sγ@C5	109 ± 8	108	109	108 ± 7
∠Sγ@C5-Zn-Sγ@C10	111 ± 8	116	113	114 ± 6

Figure 8. The distribution of coordination bond length in 1 ns NN/MM-RESP-MBG MD simulation of CCHH mode, where the dotted line represents the average value of PDB Bank statistics, the orange area and blue area represent the statistic error distribution given by PDB Bank database for Zn-S and Zn-N bond length respectively.
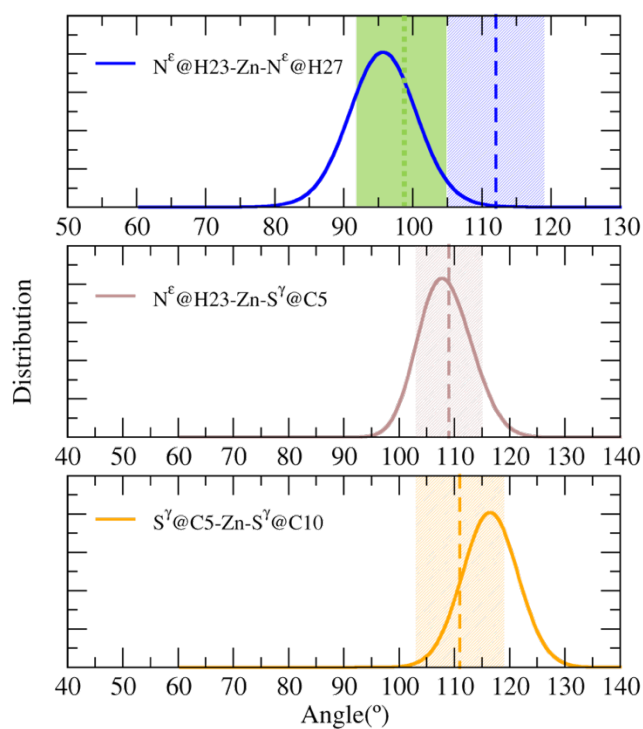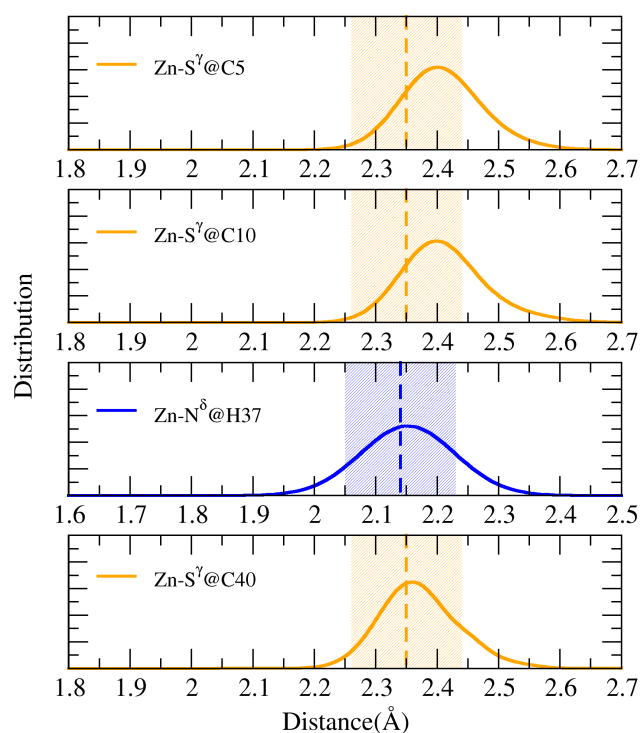
Figure 9. The distribution of coordination angle in 1 ns NN/MM-RESP-MBG MD simulation of CCHH mode, where the dotted line represents the average value of PDB Bank statistics and the orange area, brown area and blue area represents the statistic error distribution given by PDB Bank database for S-Zn-S, N-Zn-S and N-Zn-N respectively. And the green area and dot-dash line represent the relevant N-Zn-N angle error range and statistical average in 50 ps QM/MM simulation.

C.2   Structure refinement with NN/MM-RESP-MBG

In the CCCH coordination mode, the bond length of Zn-N in the crystal structure of 2L30 is only 2.01 Å, which deviates from the statistical value of PDB Bank seriously while the bond length of Zn-N is refined in NN/MM-RESP-MBG MD simulations. The comparison between the statistical average value and PDB Bank statistics and crystal structures are as shown in Table 5. As the results discussed above, the statistical average of all coordination bond lengths and angles are within the error range given by PDB Bank as shown in Figure 10 and Figure 11, indicating the accuracy of the results. The distribution of Zn-N bond length completely matches the statistic distribution of PDB Bank database. Although the distribution of the refined structural coordination bond angle slightly deviates from the statistical value, the regular tetrahedral coordination mode of the crystal structure is perfectly maintained and the average value of the angle in dynamics is completely consistent with crystal structure. It means that NN/MM-RESP-MBG has the ability to optimize and refine the abnormal crystal structure of metalloproteins.

Table 5. Comparison between the coordination bond and angle statistical average value in NN/MM-RESP-MBG MD simulation, PDB Bank statistical values and crystal structure of 2L30. The unit of bond length is Å, and the unit of angle is °.

Zinc-ligand geometry	PDB survey	NN/MM-RESP-Metal	X-ray
Zn-Sγ@C5	2.35 ± 0.09	2.4	2.34
Zn-Sγ@C8	2.35 ± 0.09	2.39	2.34
Zn-Nδ@H37	2.14 ± 0.09	2.15	2.01
Zn-Sγ@C40	2.35 ± 0.09	2.35	2.34
∠Nδ@H37-Zn-Sγ@C40	109 ± 8	113	114
∠Sγ@C5-Zn-Sγ@C8	111 ± 8	108	109
∠Sγ@C8-Zn-Sγ@C40	111 ± 8	114	113

Figure 10. The distribution of coordination bond length in 1 ns NN/MM-RESP-MBG MD simulation of CCCH mode, where the dotted line represents the average value of PDB Bank statistics, the orange area and blue area represent the statistic error distribution given by PDB Bank database for Zn-S and Zn-N bond length respectively.
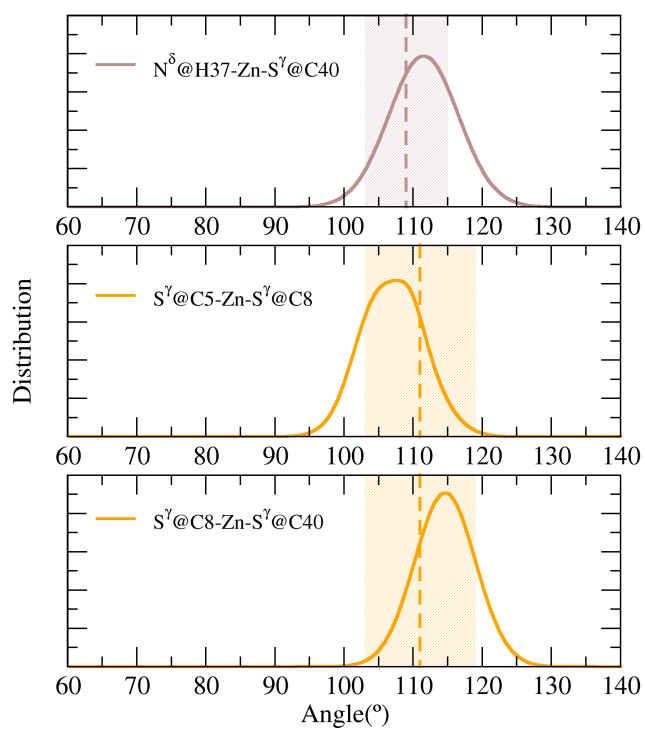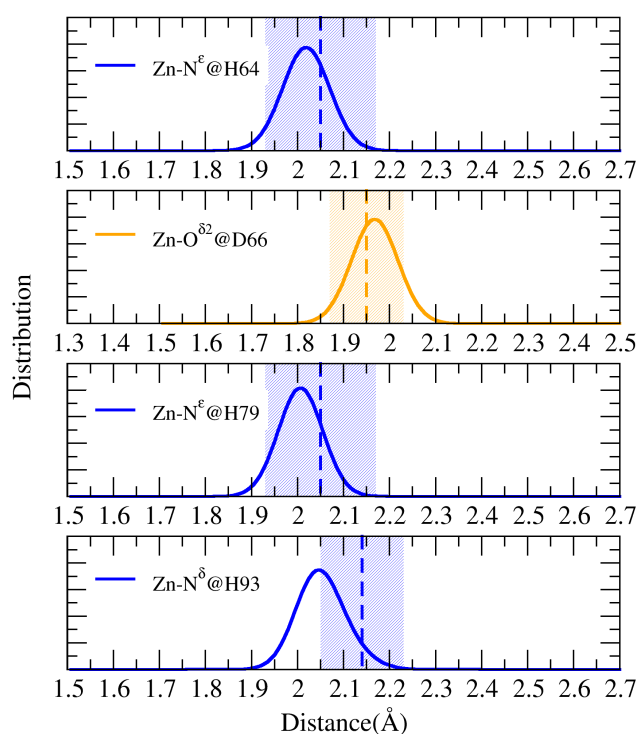
Figure 11. The distribution of coordination angle in 1 ns NN/MM-RESP-MBG MD simulation of CCHH mode, where the dotted line represents the average value of PDB Bank statistics and the orange area, brown area represents the statistic error distribution given by PDB Bank database for S-Zn-S, N-Zn-S respectively.

Finally, we analyzed the HHHO coordination mode and the result is as shown in Figure 12 and Figure 13. In the crystal structure of 1HFS, the coordination bond length between Zn^{2+} and N atoms is significantly shorter than the statistic value of PDB Bank. Among them, the two Zn-N bond lengths are only 1.78 Å and 1.83 Å. and the distance between Zn^{2+} and N^δ atom of 93th HIE residue is 2.01 Å as shown in Table 6. And not only the coordination bond length of Zn-N recovered to 2.01 Å, 2.02 Å and 2.06 Å respectively, which are in good agreement with the statistic of PDB Bank, but also the angular distribution bet Zn^{2+} and coordination atoms is consistent with the regular tetrahedral coordination structure in the crystal structure.

Table 5. Comparation between the coordination bond and angle statistical average value in NN/MM-RESP-MBG MD simulation, PDB Bank statistical values and crystal structure of 1HFS. The unit of bond length is Å, and the unit of angle is °.

| Zinc-ligand geometry	PDB survey	NN/MM-RESP-Metal	X-ray
Zn-N^ε@H64	2.05 ± 0.12	2.02	1.83
Zn-$O^{\delta 2}$@D66	1.95 ± 0.08	1.97	2
Zn-N^ε@H79	2.05 ± 0.12	2.01	1.78
Zn-N^δ@H93	2.14 ± 0.09	2.06	2.01
$\angle N^\varepsilon$@H64-Zn-$O^{\delta 2}$@D66	107 ± 12	111	105
$\angle N^\varepsilon$@H64-Zn-N^δ@H79	112 ± 7	116	119
$\angle N^\varepsilon$@H79-Zn-N^δ@H93	112 ± 7	112	113

Figure 12. The distribution of coordination bond length in 1 ns NN/MM-RESP-MBG MD simulation of HHHO mode, where the dotted line represents the average value of PDB Bank statistics, the orange area and blue area represent the statistic error distribution given by PDB Bank database for Zn-O and Zn-N bond length respectively.
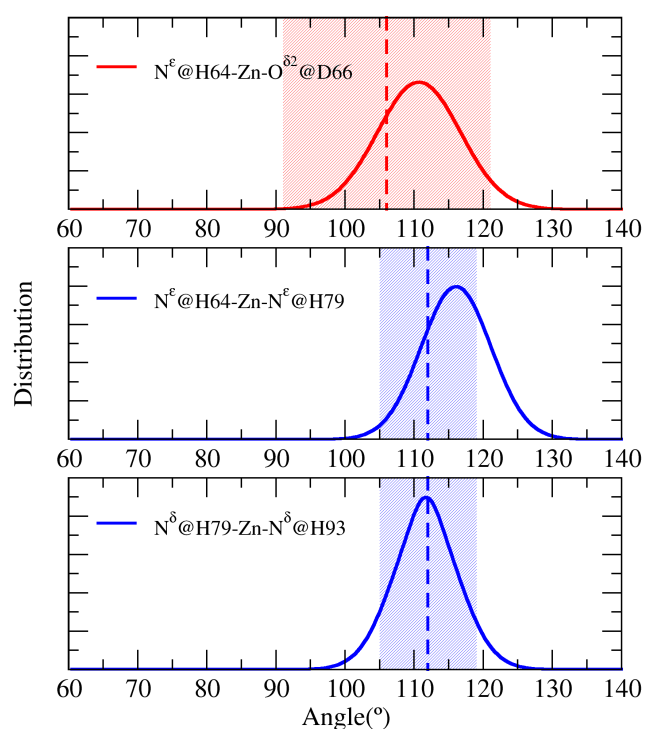
Figure 11. The distribution of coordination angle in 1 ns NN/MM-RESP-MBG MD simulation of CCHH mode, where the dotted line represents the average value of PDB Bank statistics and the pink area, blue area represents the statistic error distribution given by PDB Bank database for N-Zn-O, N-Zn-N respectively.

To further test the ability of NN/MM-RESP-MBG to refine the wrong structures of zinc-containing metalloproteins, we refined two abnormal structures sampled in 350 K MD simulations.

Here, one of them is to refine the abnormal structure of 2L30 with excessively Zn-S coordination bond. The structure of metal binding group before and after refinement is as shown in Figure 14 (a) and (b). Before the refinement, the distance between Zn^{2+} and S^γ atom of 8th CYM was 2.92 Å. In order to deal with this abnormal structure in simulation, the distance threshold for dividing MBG is set to 3.00 Å. After 30 ps NN/MM-RESP-MBG simulation, the bond length of Zn-S^γ@CYM8 has been restored to 2.40 Å and the regular tetrahedral coordination is well maintained. The time evolution of Zn-S^γ@CYM8 is as shown in Figure 14 (c).
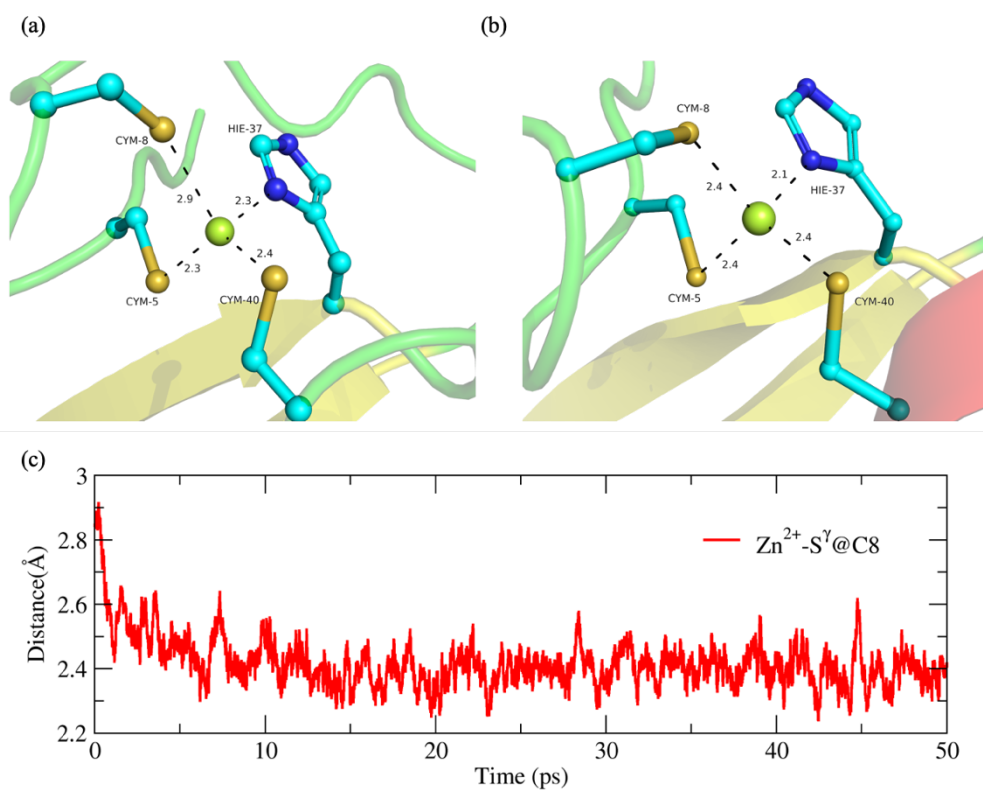
Figure 14. Structure refinement with NN/MM-RESP-MBG for 2L30 abnormal structure. (a) the abnormal initial structure, (b) the refined structure, (c) the time evolution of Zn-S distance.

Another one is to refine the abnormal structure of 1HFS with an excessively large N-Zn-O coordination angle which means the regular tetrahedral coordination structure of MBG is distorted. Before and after refinement, the structure of MBG is as shown in Figure 15 (a) and (b) respectively. It can be seen that the $N^\epsilon - Zn - O^{\delta 2}$ angle is 140° which is seriously deviated from the ideal coordination angle 109° of regular tetrahedral coordination. After less than 5 ps NN/MM-RESP-MBG MD simulation, the angle of $N^\epsilon - Zn - O^{\delta 2}$ quickly drops from 140° to 120°. If we continued the refined MD simulations, the angle will gradually return to the statistical average value of 111°.
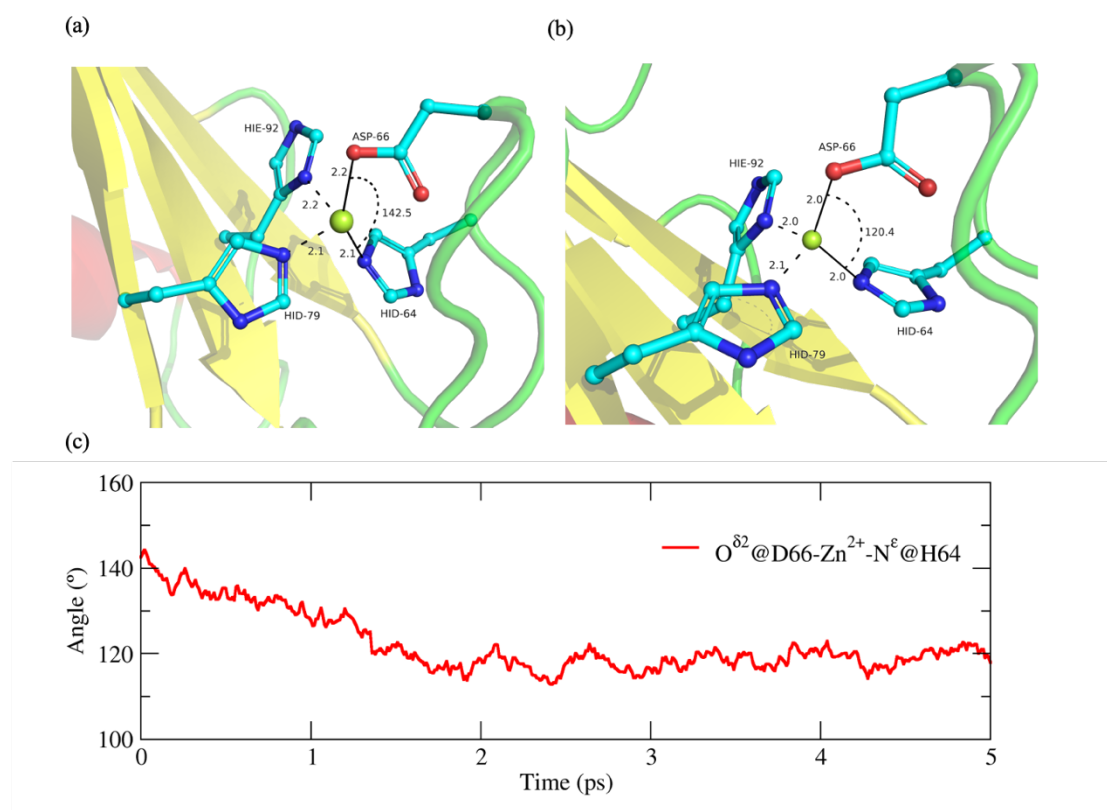
Figure 15. Structure refinement with NN/MM-RESP-MBG for 1HFS abnormal structure. (a) the abnormal initial structure, (b) the refined structure, (c) the time evolution of O-ZN-O coordination angle.

Conclusion and outlook

In this work, the NN/MM-RESP-MBG AIMD simulation method for metalloproteins is developed. In this method, the E-SOI-HDNN potential function model is used to describe the

interaction between Zn^{2+} in the metal binding group and each coordination molecules. The classical molecular force field is used to describe the long-range interaction between the external environment and MBG.

To more accurately describe the effect of the MBG on the protein, the RESP charge predicted by the E-SOI-HDNN model is used to replace the fixed charge model in the classical molecular force field. For four common coordination mode of zinc-containing metalloproteins, E-SOI-HDNN can accurately predict the atomic charge, potential energy and atomic forces of MBG while the computational efficiency is close to that of classical molecular force field. Combined with E-SOI-HDNN model and related active learning method, the reference dataset for zinc-containing metalloproteins are automated constructed. The model can not only handle stable crystal structures, but also accurately predict the energy and force of abnormal structures with strange coordination bond or angle. With the self-verification characteristics of E-SOI-HDNN model, we can judge the accuracy of NN/MM-RESP-MBG MD simulation. NN/MM-RESP-MBG method can be applied to long time MD simulations and structure refinement for metalloproteins. In 1 ns AIMD simulations of 4 zinc-containing metalloproteins with different common coordination mode, the structures of MBG are in good agreement with statistic value of PBD Bank dataset. In structure refinement MD simulations, the structure of MBG can be recover from the distorted tetrahedral coordination and the distribution of coordination bond and angle is consistent with the statistics of the PDB Bank database.

Compared with the classical molecular force field, the neural network potential function model is not limited by the function form and complex parameterization process. Its form is flexible and there are not any priori deviations. All local quantum effects, especially polarization effects and charge transfer effect can be described accurately with NN potential model. This method has advantages when dealing with molecular systems containing complex quantum effects, such as the metalloproteins or other biological macromolecular systems containing metal ions. In addition, the computational efficiency of the neural network potential function model is improved by several orders of magnitude compared to the QM calculation and QM / MM methods. For the zinc-containing metalloproteins studied in this work, on a common Linux server with a 16-core CPU and an NVIDIA GTX1080Ti GPU card, the single-step simulation with NN/MM-RESP-Metal method only takes about 0.1 s. Since the computational cost of the E-SOI-HDNN model mainly comes from the prediction in neural layer, combined with multi-GPU parallelism, the calculation efficiency can be further improved. Using this method, it is easy to deal with related problems on the time scale of nanoseconds, such as metalloproteinase catalysis, ligand binding and other important biochemical processes.

Although the neural network potential function model presented in this work has the advantages of accuracy and efficiency, there are also some shortcomings in NN/MM-RESP-MBG. First, the polarization effect of the environment on ZBG is not considered, and only the

short-range polarization effect between the Zn^{2+} and the coordination molecule is included. Secondly, the MBG area is fixed because it is hard to deal with the continuity of the potential energy surface of the total system in the current scheme. Finally, the neural network potential function model used in this work is trained with reference to the DFT calculation results. There are still rooms to improve the performance of DFT calculations when dealing with weak interactions. Therefore, a higher theory level research should be carried out.

Despite the above shortcomings, the accuracy of the NN/MM-RESP-MBG method is fully guaranteed with the calculation efficiency of the classical molecular force field in AIMD simulations of metalloproteins, it can greatly extend the application of AIMD in metalloproteins related researches.

Reference

1. Cauet, E.; Bogatko, S.; Weare, J. H.; Fulton, J. L.; Schenter, G. K.; Bylaska, E. J., Structure and dynamics of the hydration shells of the Zn(2+) ion from ab initio molecular dynamics and combined ab initio and classical molecular dynamics simulations. *J. Chem. Phys.* **2010,** *132* (19), 194502.

2. Dahlke, E. E.; Truhlar, D. G., Electrostatically embedded many-body correlation energy, with applications to the calculation of accurate second-order Moller-Plesset perturbation theory energies for large water clusters. *J. Chem. Theory Comput* **2007,** *3* (4), 1342-1348.

3. Dahlke, E. E.; Truhlar, D. G., Electrostatically embedded many-body expansion for large systems, with applications to water clusters. *J. Chem. Theory Comput* **2007,** *3* (1), 46-53.

4. Dahlke, E. E.; Truhlar, D. G., Electrostatically embedded many-body expansion for simulations. *J. Chem. Theory Comput* **2008,** *4* (1), 1-6.

5. Liu, J.; He, X.; Zhang, J. Z. H.; Qi, L. W., Hydrogen-bond structure dynamics in bulk water: insights from ab initio simulations with coupled cluster theory. *Chem. Sci* **2018,** *9* (8), 2065-2073.

6. Gresh, N., Energetics of Zn2+ Binding to a Series of Biologically Relevant Ligands - a Molecular Mechanics Investigation Grounded on Ab-Initio Self-Consistent-Field Supermolecular Computations. *J. Comput. Chem.* **1995,** *16* (7), 856-882.

7. Gresh, N.; de Courcy, B.; Piquemal, J. P.; Foret, J.; Courtiol-Legourd, S.; Salmon, L., Polarizable Water Networks in Ligand-Metalloprotein Recognition. Impact on the Relative Complexation Energies of Zn-Dependent Phosphomannose Isomerase with D-Mannose 6-Phosphate Surrogates. *J. Phys. Chem. B* **2011,** *115* (25), 8304-8316.

8. Sakharov, D. V.; Lim, C., Zn protein simulations including charge transfer and local polarization effects. *J. Am. Chem. Soc* **2005,** *127* (13), 4921-9.

9. Sakharov, D. V.; Lim, C., Force Fields Including Charge Transfer and Local Polarization Effects: Application to Proteins Containing Multi/Heavy Metal Ions. *J. Comput. Chem.* **2009,** *30* (2), 191-202.

10. Wu, J. C.; Piquemal, J. P.; Chaudret, R.; Reinhardt, P.; Ren, P. Y., Polarizable Molecular Dynamics Simulation of Zn(II) in Water Using the AMOEBA Force Field. *J. Chem. Theory Comput* **2010,** *6* (7), 2059-2070.

11. Wu, R.; Lu, Z.; Cao, Z.; Zhang, Y., A Transferable Non-bonded Pairwise Force Field to Model Zinc Interactions in Metalloproteins. *J. Chem. Theory Comput* **2011,** *7* (2), 433-443.

12. Li, P.; Merz, K. M., Jr., Taking into Account the Ion-induced Dipole Interaction in the Nonbonded Model of Ions. *J. Chem. Theory Comput* **2014,** *10* (1), 289-297.

13. Yang, Z. Z.; Cui, B. Q., Atomic charge calculation of metallobiomolecules in terms of the ABEEM method. *J. Chem. Theory Comput* **2007,** *3* (4), 1561-1568.

14. Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell, A. D., An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016,** *116* (9), 4983-5013.

15. Soniat, M.; Hartman, L.; Rick, S. W., Charge Transfer Models of Zinc and Magnesium in Water. *J. Chem. Theory Comput* **2015,** *11* (4), 1658-67.

16. Zhu, T.; Xiao, X.; Ji, C.; Zhang, J. Z., A New Quantum Calibrated Force Field for Zinc-Protein Complex. *J. Chem. Theory Comput* **2013,** *9* (3), 1788-98.

17. Li, P. F.; Merz, K. M., Metal Ion Modeling Using Classical Mechanics. *Chem. Rev.* **2017,** *117* (3), 1564-1686.

18. Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Muller, K. R., Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput* **2013,** *9* (8), 3404-19.

19. Behler, J.; Parrinello, M., Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007,** *98* (14), 146401.

20. Behler, J., Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **2011,** *13* (40), 17930-55.

21. Behler, J., Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011,** *134* (7), 074106.

22. Morawietz, T.; Sharma, V.; Behler, J., A neural network potential-energy surface for the water dimer based on environment-dependent atomic energies and charges. *J. Chem. Phys.* **2012,** *136* (6), 064103.

23. Behler, J., First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chem. Int. Ed. Engl.* **2017,** *56* (42), 12828-12840.

24. Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schutt, K. T.; Muller, K. R., Machine

learning of accurate energy-conserving molecular force fields. *Sci. Adv* **2017,** *3* (5), e1603015.

25. Schutt, K. T.; Arbabzadah, F.; Chmiela, S.; Muller, K. R.; Tkatchenko, A., Quantum-chemical insights from deep tensor neural networks. *Nat. Commun* **2017,** *8*, 13890.

26. Sauceda, H. E.; Chmiela, S.; Poltavsky, I.; Muller, K. R.; Tkatchenko, A., Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *J. Chem. Phys.* **2019,** *150* (11), 114102.

27. Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W., Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018,** *120* (14), 143001.

28. Wang, H.; Yang, W., Force Field for Water Based on Neural Network. *J. Phys. Chem. Lett* **2018,** *9* (12), 3232-3240.

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

30. Yao, K.; Herr, J. E.; Toth, D. W.; McKintyre, R.; Parkhill, J., The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci* **2018,** *9* (8), 2261-2269.

31. Yao, K.; Herr, J. E.; Parkhill, J., The many-body expansion combined with neural networks. *J. Chem. Phys.* **2017,** *146* (1), 014106.