The Journal of Chemical Physics

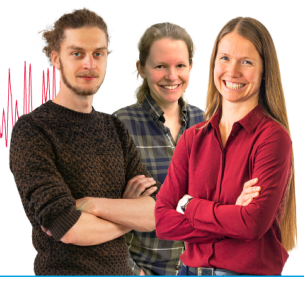
RESEARCH ARTICLE | JULY 17 2023

Improving machine learning force fields for molecular dynamics simulations with fine-grained force metrics

Special Collection: 2023 JCP Emerging Investigators Special Collection

Zun Wang ; Hongfei Wu ; Lixin Sun ; Xinheng He ; Zhirong Liu ; Bin Shao ; Tong Wang ; Tie-Yan Liu

J. Chem. Phys. 159, 035101 (2023)
https://doi.org/10.1063/5.0147023



The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

Improving machine learning force fields for molecular dynamics simulations with fine-grained force metrics

Cite as: J. Chem. Phys. 159, 035101 (2023); doi: 10.1063/5.0147023
Submitted: 18 February 2023 • Accepted: 12 May 2023 •
Published Online: 17 July 2023

Zun Wang,[1] Hongfei Wu,[1,2] Lixin Sun,[3] Xinheng He,[1] Zhirong Liu,[2] Bin Shao,[1,a)] Tong Wang,[1,b)] and Tie-Yan Liu[1]

AFFILIATIONS

[1] Microsoft Research AI4Science, Beijing 100084, China
[2] College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China
[3] Microsoft Research AI4Science, Cambridge CB1 2FB, United Kingdom

Note: This paper is part of the 2023 JCP Emerging Investigators Special Collection.
[a)] **Electronic mail:** binshao@microsoft.com
[b)] **Author to whom correspondence should be addressed:** watong@microsoft.com

ABSTRACT

Machine learning force fields (MLFFs) have gained popularity in recent years as they provide a cost-effective alternative to *ab initio* molecular dynamics (MD) simulations. Despite a small error on the test set, MLFFs inherently suffer from generalization and robustness issues during MD simulations. To alleviate these issues, we propose global force metrics and fine-grained metrics from element and conformation aspects to systematically measure MLFFs for every atom and every conformation of molecules. We selected three state-of-the-art MLFFs (ET, NequIP, and ViSNet) and comprehensively evaluated on aspirin, Ac-Ala3-NHMe, and Chignolin MD datasets with the number of atoms ranging from 21 to 166. Driven by the trained MLFFs on these molecules, we performed MD simulations from different initial conformations, analyzed the relationship between the force metrics and the stability of simulation trajectories, and investigated the reason for collapsed simulations. Finally, the performance of MLFFs and the stability of MD simulations can be further improved guided by the proposed force metrics for model training, specifically training MLFF models with these force metrics as loss functions, fine-tuning by reweighting samples in the original dataset, and continued training by recruiting additional unexplored data.

I. INTRODUCTION

Machine learning force fields (MLFFs)[1–4] are revolutionizing the realm of classical molecular simulations (MD)[5] in chemistry, biology, physics, and material science. Compared to *ab initio* MD (AIMD) simulations,[6] classical MD simulations are much faster and more scalable because they model the interatomic interactions with empirical force fields. In the past, classical MD simulations were limited by the accuracy and reliability of force fields that were trained to match a small number of experimental observations. Nowadays, MLFFs greatly enhance the accuracy and reliability by learning the interatomic interactions from *ab initio* and experimental data,[1,7,8] while keeping the low cost of empirical force fields. Geometric neural network potentials[9–18] are an example of recent MLFFs; they represent the atoms in a system as nodes, and the chemical bonds between atoms as edges in a graph, and incorporate physical symmetries, which has been proven to be data-efficient and effective in modeling many-body interactions. Therefore, MD simulations driven by MLFFs can accurately model atomic evolution at large system-scale and long time-scale phenomena and greatly enhance our understanding of chemical and biological systems.

However, MLFFs inherently suffer from generalization and robustness issues. For example, while sampling different protein conformations during MD simulations, MLFFs can encounter numerical instabilities or errors and lead to a simulation collapse. A common way of evaluating the performance of MLFFs is to compute their prediction accuracy on a test set. Metrics such as the mean absolute error (MAE) and root mean square error (RMSE) are often

used as loss functions, references for learning rate scheduling, and model selections. Some recent studies found that a simple MAE or RMSE metric on a test set is not sufficient to evaluate the generalization and robustness of MLFFs. For example, Fu *et al.*[19] and Morrow *et al.*[20] indicated that the simulations sometimes crashed, although the mean absolute error of force on the test set was small. Stocker *et al.*[21] also found that a small error on the test set cannot guarantee the simulation stability and proposed that recruiting sufficiently large training samples may improve the robustness of MLFFs. Obviously, a gap exists between the simple test-set-MAE/RMSE and the actual generalization and robustness of MLFFs while being deployed in real MD simulations.

To perform a stable long-time MD simulation, MLFFs have to make perfect force predictions for physically reasonable conformations because time integration in MD solely depends on the atomic forces. Thus, a good MLFF should ensure the following quantities being small enough: (1) the max values of force error for each atom, (2) the force error for each atom, and (3) the force error for physically reasonable conformations in the conformational space. Unfortunately, while training the MLFF model, only MAE or RMSE of forces is commonly used as the evaluation metric, which only reflects the average accuracy but understates the impact of the worst prediction case and specific evaluations on atoms. Moreover, different kinds of conformations may have different numbers of samples in the training data. The majority of conformations in the training data may bias the model, leading to poor predictions for the minority or even unrepresented conformations in the training data. For the reasons above, only evaluating test-set MAEs is inadequate and sometimes misleading. Although some studies evaluated MLFFs by performing MD simulations,[19,20] it may take a long time to run costly MD simulations before detecting errors and abnormal conformations, which hinders the MLFFs to be directly evaluated in the first place. Therefore, comprehensive metrics to make direct evaluations are required for MLFFs.

The four metrics above evaluate the global performance of the MLFFs on the MD dataset by summarizing the predictions of all samples. Thus, they are all global force metrics.

As shown in Table I, the three models show the same trend on four global force metrics for all molecules. $F_{\text{max-err}}$ value is typically an order of magnitude larger than F_{mae} when comparing among the same model and the molecule. Furthermore, the $\text{FNM}_{\text{max-err}}$ is much larger than the FNM_{mae}, implying a significant ratio of outliers.

In this study, we propose three kinds of force metrics from both global and fine-grained aspects to comprehensively evaluate the performance of MLFFs: (1) global force metrics, i.e., MAE, max error of force; (2) conformation-based force metrics, i.e., the force metrics for each kind of clusters of conformations; and (3) element-based force metrics, i.e., the force metrics for each kind of elements. Three state-of-the-art MLFFs, TorchMD-NET (ET),[16] Neural Equivariant Interatomic Potentials (NequIPs),[17] and Vector–Scalar interactive graph neural Network (ViSNet),[18] were selected for evaluation. We evaluated such MLFFs performance on three differently sized organic and biomolecular MD datasets, including the 21-atom aspirin in the revised MD17 dataset,[22] the 42-atom Ac-Ala3-NHMe in the MD22 dataset,[23] and the 166-atom protein Chignolin in our own dataset. For each molecule, MD simulations were performed at different initial conformations and velocities. With thorough analysis of the proposed force metrics and the MD simulation stability, we elucidated the underlying cause of collapsed simulations. Finally, guided by the proposed force metrics, we performed fine-tuning (reweighting samples in the training set) and continued learning (adding more samples with unrepresented conformations in the training set) to improve the performance of MLFFs and bridge the gap between ML potential test-set accuracy and MD simulations.

II. METRICS AND EVALUATIONS

We trained and evaluated three state-of-the-art MLFFs, i.e., ET, NequIP, and ViSNet for aspirin (21 atoms), Ac-Ala3-NHMe (42 atoms), and Chignolin (166 atoms) MD datasets [Figs. 1(a)–1(c)]. These three MLFF models are cutting-edge

TABLE I. Evaluation of MLFFs on global force metrics for aspirin, Ac-Ala3-NHMe, and Chignolin, respectively. F_{mae} and $F_{\text{max-err}}$ represent the mean absolute error and maximum absolute error of forces, respectively. The value in bold indicates the best performance on the corresponding metric. F_{mae} and $F_{\text{max-err}}$ are reported in the unit of kcal/(mol Å).

Molecule	No. of atoms	No. of dataset	Model	F_{mae}	FNM_{mae}	$F_{\text{max-err}}$	$\text{FNM}_{\text{max-err}}$
Aspirin	21	100 000	ET	0.2506	0.1264	1.2495	0.5553
			NequIP	0.1823	0.0902	0.9025	0.3826
			ViSNet	**0.1516**	**0.0696**	**0.7728**	**0.2840**
Ac-Ala3-NHMe	42	85 109	ET	0.1285	0.0979	0.7713	0.5132
			NequIP	0.1523	0.1150	0.8426	0.5217
			ViSNet	**0.0791**	**0.0600**	**0.4825**	**0.3205**
Chignolin	166	9543	ET	0.8421	0.5458	9.4076	4.9103
			NequIP	**0.4783**	**0.3074**	**4.3508**	**2.7698**
			ViSNet	0.6212	0.4073	4.8349	2.8204

equivariant graph neural networks for MD simulations and have achieved outstanding performance when evaluated on various publicly available datasets.

The three molecules are chosen such that MLFFs are tested on different molecule sizes. The 21-atom aspirin MD dataset consists of more than 100 thousands of samples, and the 42-atom Ac-Ala3-NHMe MD dataset contains 85, 109 samples. Notably, to evaluate model performance for proteins, we built the 166-atom Chignolin MD dataset by performing replica exchange MD to sample the whole conformational space and then quantum simulation to accurately calculate atomic forces at the density functional theory (DFT) level. The protein Chignolin dataset contains 9,543 samples (see more details in the supplementary material and Fig. S1). These models were trained with the same split of dataset, i.e., 950 training instances and 50 validation instances for aspirin, 6,000 training instances and 500 validation instances for Ac-Ala3-NHMe, and 80% training instances and 10% validation instances for Chignolin.

A. Force evaluation metrics

1. Global force metrics

The most commonly used force evaluation metric is the mean absolute error (F_{mae}) of the MLFF-predicted forces.

$$F_{\text{mae}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{3N_a}\sum_{i=1}^{N_a}\sum_{\alpha=x,\ y,\ z}\left|\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}\right|, \tag{1}$$

where N and N_a denote the number of samples in the dataset and the number of atoms in a configuration, respectively; F is the predicted forces, and $\hat{F}$ is the ground truth.

But in an MD simulation, a single bad prediction can cause the simulation to collapse, often due to an abnormal large force exerted on an atom. Therefore, the max force error ($F_{\text{max-err}}$) is a significant global force metric to evaluate the performance of the MLFFs,

$$F_{\text{max-err}} = \frac{1}{N}\sum_{n=1}^{N}\max_{\substack{i=1,\ldots,N_a,\\ \alpha=x,\ y,\ z}}\left|\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}\right|. \tag{2}$$

However, the absolute error provides a straightforward metric for evaluating the performance of MLFFs, the acceleration of error caused by the same force error on different atoms can be different, and the absolute error cannot reflect the impact of the simulation speed. For example, a force error of the same magnitude may impact the velocity of carbon atoms slightly, but the impact on the velocity of hydrogen atoms may significantly reduce the stability of the simulation. Thus, considering both absolute force error and force error normalized by mass can more comprehensively evaluate the global performance of MLFFs. We define the mean absolute force error normalized by mass (FNM_{mae}) and the max force error normalized by mass ($\text{FNM}_{\text{max-err}}$) as follows:

$$\text{FNM}_{\text{mae}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{3N_a}\sum_{i=1}^{N_a}\sum_{\alpha=x,\ y,\ z}\left|\frac{\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}}{u_i^{(n)}}\right|, \tag{3}$$

$$\text{FNM}_{\text{max-err}} = \frac{1}{N}\sum_{n=1}^{N}\max_{\substack{i=1,\ldots,N_a,\\ \alpha=x,\ y,\ z}}\left|\frac{\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}}{u_i^{(n)}}\right|, \tag{4}$$

where $u_i^{(n)}$ denotes the relative atomic mass of atom i in conformer n.

2. Conformation-based force metrics

Although the global force metrics bring insights into the performance of the MLFFs on the test set, they cannot evaluate their performance on both different elements and different conformations. Take an extreme example, if all samples in the training set and 99% of the samples in the test set are identical, while 1% of the samples in the test set are quite different from these, the global force metrics can easily reach a very low global error for the imbalance dataset. However, the model would perform poorly in the samples of the 1% test set actually. Thus, it is insufficient to assess the performance of MLFFs merely depending on global force metrics.

To alleviate this issue, we propose force metrics on clusters of conformations and then analyze the absolute errors and FNM errors within each cluster. We employed the K-means clustering algorithm[24] to categorize the conformations into multiple clusters for three molecules, respectively. These clusters were then visualized in a 2D space with Principal Component Analysis (PCA).[25] The number of the clusters was determined to be 3, 4, and 5 for aspirin, Ac-Ala3-NHMe, and Chignolin, respectively.

Based on these clusters, conformation-based force metrics are defined as follows:

$$F_{\text{mae}}^{(k)} = \frac{1}{N^{(k)}}\sum_{n\in\text{Clus}_k}\frac{1}{3N_a}\sum_{i=1}^{N_a}\sum_{\alpha=x,\ y,\ z}\left|\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}\right|, \tag{5}$$

$$F_{\text{max-err}}^{(k)} = \frac{1}{N^{(k)}}\sum_{n\in\text{Clus}_k}\max_{\substack{i=1,\ldots,N_a,\\ \alpha=x,\ y,\ z}}\left|\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}\right|, \tag{6}$$

$$\text{FNM}_{\text{mae}}^{(k)} = \frac{1}{N^{(k)}}\sum_{n\in\text{Clus}_k}\frac{1}{3N_a}\sum_{i=1}^{N_a}\sum_{\alpha=x,\ y,\ z}\left|\frac{\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}}{u_i^{(n)}}\right|, \tag{7}$$

$$\text{FNM}_{\text{max-err}}^{(k)} = \frac{1}{N^{(k)}}\sum_{n\in\text{Clus}_k}\max_{\substack{i=1,\ldots,N_a,\\ \alpha=x,\ y,\ z}}\left|\frac{\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}}{u_i^{(n)}}\right|, \tag{8}$$

where Clus_k denotes the k-th cluster and $N^k = |\text{Clus}_k|$ represents the number of data points in the k-th cluster.

Furthermore, we propose the ratio of bad points (RBPs) to quantify the extent of outliers in a cluster,

$$\text{RBP}^{(k)} = \frac{\left|\{n|n \in \text{Clus}_k \wedge F_{\text{mae}}^{(n)} > F_{\text{mae}}\}\right|}{|\{n|n \in \text{Clus}_k\}|}. \tag{9}$$

This RBP metric provides a unique viewpoint on how well the model performs for each cluster. A good MLFF is supposed to have balanced prediction accuracies and, thus, exhibits comparable RBP values among all clusters.

As shown in Figs. 1(d)–1(f), the configurations of aspirin and Chignolin are less continuous in the first two principal components (2D-PC) space than those of Ac-Ala3-NHMe, indicating a more complex manifold of training data.

Unlike the global force metrics (Table I), where all four metrics are positively correlated with each other, the conformation-based metrics have more interesting trends [Figs. 1(g)–1(l)]. The

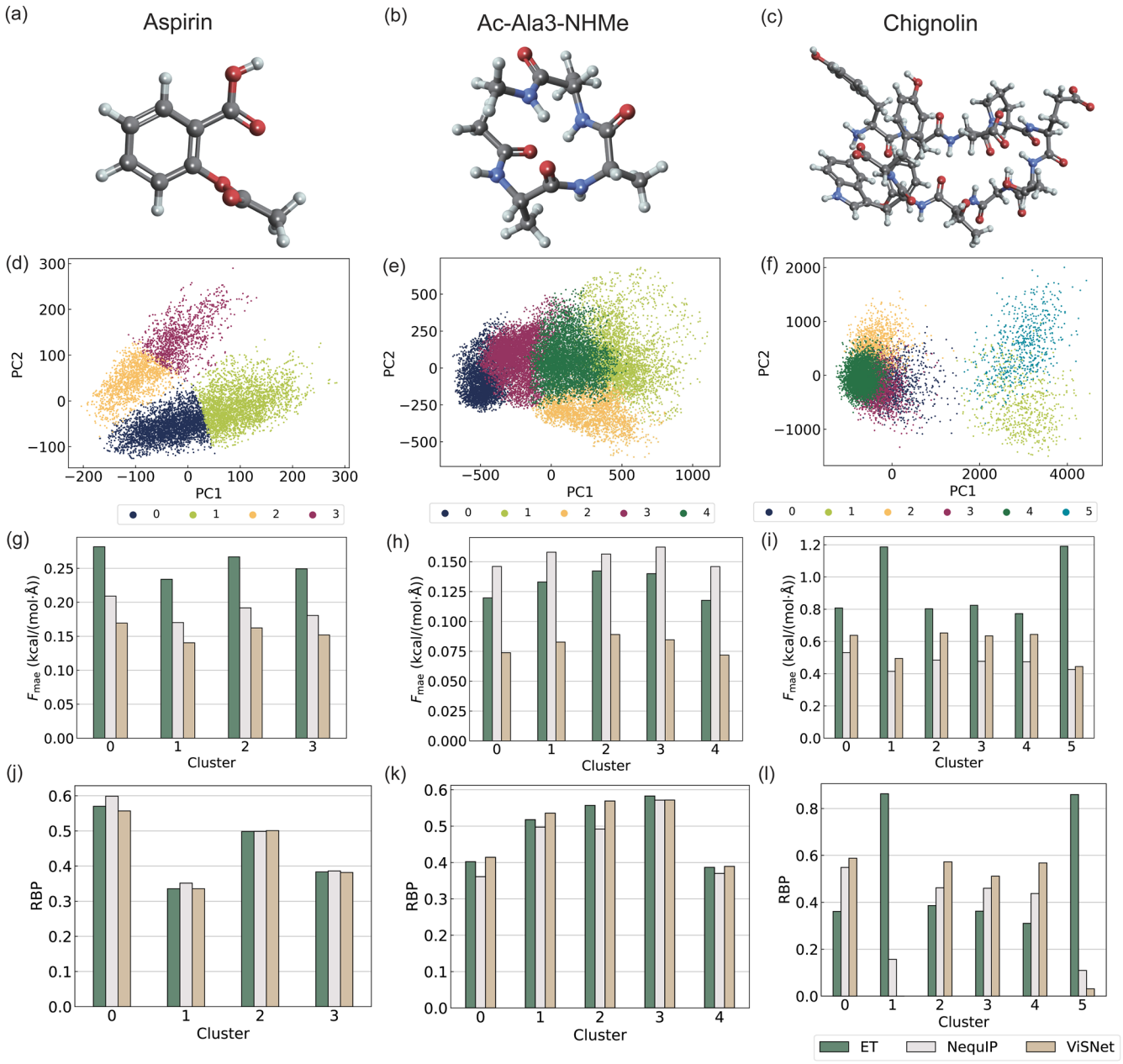


FIG. 1. Evaluation of MLFFs on conformation-based force metrics for aspirin, Ac-Ala3-NHMe, and Chignolin, respectively. The molecular structures of aspirin (a), Ac-Ala3-NHME (b), and Chignolin (c). The visualization of molecular conformations projected on the first two PC spaces for aspirin (d), Ac-Ala3-NHMe (e), and Chignolin (f). Conformations are clustered using K-means and projected into a 2D space using PCA. The different clusters are drawn with different colors. (g)–(i) Evaluation of the mean absolute errors of force for different clusters of the molecules. (j)–(l) Evaluation of the ratio of bad points (RBPs) for different clusters of the molecules.

conformation-based FNM errors for a cluster do not always positively correlate with the absolute error, as the ground truth of forces in different clusters can vary. The $F_{\text{mae}}^{(k)}$ values of the three models exhibit a similar trend for aspirin and Ac-Ala3-NHMe but not for Chignolin. All models achieved the lowest $F_{\text{mae}}^{(k)}$ on cluster 0 of aspirin and on cluster 4 of Ac-Ala3-NHMe. However, for Chignolin, ET made poor performance on clusters 1 and 5, while NequIP and ViSNet performed well on the two clusters.

TABLE II. Evaluation of MLFFs on element-based force metrics for aspirin, Ac-Ala3-NHMe, and Chignolin, respectively. The value in bold indicates the best performance on the corresponding metric among different models. F_{mae} and $F_{max\text{-}err}$ are reported in the unit of kcal/(mol Å).

		Hydrogen (H)			Carbon (C)			Oxygen (O)			Nitrogen (N)		
Molecule	Metrics	ET	NequIP	ViSNet	ET	NequIP	ViSNet	ET	NequIP	ViSNet	ET	NequIP	ViSNet
Aspirin	F_{mae}	0.1245	0.0883	**0.0656**	0.3447	0.2448	**0.2134**	0.2912	0.2295	**0.1847**	···	···	···
	FNM_{mae}	0.1245	0.0883	**0.0656**	0.0287	0.0204	**0.0178**	0.0182	0.0143	**0.0115**	···	···	···
	$F_{max\text{-}err}$	0.4665	0.3209	**0.2381**	1.1969	0.8628	**0.7480**	0.8296	0.6472	**0.5197**	···	···	···
	$FNM_{max\text{-}err}$	0.4665	0.3209	**0.2381**	0.0997	0.0719	**0.0623**	0.0518	0.0404	**0.0325**	···	···	···
Ac-Ala3-NHMe	F_{mae}	0.0807	0.0946	**0.0494**	0.1833	0.2182	**0.1137**	0.1423	0.1855	**0.0884**	0.2135	0.2385	**0.1292**
	FNM_{mae}	0.0807	0.0946	**0.0494**	0.0153	0.0182	**0.0095**	0.0089	0.0116	**0.0055**	0.0152	0.0170	**0.0092**
	$F_{max\text{-}err}$	0.4328	0.4421	**0.2708**	0.6842	0.7793	**0.4305**	0.3874	0.4915	**0.2418**	0.5940	0.6278	**0.3604**
	$FNM_{max\text{-}err}$	0.4328	0.4421	**0.2708**	0.0570	0.0649	**0.0359**	0.0242	0.0307	**0.0151**	0.0424	0.0448	**0.0257**
Chignolin	F_{mae}	0.5117	**0.2880**	0.3838	1.0945	**0.5768**	0.7702	1.0144	**0.7234**	0.8467	1.2994	**0.7402**	0.9472
	FNM_{mae}	0.5117	**0.2880**	0.3838	0.0912	**0.0481**	0.0642	0.0634	**0.0452**	0.0529	0.0928	**0.0529**	0.0677
	$F_{max\text{-}err}$	4.1458	**2.3468**	2.3989	8.8916	**3.8011**	4.5203	5.5665	**3.5319**	3.6970	5.3953	**2.5375**	3.1729
	$FNM_{max\text{-}err}$	4.1458	**2.3468**	2.3989	0.7410	**0.3168**	0.3767	0.3479	**0.2207**	0.2311	0.3854	**0.1812**	0.2266

For aspirin and Ac-Ala3-NHMe, all three models show comparable RBP among different clusters, while the differences among clusters of Chignolin became large. A lower $F_{mae}^{(k)}$ does not necessarily correlate with a lower RBP. In addition, it is surprising that the RBP for the ViSNet model on the Chignolin dataset for cluster 1 is 0.

These observations highlight the importance of conformation-based metrics in providing more elaborate measures of MLFFs' performance among different clusters in the conformational space.

Further details of the clustering and the FNM errors of the cluster-related metrics can be found in Figs. S2–S4 and Tables S4–S6 of the supplementary material (SM).

3. Element-based force metrics

In addition, we proposed element-based force metrics, taking inspiration from NequIP.[17] Similar to conformation-based metrics, element-based metrics for element S are defined as follows:

$$F_{mae}^{(S)} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{3N_a^{(S)}}\sum_{i\in S}\sum_{\alpha=x,\,y,\,z}\left|\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}\right|, \tag{10}$$

$$F_{max-err}^{(S)} = \frac{1}{N}\sum_{n=1}^{N}\max_{\substack{i=1,\ldots,N_a^{(S)},\\ \alpha=x,\,y,\,z}}\left|\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}\right|, \tag{11}$$

$$FNM_{mae}^{(S)} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{3N_a^{(S)}}\sum_{i\in S}\sum_{\alpha=x,\,y,\,z}\left|\frac{\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}}{u_i^{(n)}}\right|, \tag{12}$$

$$FNM_{max-err}^{(S)} = \frac{1}{N}\sum_{n=1}^{N}\max_{\substack{i=1,\ldots,N_a^{(S)},\\ \alpha=x,\,y,\,z}}\left|\frac{\hat{F}_{i\alpha}^{(n)} - F_{i\alpha}^{(n)}}{u_i^{(n)}}\right|, \tag{13}$$

where $N_a^{(S)} = |S|$ denotes the number of atoms belonging to element S in a configuration.

As shown in Table II, the absolute error of element-based metrics for heavy atoms is higher than that for hydrogen atoms. However, the results for element-based errors normalized by relative atomic mass may differ. The FNM_{mae} and $FNM_{max\text{-}err}$ values of hydrogen atoms are significantly larger compared to those of heavy atoms in three molecules, which indicates the velocity of hydrogen atoms in simulations may be more unstable. These metrics offer a more reliable evaluation of the performance of MLFFs than F_{mae}.

B. Comprehensive validations via MD simulations

Directly measuring the stability and accuracy of MD simulations can offer the most accurate metrics for MLFFs' performance in MD simulations. This metric is often not used due to the high computational cost to obtain AIMD ground truth. Here, we performed an MLFF-driven MD of 1 ns for aspirin and Ac-Ala3-NHMe, and 0.01 ns for Chignolin. All simulations were run with a Berendsen NVT thermostat at 500 K and a time step of 1 fs. For each molecule, 30 different initial structures were randomly selected from the original MD datasets.

1. Stability and sampling efficiency

When evaluating MLFF-driven MD simulations, it is important to ensure the stability of the simulations. It is meaningless to evaluate whether a simulation achieves equilibrium with sufficient sampling and reasonable physical observations if the simulation is unstable.

We followed the definition of stability in Ref. 19 that when any bond length in the current frame is beyond 0.3 Å compared with the equilibrium bond length, the conformation is recognized as an unstable one. For all MLFFs trained in this study, the simulations for aspirin and Ac-Ala3-NHMe were stable, while the simulations for Chignolin, the largest molecular system with the smallest size of dataset, were always unstable. Interestingly, as shown in Fig. 2, although NequIP achieved the best force

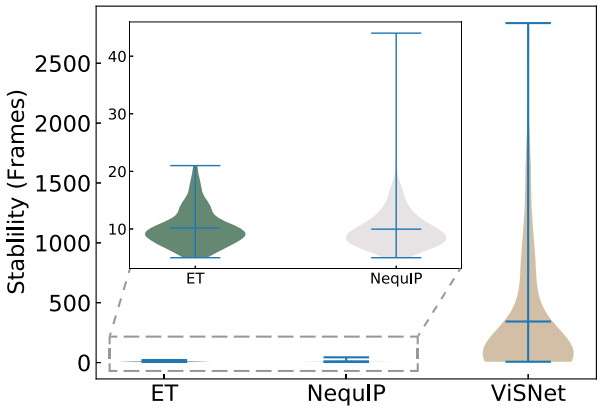


FIG. 2. Analysis of the stability of MD simulations for Chignolin driven by MLFFs. The inset provides a closer look at the results of the ET and NequIP. The shaded area in the violin plots shows the distribution of the number of frames before collapse.

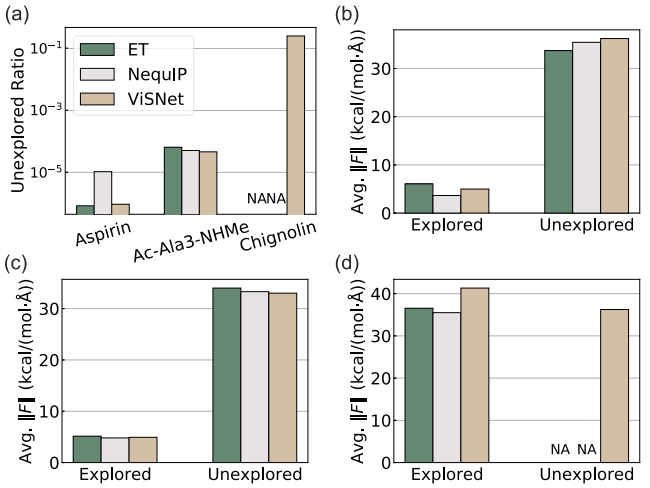


FIG. 3. Analysis of conformation exploration during MD simulations for aspirin, Ac-Ala3-NHMe, and Chignolin driven by MLFFs. (a) The proportion of sampled conformations during simulation that locate in unexplored regions of the original dataset. NA denotes not applicable, as the trajectories were terminated prematurely due to their rapid collapse. (b)–(d) Average norm of force in explored or unexplored regions for aspirin, Ac-Ala3-NHMe, and Chignolin, respectively.

metrics for Chignolin, the stability of the simulation trajectories driven by ViSNet was significantly stronger than the other two models.

We further analyzed the sampling efficiency by proposing the concepts of explored and unexplored ratios for the simulation trajectories. The radius of each cluster can be defined as the maximum distance from the centroid to the farthest point of the cluster. A conformation can be defined as an explored or unexplored one based on its distance to all the cluster centers. If the distance between a conformation and the closest cluster center is larger than the radius of this cluster, the conformation is defined as unexplored; otherwise, it is defined as explored and considered to be located within that closest cluster. To visualize all conformations sampled by the simulations, we mapped the data point to the 2D-PC space.

As shown in Fig. 3(a), unexplored conformations are present in aspirin and Ac-Ala3-NHMe MLFF-MD simulations, demonstrating the generalization capability of the MLFFs toward unseen conformations. On the other hand, the Chignolin trajectories driven by ET and NequIP never encounter unexplored region due to the instability of the trajectories at an early stage, leaving insufficient time to sample unexplored conformations. It is worth noticing that the force magnitudes in the unexplored areas are significantly larger than those in the explored regions for aspirin and Ac-Ala3-NHMe [Figs. 3(b) and 3(c)]. However, this trend is not observed in Chignolin [Fig. 3(d)]. The force magnitudes are similar for explored and unexplored areas of Chignolin, which can be attributed to the poor force accuracy and generalization for this molecule.

2. Accuracy on statistical properties for aspirin and Ac-Ala3-NHMe

For aspirin and Ac-Ala3-NHMe, we analyzed the statistical properties. Figures 4 and 5 show the root mean square deviation (RMSD), root mean square fluctuation (RMSF),[26] potential energy, and contact map of simulation trajectories of aspirin and Ac-Ala3-NHMe, respectively.

All three MLFF models are able to run fully stable MD simulations. The potential energies are stable during MD simulations with reasonable fluctuations [Figs. 4(e)–4(g) and 5(e)–5(g)]. RMSD measures the difference between a target structure and a reference structure. Its variation over time reflects how the structure and its components evolve in the MD simulations. All three models reach an equilibrium RMSD value of 0.10 Å within a 200 ps and have a consistent variation of 0.02 Å in the 1 ns second simulations [see Figs. 4(h)–4(j) and 5(h)–5(j)]. RMSF calculates the average deviation of each atom from its average position throughout the simulation. This measurement highlights which amino acids are more flexible and play the most significant role in molecular motion. As shown in Figs. 4(k)–4(m) and 5(k)–5(m), all three models share similar trend on which atoms having largest values of RMSF in aspirin and Ac-Ala3-NHMe, except for some subtle differences among the medium values ranges. Finally, the contact map displays the distance between each pair of atoms. The contact maps of conformations sampled by these models (represented by the lower triangular matrices) are in agreement with the original training dataset (represented by the upper triangular matrices) in Figs. 4(b)–4(d) and 5(b)–5(d).

The similar trends in RMSD and RMSF and the agreement of the contact map with AIMD data indicate that these three models perform similarly well on small and medium-sized molecules, despite the small differences in their force metrics.

3. Analysis of collapsed simulation trajectories for Chignolin

For Chignolin, we further investigated the detailed collapsed information and the reason why the simulation collapsed.

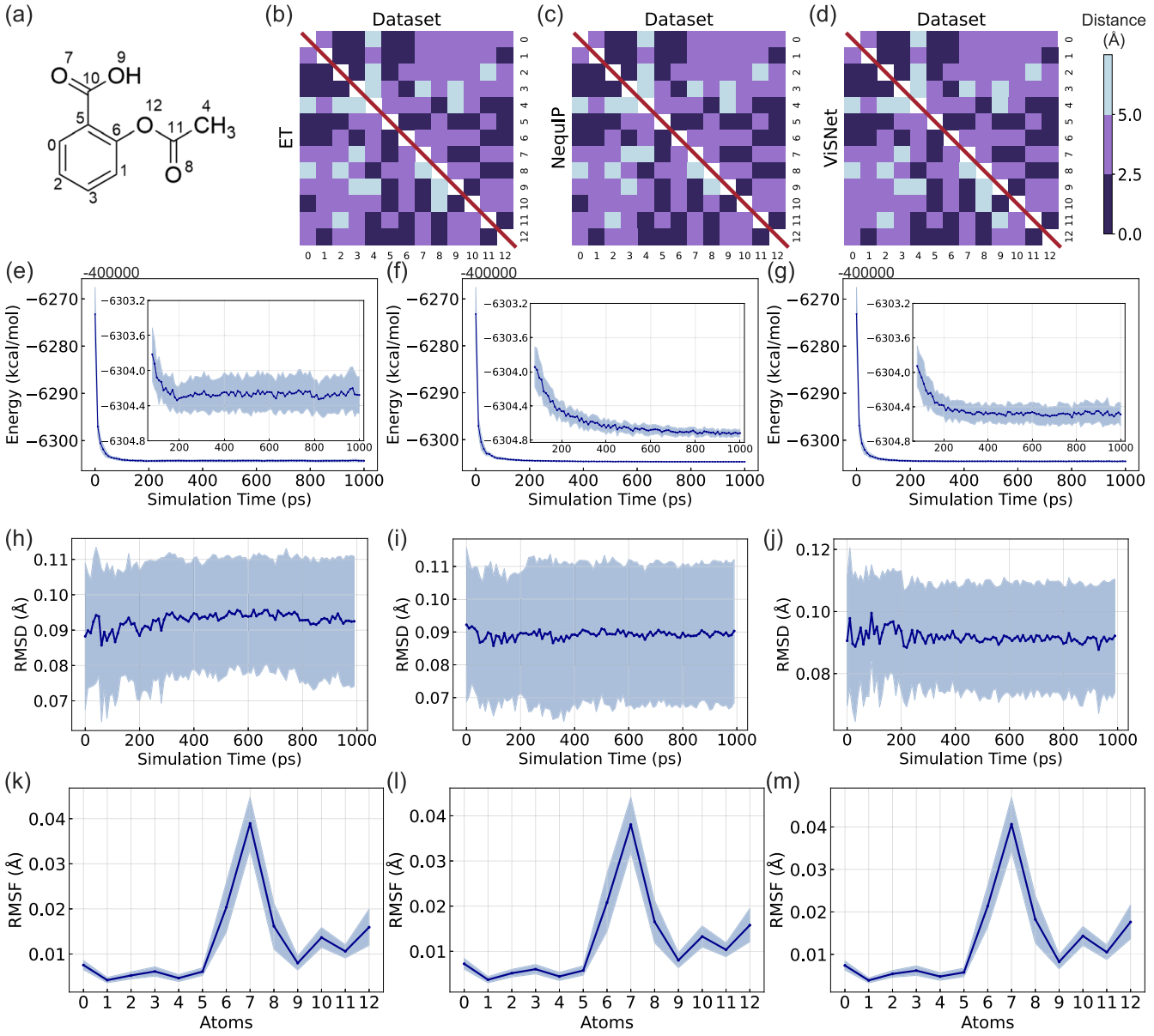


FIG. 4. Statistic analysis of MD simulations for aspirin driven by MLFFs. (a) The chemical structure of aspirin and the indices of heavy atoms. (b)–(d) The contact maps made by MLFFs (lower triangle) compared with those of the original dataset (upper triangle). (e)–(g) The potential energy of the simulation trajectories over time. (h)–(j) The root mean squared deviation (RMSD) of the simulation trajectories over time. (k)–(m) The root mean square fluctuations (RMSF) of the trajectories simulated by three MLFFs. From (e) to (m), the average values were calculated from multiple trajectories starting from different initial structures, and the shaded areas indicate the standard deviation range.

Figures 6(a), 6(c), and 6(e) show the fluctuation of maximum forces in each simulation frame. The instability can be seen as the forces increase, and this phenomenon is more pronounced for models with low F_{mae}, especially for ViSNet and NequIP. We infer that the model with lower F_{mae} may be more sensitive to abnormal structures. As shown in Figs. 6(b), 6(d), and 6(f), the bond length between the hydrogen atom and the corresponding heavy atom becomes abnormal, leading to the simulation collapse. Table II shows that the FNM errors of hydrogen atoms are larger than those for heavy atoms. This is because hydrogen atoms are lighter and more sensitive to slight deviations in the force value.

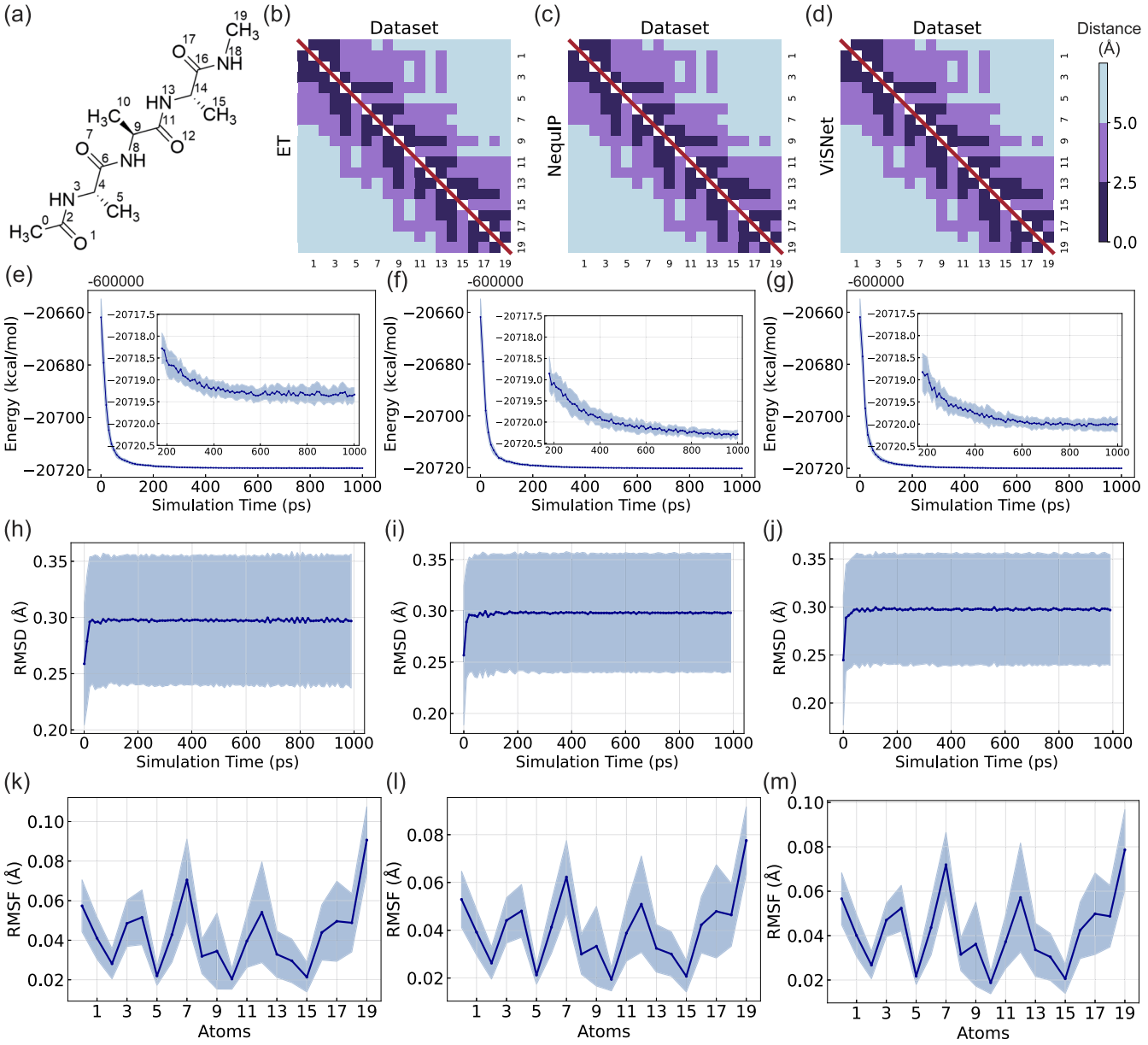


FIG. 5. Statistic analysis of MD simulations for Ac-Ala3-NHMe driven by MLFFs. (a) The chemical structure of aspirin and the indices of heavy atoms. (b)–(d) The contact maps made by MLFFs (lower triangle) compared with those of the original dataset (upper triangle). (e)–(g) The potential energy of the simulation trajectories over time. (h)–(j) The root mean squared deviation (RMSD) of the simulation trajectories over time. (k)–(m) The root mean square fluctuations (RMSF) of the trajectories simulated by three MLFFs. From (e) to (m), the average values were calculated from multiple trajectories starting from different initial structures, and the shaded areas indicate the standard deviation range.

We also analyzed the progression of errors leading up to the unstable conformations by calculating the ground truth *ab initio* forces for ten frames prior to simulation collapse (Fig. 7). Interestingly, the force errors, specifically, F_{mae} and $F_{\text{max-err}}$, accumulated over time for both ET and NequIP, while for ViSNet, the error remained relatively small in the previous frames but suddenly increased in the last several frames when unexplored regions were sampled.

III. GUIDING MLFF'S TRAINING BY FORCE METRICS

The previously discussed results shed light on areas where improvements can be made to enhance the performance of the

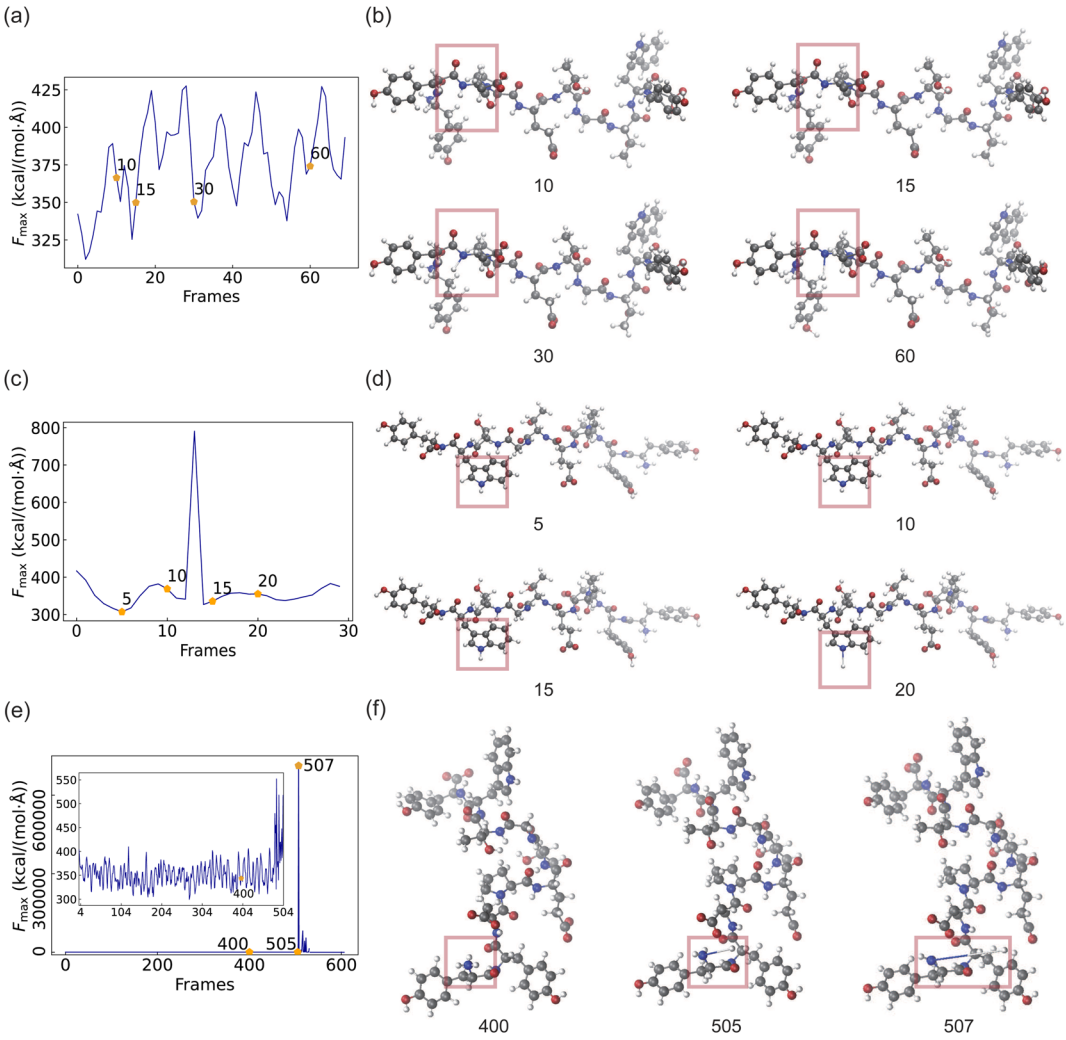


FIG. 6. Structural analysis of the collapsed simulation trajectories of Chignolin. (a) The variation of the maximum force during the simulation driven ET. (b) The conformations corresponding to the frames labeled with orange dots in (a). The transparent red boxes highlight the sub-structures that display a gradual appearance of abnormal chemical behavior over the simulation time. The bond length of N–H increases and leads to unstable simulations. (c) The variation of the maximum force during the simulation driven by NequIP. (d) The conformations corresponding to the frames labeled with orange dots in (c). The N–H bond is compressed in the tenth frame and then the hydrogen atom is pushed away by the repulsive force. (e) The variation of the maximum force during the simulation driven by ViSNet. (f) The conformations corresponding to the frames labeled with orange dots in (e). The N–H bond becomes abnormal abruptly.

MLFFs. To that end, we propose three approaches to enhance the MLFFs: training the model with the force metrics we proposed as the loss function, fine-tuning by reweighting samples in the original dataset, and continuing training by recruiting additional unexplored data.

For the first approach, we chose the $F_{\text{mae}}^{(k)}$ and the $F_{\text{mae}}^{(S)}$ as loss functions, respectively, to retrain the ViSNet model on the Chignolin dataset. As shown in Table III, ViSNet trained with these two loss functions outperforms the original model on Chignolin.

The second approach, continued learning, can be considered a form of active learning.[7,27,28] Furthermore, different from previous active learning approaches that usually add samples with high uncertainties to the training set,[29] we added samples that are located in the unexplored region of the original dataset. We sampled more conformations of Chignolin by running replica exchange MD simulations. 3,000 new unexplored conformations of Chignolin were added to the training set, and the model of ViSNet was continually trained (more details could be found in the supplementary material).

The third approach, fine-tuning, improved the original ViSNet model by reweighting all samples in the original dataset. The weights were updated to increase the importance of under-represented conformations, similar to methods that handle imbalanced datasets.[30,31] The unsampling or downsampling weights were determined by the

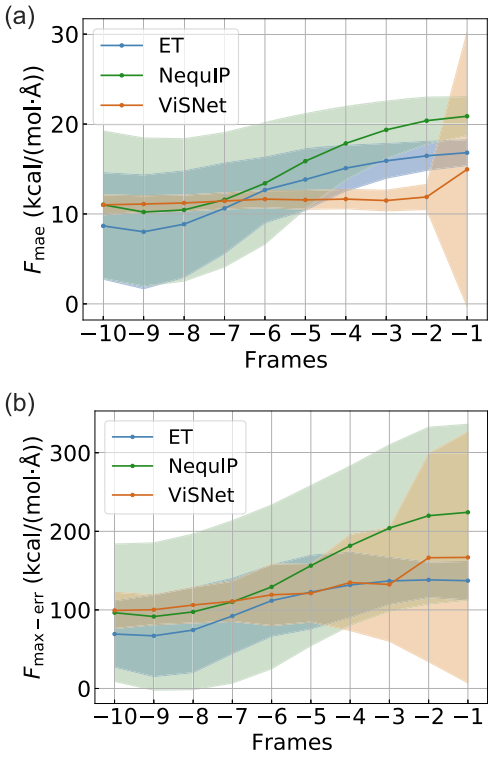


FIG. 7. The variation of the average value and standard deviation of F_{mae} (a) and $F_{max\text{-}err}$ (b) of the ten frames prior to the emergence of an abnormal bond length.

TABLE III. F_{mae}s of ViSNet for Chignolin with different loss functions. F_{mae}s are reported in the unit of kcal/(mol Å).

Loss function	F_{mae}
Global force metrics	0.6212
Conformation-based force metrics	0.3677
Element-based force metrics	0.3868

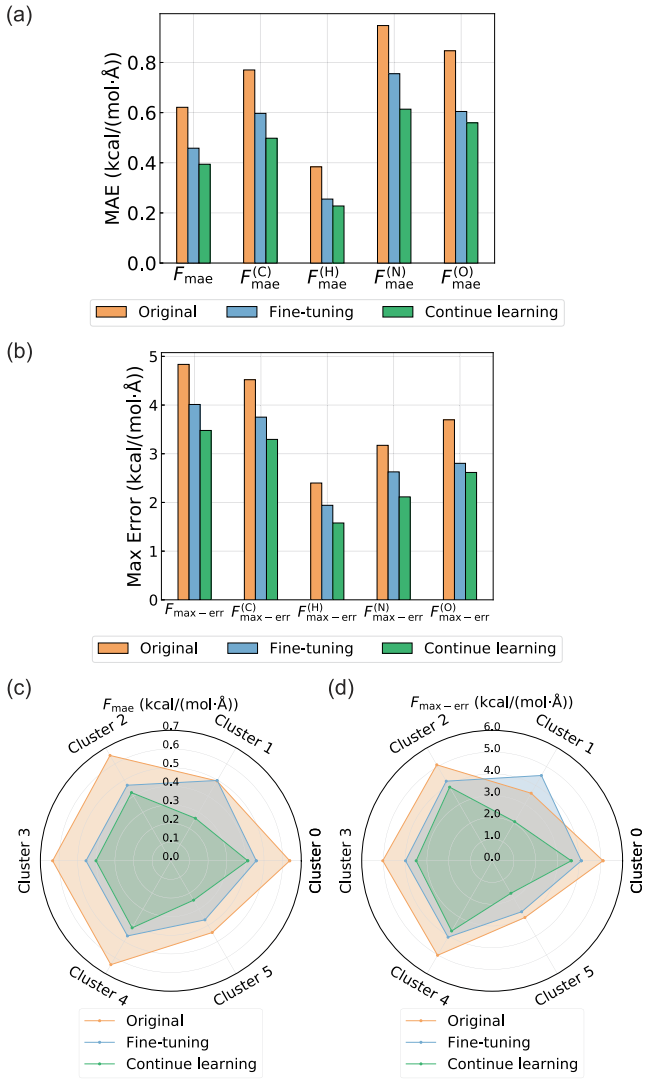


FIG. 8. Evaluation of the global and fine-grained force metrics after fine-tuning and continued learning on Chignolin, respectively. (a) The global and element-based F_{mae}; (b) The global and element-based $F_{max\text{-}err}$; (c) The F_{mae} for different clusters; and (d) The $F_{max\text{-}err}$ for different clusters.

RBP calculated for each cluster. If a cluster had a higher RBP compared to other clusters, larger weights were assigned to the samples in the cluster.

Figures 8(a) and 8(b) compared the global and element-based force metrics for the original ViSNet, ViSNet with continued learning and ViSNet with fine-tuning on the same original test set. Both continued learning and fine-tuning approaches achieved lower F_{mae}s and $F^{(S)}_{mae}$s values compared to the original model, indicating an improvement in performance.

The radar charts shown in Figs. 8(c) and 8(d) illustrate the significant improvements in conformation-based force metrics for the Chignolin with the two approaches. As we expected, the F_{mae} and $F_{max\text{-}err}$ values on different clusters derived from the continued learning model showed a similar shape on the radar charts to those of the original model but with smaller errors. As a contrast, the fine-tuned model exhibited different shapes in the radar charts. Furthermore, the performance improvement is much significant for the cluster that had a larger RBP value in the original model.

Furthermore, we assessed the stability of the simulations with the two kinds of updated models. We set the maximum simulation steps to 10 000 and run MD simulations from the same 30 initial conformations used in the original ViSNet. Five parallel simulations with random velocities allocated were run for each initial conformation. As shown in Fig. 9, the two kinds of updated models have made significant improvements to the stability of simulations, especially for the model with continuous learning. Notably, the model with continued learning finished 10 000 step for 9.3% of the simulation

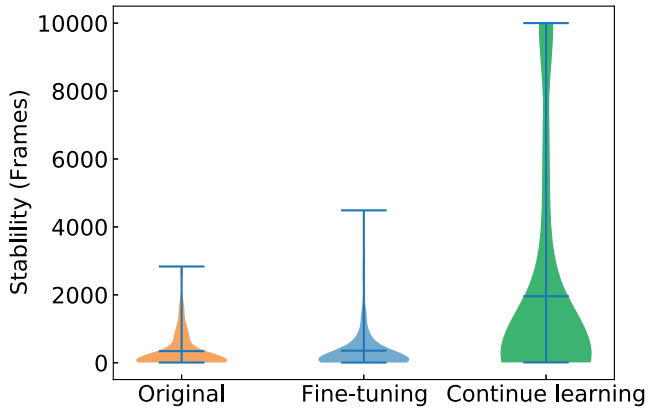


FIG. 9. Analysis of the stability of MD simulations for Chignolin driven by the original ViSNet, ViSNet with fine-tuning, and ViSNet with continued learning. The shaded area in the violin plots shows the distribution of the number of frames before collapse.

runs, which indicates that adding more data with novel conformations can significantly enhance the robustness and generalization of the MLFF for MD simulation.

IV. CONCLUSION

We show that an accurate force prediction by MLFFs for all kinds of atom types and all possible conformations plays a crucial role in its usefulness in MD simulations. To this end, we propose global force metrics and fine-grained metrics from element and conformation aspects to systematically measure MLFFs for every atom and every conformation of molecules. Such force metrics can directly examine MLFFs without running costly MD simulations, reducing the computational cost of MLFF evaluation.

We further analyzed the stability and sampling efficiency of MD simulations driven by MLFFs and detected the reason that led to simulation collapse.

Finally, guided by the proposed force metrics, we designed continued learning and fine-tuning approaches to improve MLFF's performance. Both methods are proven to be effective in improving the MD performance of the MLFFs. Our study highlights the potential and limitations of MLFFs for practical applications in MD simulations and sheds light on developing more robust and accurate MLFFs in the future.

SUPPLEMENTARY MATERIAL

The supplementary material includes details of datasets, training settings, and analysis methods and also contains more results from force metrics analysis.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Z.W. and H.W. contributed equally to this work.

T.W. led, conceived and designed the study. Z.W. and H.W. carried out model training, force metric evaluation, MD simulations and model fine-tuning and continued learning. L.S. carried out model training for NequIP. X.H. carried out construction and analysis for Chignolin dataset. Z.W. wrote the original manuscript. T.W., L.S., H.W., Z.L., B.S., and T. L. revised the manuscript. All authors approved the final manuscript.

Zun Wang: Data curation (equal); Formal analysis (lead); Methodology (equal); Resources (equal); Validation (lead); Visualization (equal); Writing – original draft (lead). **Hongfei Wu**: Data curation (supporting); Formal analysis (equal); Validation (equal); Visualization (equal); Writing – original draft (supporting). **Lixin Sun**: Formal analysis (supporting); Methodology (supporting); Writing – review & editing (supporting). **Xinheng He**: Data curation (equal). **Zhirong Liu**: Writing – review & editing (equal). **Bin Shao**: Writing – review & editing (supporting). **Tong Wang**: Conceptualization (lead); Formal analysis (equal); Investigation (lead); Methodology (lead); Project administration (lead); Resources (lead); Supervision (lead); Writing – review & editing (lead). **Tie-Yan Liu**: Writing – review & editing (supporting).

DATA AVAILABILITY

Aspirin and Ac-Ala3-NHMe MD datasets can be downloaded from https://figshare.com/articles/dataset/Revised_MD17_dataset_rMD17_/12672038 and http://www.sgdml.org/#datasets, respectively. Our designed Chignolin MD dataset is available at https://msracb.blob.core.windows.net/pub/Chig-MD-10K.zip.

REFERENCES

[1] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).

[2] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).

[3] L. Zhang, J. Han, H. Wang, W. Saidi, R. Car, and E. Weinan, "End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems," in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31.

[4] A. Kabylda, V. Vassilev-Galindo, S. Chmiela, I. Poltavsky, and A. Tkatchenko, "Towards linearly scaling and chemically accurate global machine learning force fields for large molecules," arXiv:2209.03985 (2022).

[5] B. J. Alder and T. E. Wainwright, "Studies in molecular dynamics. I. General method," J. Chem. Phys. **31**, 459–466 (1959).

[6] R. Car and M. Parrinello, "Unified approach for molecular dynamics and density-functional theory," Phys. Rev. Lett. **55**, 2471 (1985).

[7] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).

[8]R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B **99**, 014104 (2019).
[9]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8**, 13890 (2017).
[10]K. Schütt, P.-J. Kindermans, H. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2017), pp. 992–1002.
[11]K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).
[12]J. Gasteiger, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," in International Conference on Learning Representations, 2019.
[13]J. Klicpera, S. Giri, J. T. Margraf, and S. Günnemann, "Fast and uncertainty-aware directional message passing for non-equilibrium molecules," in NeurIPS-W, 2020.
[14]K. Schütt, O. Unke, and M. Gastegger, "Equivariant message passing for the prediction of tensorial properties and molecular spectra," in *International Conference on Machine Learning* (PMLR, 2021), pp. 9377–9388.
[15]J. Gasteiger, F. Becker, and S. Günnemann, "GemNet: Universal directional graph neural networks for molecules," in *Advances in Neural Information Processing Systems* (Curran Associates Inc., 2021), Vol. 34, pp. 6790–6802.
[16]P. Thölke and G. De Fabritiis, "TorchMD-NET: Equivariant transformers for neural network based molecular potentials," arXiv:2202.02541 (2022).
[17]S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," Nat. Commun. **13**, 2453 (2022).
[18]Y. Wang, S. Li, X. He, M. Li, Z. Wang, N. Zheng, B. Shao, T. Wang, and T.-Y. Liu, "ViSNet: A scalable and accurate geometric deep learning potential for molecular dynamics simulation," arXiv:2210.16518 (2022).
[19]X. Fu, Z. Wu, W. Wang, T. Xie, S. Keten, R. Gomez-Bombarelli, and T. Jaakkola, "Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations," arXiv:2210.07237 (2022).
[20]J. D. Morrow, J. L. Gardner, and V. L. Deringer, "How to validate machine-learned interatomic potentials," J. Chem. Phys. **158**, 121501 (2023).
[21]S. Stocker, J. Gasteiger, F. Becker, S. Günnemann, and J. T. Margraf, "How robust are modern graph neural network potentials in long and hot molecular dynamics simulations?," Mach. Learn.: Sci. Technol. **3**, 045010 (2022).
[22]A. S. Christensen and O. A. Von Lilienfeld, "On the role of gradients for machine learning of molecular energies and forces," Mach. Learn.: Sci. Technol. **1**, 045018 (2020).
[23]S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller, "Accurate global machine learning force fields for molecules with hundreds of atoms," Sci. Adv. **9**, eadf0873 (2023).
[24]J. MacQueen, "Classification and analysis of multivariate observations," in *5th Berkeley Symposium on Math. Statist. Probability* (University of California, Los Angeles, 1967), pp. 281–297.
[25]H. Hotelling, "Analysis of a complex of statistical variables into principal components," J. Educ. Psychol. **24**, 417 (1933).
[26]R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "MDTraj: A modern open library for the analysis of molecular dynamics trajectories," Biophys. J. **109**, 1528–1532 (2015).
[27]B. Settles, Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
[28]S. Fujikake, V. L. Deringer, T. H. Lee, M. Krynski, S. R. Elliott, and G. Csányi, "Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures," J. Chem. Phys. **148**, 241714 (2018).
[29]J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky, "On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events," npj Comput. Mater. **6**, 20 (2020).
[30]N. V. Chawla, "Data mining for imbalanced datasets: An overview," Data Min. Knowl. Discov. 875–886 (2009).
[31]H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng. **21**, 1263–1284 (2009).