Forces are not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations

Xiang Fu[*] MIT
Zhenghao Wu Northwestern University
Wujie Wang MIT
Tian Xie Microsoft Research
Sinan Keten Northwestern University
Rafael Gomez-Bombarelli MIT
Tommi Jaakkola MIT

Abstract

Molecular dynamics (MD) simulation techniques are widely used for various natural science applications. Increasingly, machine learning (ML) force field (FF) models begin to replace *ab-initio* simulations by predicting forces directly from atomic structures. Despite significant progress in this area, such techniques are primarily benchmarked by their force/energy prediction errors, even though the practical use case would be to produce realistic MD trajectories. We aim to fill this gap by introducing a novel benchmark suite for ML MD simulation. We curate representative MD systems, including water, organic molecules, peptide, and materials, and design evaluation metrics corresponding to the scientific objectives of respective systems. We benchmark a collection of state-of-the-art (SOTA) ML FF models and illustrate, in particular, how the commonly benchmarked force accuracy is not well aligned with relevant simulation metrics. We demonstrate when and how selected SOTA methods fail, along with offering directions for further improvement. Specifically, we identify stability as a key metric for ML models to improve. Our benchmark suite comes with a comprehensive open source codebase[2] for training and simulation with ML FFs to facilitate further work.

1 Introduction

Molecular Dynamics (MD) simulations provide atomistic insights into physical phenomena in materials and biological systems. Such simulations are typically based on force fields (FFs) that characterize the underlying potential energy surface (PES) of the system and then use Newtonian forces to simulate long trajectories [14]. The PES itself is challenging to compute and would ideally be done through quantum chemistry which is computationally expensive. Traditionally, the alternative has been parameterized force fields that are surrogate models built from empirically chosen functional forms [21]. Recently, machine learning (ML) force fields [59] have shown promise to accelerate MD simulations by orders of magnitude while being quantum chemically accurate. The evidence supporting the utility of ML FFs is often based on their accuracy in

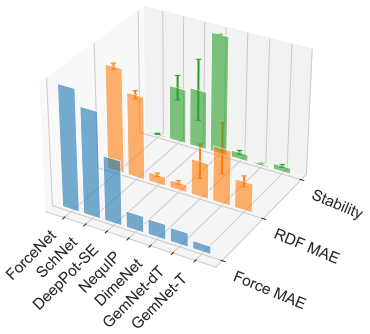

Figure 1: Results on water-10k.

[*]Correspondence to: Xiang Fu at xiangfu@mit.edu
[2]https://github.com/kyonofx/MDsim

NeurIPS 2022 AI for Science Workshop.

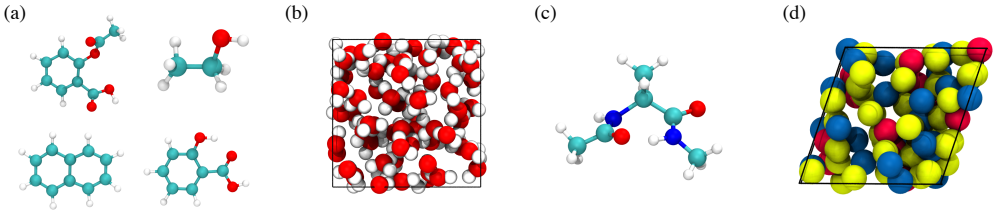

Figure 2: Visualization of the benchmarked systems. (a) MD17 molecules: Aspirin, Ethanol, Naphthalene, and Salicylic acid. (b) 64 water molecules. (c) Alanine dipeptide. (d) LiPS.

Table 1: Dataset summary. PBC stands for periodic boundary conditions. *Simulation of alanine dipeptide uses Metadynamics with implicit solvation.

Dataset	System Type	PBC	#Atoms	Simulation Length	Objective
MD17	Small molecule	✗	9-21	300 ps (600k steps)	Interatomic distances
Water	liquid	✓	192	500 ps (500k steps)	RDF, Diffusivity
Alanine dipeptide	Peptide	✗	23	5 ns (2.5M steps)*	Dihedral angle analysis
LiPS	solid-state materials	✓	83	50 ps (200k steps)	RDF, Diffusivity

reconstituting forces across test cases [12]. The evaluations invariably do not involve simulations. **However, we show that force accuracy alone does not suffice for effective simulation** (Figure 1, Force MAE and RDF MAE are the lower the better. Stability is the higher the better.).

MD simulation not only describes microscopic details on how the system evolves, but also entails macroscopic observables that characterize system properties. Calculating meaningful observables often requires long simulations to sample the underlying equilibrium distribution. These observables are designed to be predictive of material properties such as diffusivity in electrolyte materials [67], and reveal detailed physical mechanisms, such as the folding kinetics of protein dynamics [33, 36]. Although these observables are critical products of MD simulations, systematic evaluations have not been sufficiently studied in existing ML FF literature. To gain insight into the performance of existing models in a simulation setting, we propose a series of simulation-based benchmark protocols. Compared to the popular multistep prediction task in the learned simulator community [43], MD observables focus on distributional quantities. The exact recovery of the trajectories given the initial conditions is not the ultimate goal.

Evaluating learned models through MD simulations requires careful design over **the selection of systems, the simulation protocol, and the evaluation metrics**: (1) A benchmark suite should cover diverse and representative systems to reflect the various challenges in different MD applications. (2) Simulations can be computationally expensive. An ideal benchmark needs to balance the cost of evaluation and the complexity of the system so that meaningful metrics can be obtained with reasonable time and computation. (3) Selected systems should be well studied in the simulation domain, and chosen metrics should characterize the system's important degrees of freedom or geometric features.

Are current state-of-the-art (SOTA) ML FFs capable of simulating a variety of MD systems? What might cause a model to fail in simulations? In this paper, our aim is to answer these questions with a novel benchmark study. The contributions of this paper include:

- We introduce a novel benchmark suite for ML MD simulation with simulation protocols and quantitative metrics. We perform extensive experiments to benchmark a collection of SOTA ML models. We provide a complete codebase for training and simulating MD with ML FFs to lower the barrier to entry and facilitate future work.
- We show that many existing models are inadequate when they are evaluated on simulation-based benchmarks, even when they show accurate force prediction (as shown in Figure 1).
- By performing and analyzing MD simulations, we summarize common failure modes and discuss the causes and potential solutions to motivate future research.

Table 2: Models benchmarked in this work. The translation/rotation symmetries are respected by the feature representation at every layer. Number of parameters on the MD17 dataset are reported.

Model	Symmetry Principle of Geometric Features	Energy Conservation	#Parameters
DeepPot-SE [72]	E(3)-invariant	✓	1.04M
SchNet [46]	E(3)-invariant	✓	0.12M
DimeNet [17]	E(3)-invariant	✓	2.1M
ForceNet [25]	Translation-invariant	✗	11.37M
GemNet-T [16]	SE(3)-equivariant	✗	1.89M
GemNet-dT [16]	SE(3)-equivariant	✓	2.31M
NequIP [4]	E(3)-equivariant	✓	1.05M

2 Datasets and Metrics

We refer interested readers to Appendix A for a discussion on related work and Appendix B for a introduction to the problem setting of learning to simulate MD. We note that, popular benchmark datasets, such as MD17, focus on the force prediction task for gas-phase small molecules. However, successes in these tasks are not sufficient evidence for (1) capturing complex interatomic interactions that are critical for condensed phase systems; and (2) recovery of critical simulation observables that cannot be directly indicated by force prediction accuracy. We wish to extend existing benchmarks to expand coverage of the types of systems that manifest complex intermolecular interactions at multiple scales. Furthermore, beyond force predictions, we conduct simulations and benchmark observables that reflect the actual simulation quality, along with stability and computational efficiency. We choose four representative MD systems (summarized in Table 1): small molecules (MD17), liquid water (water), peptide (alanine dipeptide), and solid-state materials (LiPS). To facilitate relatively easy and fast benchmarking, we limit the size of the systems to less than 200 atoms. Detailed description for each dataset is in Appendix D.

Quantifying simulation stability. ML FFs can produce unstable dynamics, as the learned force field may not extrapolate robustly to the undersampled configuration space. As a result, the trajectories can enter nonphysical states that are not meaningful for observable calculations. Therefore, we closely monitor how much the simulated structure deviates from the physical configurations, with the RDF for condensed phase systems and bond lengths for flexible molecules. We say that a simulation becomes "unstable" when the deviation exceeds a threshold, which implies sampling of highly nonphysical structures. We then use the time duration for which a model remains stable in simulations to measure its stability. The ensemble statistics are only computed over the stable part of the simulated trajectories. Details on the stability criterion are included in Appendix C.

3 Experiments

Benchmarked models. We adopt the Open Catalyst Project implementation of SchNet [46], DimeNet [17], ForceNet [25], GemNet-T/dT [16], and the official implementation of DeepPot-SE [72] and NequIP [4]. A summary of all benchmarked models is in Table 2. These models have been popular in previous benchmark studies for force/energy prediction. They use different representations for atomistic interactions and respect different levels of euclidean symmetry. We follow all original hyperparameters introduced in the respective papers and only make minimal adjustments when the training is unstable. More details on the hyperparameters can be found in Appendix E.

Key observations. We make two key observations as evidenced in the experimental results:

1. Despite being widely used, force prediction is not sufficient for evaluating ML FFs. It generally does not align with simulation stability and performance in estimating ensemble properties.
2. While often neglected, stability can be a major bottleneck for ML FFs. Lower force error and more training data does not necessarily give rise to more stable simulations, suggesting stability as a fundamental consideration for comparison and model design.

We next go through experimental results of all four datasets in detail to demonstrate the key observations while making other observations.

MD17. As shown in Table 3, more recent models that lie on the right side of the table generally achieve a lower force error, but may lack stability. Figure 3 selects results from Table 3 to demonstrate

Table 3: Results on MD17. Darker green color indicates better performance. For all results, force MAE is reported in the unit of [meV/Å], and stability is reported in the unit of [ps]. The distribution of interatomic distances $h(r)$ MAE is unitless. FPS stands for frames per second. For all metrics (↓) indicates the lower the better, and (↑) indicates the higher the better. Standard deviation from 5 simulations is in subscript for applicable metrics.

Molecule	Model	DeepPot-SE	SchNet	DimeNet	ForceNet	GemNet-T	GemNet-dT	NequIP
Aspirin	Force (↓)	21.0	35.6	10.0	22.1	3.3	5.1	2.3
	Stability (↑)	$9_{(15)}$	$26_{(23)}$	$54_{(12)}$	$182_{(144)}$	$72_{(50)}$	$192_{(132)}$	$300_{(0)}$
	$h(r)$ (↓)	$0.65_{(0.47)}$	$0.36_{(0.57)}$	$0.04_{(0.00)}$	$0.56_{(0.15)}$	$0.04_{(0.02)}$	$0.04_{(0.01)}$	$0.02_{(0.00)}$
	FPS (↑)	88.0	108.9	20.6	137.3	28.2	56.8	8.4
Ethanol	Force	8.9	16.8	4.2	14.9	2.1	1.7	1.3
	Stability	$300_{(0)}$	$247_{(106)}$	$26_{(10)}$	$300_{(0)}$	$169_{(98)}$	$300_{(0)}$	$300_{(0)}$
	$h(r)$	$0.09_{(0.00)}$	$0.21_{(0.11)}$	$0.15_{(0.03)}$	$0.86_{(0.05)}$	$0.10_{(0.02)}$	$0.09_{(0.00)}$	$0.08_{(0.00)}$
	FPS	101.0	112.6	21.4	141.1	27.1	54.3	8.9
Naphthalene	Force	13.4	22.5	5.7	9.9	1.5	1.9	1.1
	Stability	$246_{(109)}$	$18_{(2)}$	$85_{(68)}$	$300_{(0)}$	$8_{(2)}$	$25_{(10)}$	$300_{(0)}$
	$h(r)$	$0.11_{(0.00)}$	$0.09_{(0.00)}$	$0.10_{(0.01)}$	$1.02_{(0.00)}$	$0.13_{(0.00)}$	$0.12_{(0.01)}$	$0.12_{(0.01)}$
	FPS	109.3	110.9	19.1	140.2	27.7	53.5	8.2
Salicylic Acid	Force	14.9	26.3	9.6	12.8	4.0	4.0	1.6
	Stability	$300_{(0)}$	$300_{(0)}$	$73_{(82)}$	$1_{(0)}$	$26_{(24)}$	$94_{(109)}$	$300_{(0)}$
	$h(r)$	$0.03_{(0.00)}$	$0.03_{(0.00)}$	$0.06_{(0.02)}$	$0.35_{(0.00)}$	$0.08_{(0.04)}$	$0.07_{(0.03)}$	$0.03_{(0.00)}$
	FPS	94.6	111.7	19.4	143.2	28.5	52.4	8.4

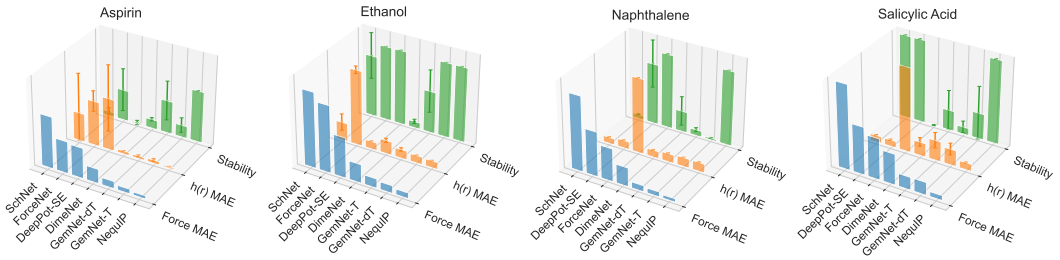

Figure 3: Head-to-head comparison of force MAE vs. Stability and $h(r)$ MAE on MD17 molecules. Models are on the x-axis and are sorted according to force error in descending order. High stability and low $h(r)$ MAE mean better performance. Error bars indicate 95% confidence intervals.

the non-aligned trends of force prediction performance vs. simulation performance, which supports **key observation 1**. GemNet-T/dT can attain a very low force error for all four molecules, but often collapse before the simulation finishes. This observation constitutes **key observation 2**. We note that although the stable portion of simulated trajectories produced by GemNet-T/dT can recover the $h(r)$ relatively accurately, stability will become a bigger issue when the statistics of interest require long simulations, as demonstrated in other experiments. On the other hand, despite having a relatively high force error, DeepPot-SE performs very well on simulation-based metrics on all molecules except for `Aspirin` (Figure 3). With the highest molecular weight, `Aspirin` is indeed the hardest task in MD17 in the sense that all models attain high force prediction errors on it.

We further observe that good stability does not imply accurate recovery of trajectory statistics. Although ForceNet remains stable for `Ethanol` and `Naphthalene`, the extracted $h(r)$ deviates a lot from the reference (Table 3), indicating that ForceNet does not learn the underlying PES correctly, possibly due to its lack of energy conservation and rotational equivariance. Overall, NequIP is the best-performing model on MD17. It achieves the best performance in both force prediction and simulation-based metrics for all molecules while requiring the highest computational cost. More detailed results on MD17 including individual $h(r)$ are included in Appendix E.

Water. Under different challenges posed by a condensed phase system, **key observation 1 and 2** are still evident according to Table 4: GemNet-T/dT and DimeNet are the top-3 models in terms of force prediction, but all lack stability. With insufficient simulation time due to early collapse, the RDFs computed from their simulated trajectories are of high variance. The water diffusivity coefficient requires long (100 ps in our experiments) trajectories to estimate and thus cannot be extracted for

Table 4: Results on Water-10k. RDF MAE is unit-less. Diffusivity is computed by averaging 5 runs from 5 random initial configurations and its MAE is reported in the unit of $[10^{-9}\ \mathrm{m}^2/\mathrm{s}]$. The reference diffusivity coefficient is $2.3 \times 10^{-9}\ \mathrm{m}^2/\mathrm{s}$.

	DeepPot-SE	SchNet	DimeNet	ForceNet	GemNet-T	GemNet-dT	NequIP
Force (↓)	5.8	9.5	1.4	10.9	0.7	1.3	1.5
Stability (↑)	$247_{(147)}$	$232_{(59)}$	$30_{(10)}$	$7_{(3)}$	$25_{(7)}$	$7_{(3)}$	$500_{(0)}$
$\mathrm{RDF}_{(O,O)}$ (↓)	$0.07_{(0.01)}$	$0.63_{(0.04)}$	$0.27_{(0.15)}$	$0.79_{(0.03)}$	$0.22_{(0.05)}$	$0.42_{(0.22)}$	$0.06_{(0.02)}$
$\mathrm{RDF}_{(H,H)}$ (↓)	$0.06_{(0.02)}$	$0.30_{(0.02)}$	$0.18_{(0.08)}$	$0.55_{(0.01)}$	$0.16_{(0.03)}$	$0.35_{(0.25)}$	$0.05_{(0.01)}$
$\mathrm{RDF}_{(H,O)}$ (↓)	$0.19_{(0.05)}$	$0.57_{(0.04)}$	$0.21_{(0.04)}$	$1.34_{(0.03)}$	$0.20_{(0.04)}$	$0.42_{(0.27)}$	$0.27_{(0.07)}$
Diffusivity (↓)	0.04	1.90	-	-	-	-	0.18
FPS (↓)	91.0	78.9	17.9	67.6	11.3	33.7	3.9

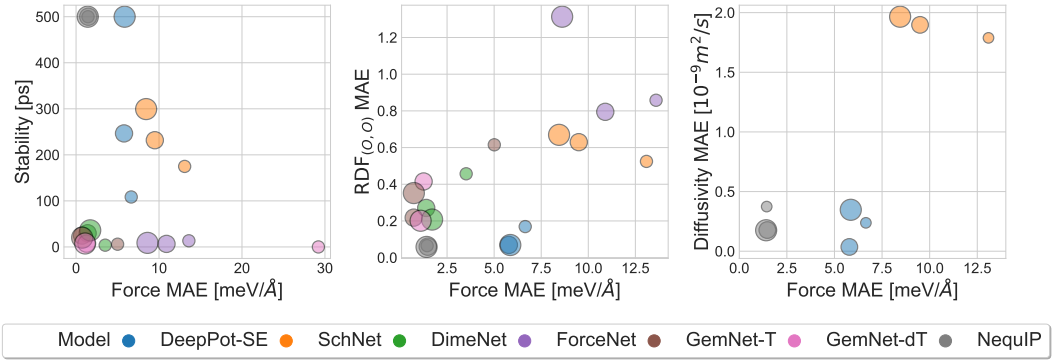

Figure 4: Comparison of force MAE vs. stability (Left), force MAE vs. RDF MAE (Middle), and force MAE vs. Diffusivity MAE (Right) on the water benchmark. Each model is trained with three dataset sizes. The color of a point indicates the model identity, while the point size indicates the training dataset size (small: 1k, medium: 10k, large: 90k). Metrics infeasible to extract from certain model/dataset size (e.g., Diffusivity for unstable models) are not included.

unstable models. Like MD17, DeepPot-SE does not achieve the best force prediction performance but demonstrates decent stability and highly accurate recovery of simulation statistics.

Figure 4 further compares model performance with different training dataset sizes. **Key observation 1 and 2** are clearly shown: Models located on the left of each scatter plot have very low force error but may have poor stability or high error in simulation statistics. More specifically, although more training data almost always improve force prediction performance, its effect on simulation performance is not entirely clear. On the one hand, GemNet-T/dT, DimeNet, and ForceNet are not stable even when under the highest training data budget. On the other hand, we observe a clear improvement of DeepPot-SE when more training data is used. NequIP is again the best performing model, achieving very low force error, excellent stability, and accurate recovery of ensemble statistics, even under the lowest data budget of 1,000 training+validation structures. However, when the training dataset is sufficiently large (90k), DeepPot-SE has equally good results as NequIP while being more than 20 times faster – dataset size also influences the model of choice for a certain dataset. More results, including tables for water-1k/90k and a study on model size, are included in Appendix E.

Alanine dipeptide poses unique challenges in sampling different metastable states separated by high free energy barriers. Table 5 shows all models have high force errors due to the random forces introduced by the lack of explicit account of water molecules. Although the force errors are in the same order of magnitude, all models except NequIP fail to simulate stably. The FES reconstruction task requires stable simulation for the entire 5 ns. NequIP is the only model that manages to finish five simulations out of six but produces inaccurate statistics. All other models are not stable enough to produce meaningful results. We further analyze the results on this task in Section 4.

LiPS. Compared to flexible molecules and liquid water, this solid material system features slower kinetics. From Table 6 we observe that most models are capable of finishing 50-ns simulations stably. In this dataset, the performance on diffusivity estimation and force prediction align well. We observe that both GemNet-T and GemNet-dT show excellent force prediction, stability, and recovery of observables, while GemNet-dT is 2.6 times faster. The better efficiency comes from the

Table 5: Results on alanine dipeptide. #Finished is the number of simulations stable for 5 ns. MAE of $F(\phi)$ and $F(\psi)$ are reported in the unit of $[\mathrm{kJ/mol}]$.

	DeepPot-SE	SchNet	DimeNet	ForceNet	GemNet-T	GemNet-dT	NequIP
Force (↓)	272.1	217.0	239.0	284.7	233.5	219.7	215.6
#Finished (↑)	0/6	0/6	0/6	0/6	0/6	0/6	5/6
Stability (↑)	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$18_{(27)}$	$0_{(0)}$	$4168_{(1860)}$
$F(\phi)$ (↓)	-	-	-	-	-	-	$108_{(2)}$
$F(\psi)$ (↓)	-	-	-	-	-	-	$126_{(4)}$
FPS (↑)	54.3	42.4	12.1	99.1	15.0	36.5	8.3

Table 6: Results on LiPS. Li-ion Diffusivity coefficient is computed by averaging 5 runs from 5 random initial configurations. The reference Li-ion diffusivity coefficient is $1.35 \times 10^{-9}\ \mathrm{m^2/s}$

	DeepPot-SE	SchNet	DimeNet	ForceNet	GemNet-T	GemNet-dT	NequIP
Force (↓)	40.5	28.8	3.2	12.8	1.3	1.4	3.7
Stability (↑)	$4_{(3)}$	$50_{(0)}$	$48_{(4)}$	$26_{(8)}$	$50_{(0)}$	$50_{(0)}$	$50_{(0)}$
RDF (↓)	$0.27_{(0.15)}$	$0.04_{(0.00)}$	$0.05_{(0.01)}$	$0.51_{(0.08)}$	$0.04_{(0.00)}$	$0.04_{(0.00)}$	$0.04_{(0.01)}$
Diffusivity (↓)	-	0.38	0.30	-	0.24	0.28	0.34
FPS (↑)	66.1	35.2	14.8	72.1	16.9	43.5	8.2

direct prediction of atomic forces $\boldsymbol{F}$ instead of taking the derivative $\boldsymbol{F} = \partial E / \partial \boldsymbol{x}$, which also makes GemNet-dT not energy-conserving – a potential issue we further discuss in Section 4.

Implications on model architecture. More recent models utilizing SE(3)/E(3)-equivariant representations and operations such as GemNet and NequIP are more expressive and can capture interatomic interactions more accurately. This is reflected by their very low force error and accurate recovery of ensemble statistics when not bottlenecked by stability. Moreover, NequIP shows that excellent accuracy and stability can be simultaneously achieved. The stability may come from parity-equivariance and the explicit architecture in manipulating higher-order geometric tensors. We believe further investigations into the extrapolation behavior induced by different equivariant geometric representations and operations [3] is a fruitful direction in designing more powerful ML FFs.

4 Failure Modes: Causes and Future Directions

A case study on alanine dipeptide simulation. NequIP achieves decent performance on all our tasks but fails on alanine dipeptide. It is also the only model that can simulate stably for 5 ns. Figure 5 (a) demonstrates how NequIP fails to reconstruct the FES: it does not manage to sample much of the transition regions and the configuration space with $\phi \in [0, 180°]$. Figure 5 (b) demonstrates the reconstructed FES, which significantly deviates from the reference. This failure can be partially explained by Figure 5 (c), the training data distribution produced by the reference potential. The relatively high-energy (low-density) regions are exactly those that are not reachable by NequIP. Even though our MD trajectory is well-equilibrated, the relative difference in populations of different meta-stable states creates data imbalance, making it more challenging for the model to learn PES for higher-energy configurations where density is relatively low. In our experiments, we observe that simulations starting from the low-density meta-stable state (e.g., black star marked in Figure 5 (c)) tend to fail. This implies that generalization across different regions in the conformational space is an important challenge for ML FFs. To prevent ML FFs from sampling nonphysical regions, which is a common precursor to failed simulation (Figure 6), one can deliberately include distorted and off-equilibrium geometries in the training data to improve model robustness[51]. Alternatively, one can resort to active learning [66, 61, 48] to acquire new data points based on model uncertainty.

Energy conservation. Models that directly predict forces may not conserve energy. Figure 5 (d) demonstrates the evolution of model-predicted total energy for selected models on LiPS, in a microcanonical (NVE) ensemble. The energy of an isolated system in the NVE ensemble is in principle conserved. We observe that GemNet-T conserves energy, whereas GemNet-dT fails to conserve the predicted total energy. The existence of non-conservative forces breaks the time reversal symmetry and, therefore, may not be able to reach equilibrium for observable calculation. However, in our experiment, GemNet-dT performs well on the LiPS dataset when coupled with a thermostat. Previous

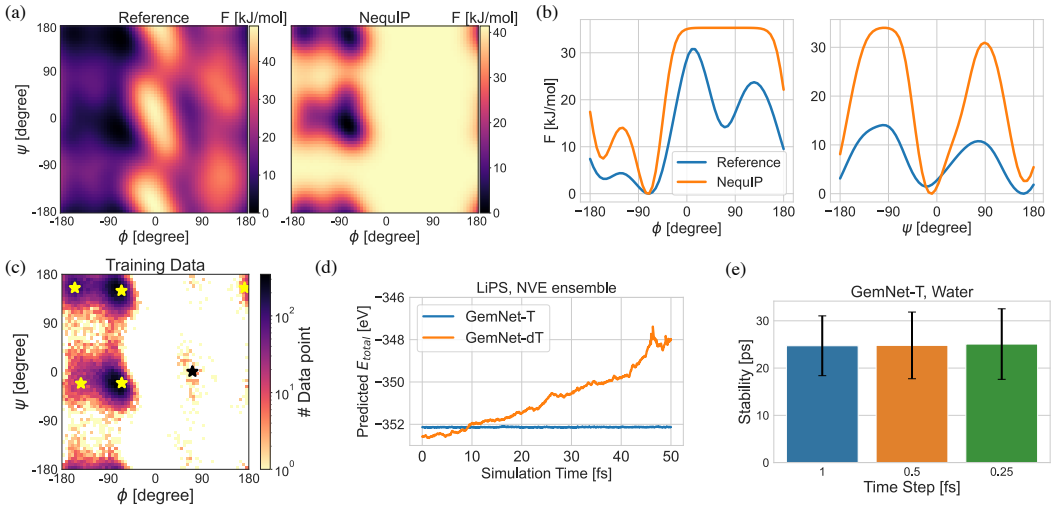

Figure 5: (a) Ramachandran plots of the alanine dipeptide FES reconstructed from 5-ns reference simulation vs. 5-ns NequIP simulation, both using MetaDynamics. (b) $F(\phi)$ and $F(\psi)$ of alanine dipeptide extracted from reference simulation vs. from NequIP simulation. (c) (ϕ, ψ) distribution of the alanine dipeptide training dataset. The six initialization points are marked with stars. NequIP fails to remain stable when the simulation starts from the point marked with black color. (d) Model-predicted total energy as a function of simulation time when simulating the LiPS system using the NVE ensemble. (e) On water-10k, stability does not improve when the time step is reduced for GemNet-T.

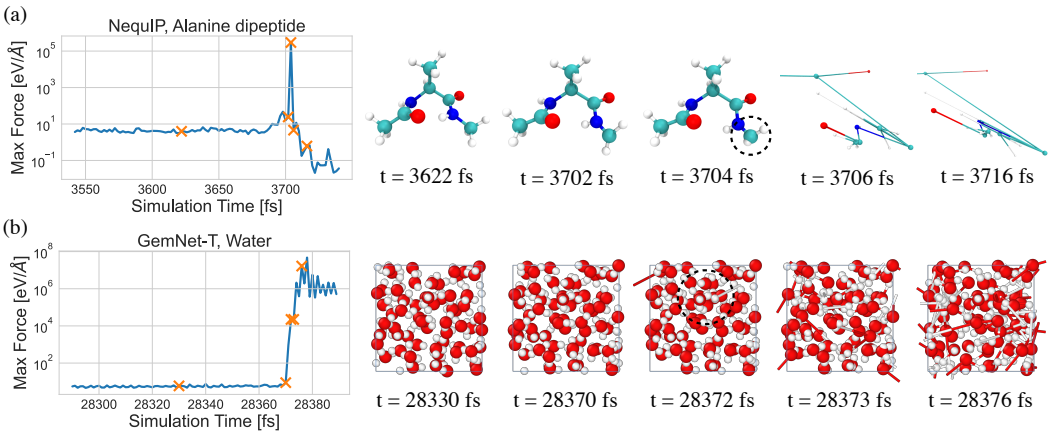

Figure 6: Examples of simulation collapse when applying (a) NequIP to alanine dipeptide and (b) GemNet-T to water. The y-axis shows the maximum force observed on any atom in the system at a certain time step. An orange cross indicates visualized time steps. Notable nonphysical regions are circled. The collapse usually happens within a very short period of time after the initial local errors.

works [30] also found that energy conservation is not required for SOTA performance on OC20. The usability of non-conservative FFs in simulations requires further careful investigations.

Simulation instability is a major bottleneck for highly accurate models such as GemNet-T to fail on several simulation tasks. Moreover, in our water experiments, we find a larger amount of training data does not resolve this issue for GemNet-T/dT and DimeNet (Figure 4). We further experiment with smaller simulation time steps for GemNet-T on water (Figure 5 (e)), but stability still does not improve. On the other hand, [51] demonstrates that the stability of GemNet improves with larger training sets on QM7-x, which includes high-energy off-equilibrium geometries obtained from normal mode sampling. We hypothesize that including these distorted geometries may improve the model's robustness against going into nonphysical configurations. We also observe that the

simulation can collapse within a short time window after a long period of stable simulation, as visualized in Figure 6. In both cases, we observe that the nonphysical configurations first emerge at local regions (circled), which cascade to the entire system very quickly as extremely large forces are being produced and subsequently integrated into the dynamics. At the end of the visualization, the bonds in the alanine dipeptide system are broken. Therefore, the local-descriptor-based NequIP model predicts very small forces. For the water system, the particles are packed in a finite periodic box. The nonphysical configurations exhibit incorrect coordination structures and extremely large forces. Regarding stability, past works found adding noise paired with a denoising objective during training helpful in improving out-of-distribution test performance on OC20 [19], and in stabilizing learned simulations [43]. Another relevant line of work in coarse-grained MD simulation has studied regularization with an empirical "prior energy" [63] and post-prediction refinement [15] to battle simulation instability.

5 Conclusion and Outlook

We have introduced a diverse suite of MD simulation tasks and conducted a thorough comparison of SOTA ML FFs to reveal novel insights into ML for MD simulation. As shown in our experiments, benchmarking only force error is not sufficient, and simulation-based metrics should be used to reflect the practical utility of a model. We demonstrate case studies on the failure of existing training schemes/models to better understand their limitations and emphasize the importance of simulation stability. Our experiments also show that the performance of a model can be highly case-dependent. For more challenging MD systems, more expressive atomistic representations may be required. For example, recent work has explored non-local descriptors [27] aiming at capturing long-range interactions in large molecules. Strictly local equivariant representations [37] are studied for very large systems where computational scalability is critical. New datasets [11] and benchmarks have been playing an important role in inspiring future work in the thriving field of ML MD simulation.

The possibility of ML in advancing MD simulation is not limited to ML force fields. Enhanced sampling methods enable fast sampling of rare events and have been augmented with ML techniques [44, 52, 24]. Differentiable simulations [45, 64, 10, 26, 20] offer a principled way of learning the force field by directly training the simulation process to reproduce experimental observables [64, 65, 53]. We hope our datasets and benchmarks will encourage future developments in all related aspects to push the frontier in ML for MD simulations.

References

[1] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, September 2015.

[2] Nongnuch Artrith, Alexander Urban, and Gerbrand Ceder. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Physical Review B*, 96(1):014112, 2017.

[3] Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor NC Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of e (3)-equivariant atom-centered interatomic potentials. *arXiv preprint arXiv:2205.06643*, 2022.

[4] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):1–11, 2022.

[5] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

[6] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021.

[7] Yaoyi Chen, Andreas Krämer, Nicholas E Charron, Brooke E Husic, Cecilia Clementi, and Frank Noé. Machine learning implicit solvation for molecular dynamics. *The Journal of Chemical Physics*, 155(8):084101, 2021.

[8] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

[9] Daniel A. Colón-Ramos, Patrick La Riviere, Hari Shroff, and Rudolf Oldenbourg. Transforming the development and dissemination of cutting-edge microscopy and computation. *Nat Methods*, 16(8):667–669, August 2019.

[10] Stefan Doerr, Maciej Majewski, Adrià Pérez, Andreas Kramer, Cecilia Clementi, Frank Noe, Toni Giorgino, and Gianni De Fabritiis. Torchmd: A deep learning framework for molecular simulations. *Journal of chemical theory and computation*, 17(4):2355–2363, 2021.

[11] Peter Eastman, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni De Fabritiis, and Thomas E. Markland. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials, 2022.

[12] Felix A Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S Schoenholz, George E Dahl, Oriol Vinyals, Steven Kearnes, Patrick F Riley, and O Anatole Von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid dft error. *Journal of chemical theory and computation*, 13(11):5255–5264, 2017.

[13] Michael Feig. Kinetics from implicit solvent simulations of biomolecules as a function of viscosity. *Journal of chemical theory and computation*, 3(5):1734–1748, 2007.

[14] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001.

[15] Xiang Fu, Tian Xie, Nathan J Rebello, Bradley D Olsen, and Tommi Jaakkola. Simulate time-integrated coarse-grained molecular dynamics with geometric machine learning. *arXiv preprint arXiv:2204.10348*, 2022.

[16] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802, 2021.

[17] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020.

[18] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[19] Jonathan Godwin, Michael Schaarschmidt, Alexander L Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Veličković, James Kirkpatrick, and Peter Battaglia. Simple gnn regularisation for 3d molecular property prediction and beyond. In *International conference on learning representations*, 2021.

[20] Joe G Greener and David T Jones. Differentiable molecular simulation can learn all the parameters in a coarse-grained force field for proteins. *PloS one*, 16(9):e0256990, 2021.

[21] Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996.

[22] Teresa Head-Gordon, Martin Head-Gordon, Michael J Frisch, Charles Brooks III, and John Pople. A theoretical study of alanine dipeptide and analogs. *International Journal of Quantum Chemistry*, 36(S16):311–322, 1989.

[23] Johannes Hoja, Leonardo Medrano Sandonas, Brian G Ernst, Alvaro Vazquez-Mayagoitia, Robert A DiStasio Jr, and Alexandre Tkatchenko. Qm7-x, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Scientific data*, 8(1):1–11, 2021.

[24] Lars Holdijk, Yuanqi Du, Ferry Hooft, Priyank Jaini, Bernd Ensing, and Max Welling. Path integral stochastic optimal control for sampling transition paths. *arXiv preprint arXiv:2207.02149*, 2022.

[25] Weihua Hu, Muhammed Shuaibi, Abhishek Das, Siddharth Goyal, Anuroop Sriram, Jure Leskovec, Devi Parikh, and C Lawrence Zitnick. Forcenet: A graph neural network for large-scale quantum calculations. *arXiv preprint arXiv:2103.01436*, 2021.

[26] John Ingraham, Adam Riesselman, Chris Sander, and Debora Marks. Learning protein structure with a differentiable simulator. In *International Conference on Learning Representations*, 2018.

[27] Adil Kabylda, Valentin Vassilev-Galindo, Stefan Chmiela, Igor Poltavsky, and Alexandre Tkatchenko. Towards linearly scaling and chemically accurate global machine learning force fields for large molecules. *arXiv preprint arXiv:2209.03985*, 2022.

[28] George A Kaminski, Richard A Friesner, Julian Tirado-Rives, and William L Jorgensen. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*, 105(28):6474–6487, 2001.

[29] Alireza Khorshidi and Andrew A Peterson. Amp: A modular approach to machine learning in atomistic simulations. *Computer Physics Communications*, 207:310–324, 2016.

[30] Adeesh Kolluru, Muhammed Shuaibi, Aini Palizhati, Nima Shoghi, Abhishek Das, Brandon Wood, C Lawrence Zitnick, John R Kitchin, and Zachary W Ulissi. Open challenges in developing generalizable large scale machine learning models for catalyst discovery. *arXiv preprint arXiv:2206.02005*, 2022.

[31] Dávid Péter Kovács, Cas van der Oord, Jiri Kucera, Alice EA Allen, Daniel J Cole, Christoph Ortner, and Gábor Csányi. Linear atomic cluster expansion force fields for organic molecules: beyond rmse. *Journal of chemical theory and computation*, 17(12):7696–7711, 2021.

[32] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[33] Thomas J Lane, Gregory R Bowman, Kyle Beauchamp, Vincent A Voelz, and Vijay S Pande. Markov state model reveals folding and functional dynamics in ultra-long md trajectories. *Journal of the American Chemical Society*, 133(45):18413–18419, 2011.

[34] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

[35] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022.

[36] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

[37] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *arXiv preprint arXiv:2204.05249*, 2022.

[38] Cheol Woo Park, Mordechai Kornbluth, Jonathan Vandermause, Chris Wolverton, Boris Kozinsky, and Jonathan P Mailoa. Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *npj Computational Materials*, 7(1):1–9, 2021.

[39] Jay W. Ponder and David A. Case. Force Fields for Protein Simulations. In *Advances in Protein Chemistry*, volume 66, pages 27–85. Elsevier, 2003.

[40] Zhuoran Qiao, Anders S Christensen, Matthew Welborn, Frederick R Manby, Anima Anandkumar, and Thomas F Miller III. Unite: Unitary n-body tensor equivariant network with applications to quantum chemistry. *arXiv preprint arXiv:2105.14655*, 2021.

[41] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

[42] David Rosenberger, Justin S Smith, and Angel E Garcia. Modeling of peptides with classical and novel machine learning force fields: A comparison. *The Journal of Physical Chemistry B*, 125(14):3598–3612, 2021.

[43] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, pages 8459–8468. PMLR, 2020.

[44] Elia Schneider, Luke Dai, Robert Q Topper, Christof Drechsel-Grau, and Mark E Tuckerman. Stochastic neural network approach for learning high-dimensional free energy surfaces. *Physical review letters*, 119(15):150601, 2017.

[45] Samuel Schoenholz and Ekin Dogus Cubuk. Jax md: a framework for differentiable physics. *Advances in Neural Information Processing Systems*, 33:11428–11441, 2020.

[46] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

[47] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[48] Daniel Schwalbe-Koda, Aik Rui Tan, and Rafael Gómez-Bombarelli. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nature communications*, 12(1):1–12, 2021.

[49] Hythem Sidky, Wei Chen, and Andrew L Ferguson. Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Molecular Physics*, 118(5):e1737742, 2020.

[50] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, 8(4):3192–3203, 2017.

[51] Sina Stocker, Johannes Gasteiger, Florian Becker, Stephan Günnemann, and Johannes Margraf. How robust are modern graph neural network potentials in long and hot molecular dynamics simulations? 2022.

[52] Mohammad M Sultan, Hannah K Wayment-Steele, and Vijay S Pande. Transferable neural networks for enhanced sampling of protein dynamics. *Journal of chemical theory and computation*, 14(4):1887–1894, 2018.

[53] Stephan Thaler and Julija Zavadlav. Learning neural network potentials from experimental data via differentiable trajectory reweighting. *Nature Communications*, 12(1):1–10, 2021.

[54] Philipp Thölke and Gianni De Fabritiis. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2021.

[55] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

[56] Richard Tran, Janice Lan, Muhammed Shuaibi, Siddharth Goyal, Brandon M Wood, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysis. *arXiv preprint arXiv:2206.08917*, 2022.

[57] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, February 2014.

[58] Oliver T Unke, Stefan Chmiela, Michael Gastegger, Kristof T Schütt, Huziel E Sauceda, and Klaus-Robert Müller. Spookynet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature communications*, 12(1):1–14, 2021.

[59] Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Sch utt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021.

[60] Oliver T Unke and Markus Meuwly. A reactive, scalable, and transferable model for molecular energies from a neural network approach based on local information. *The Journal of chemical physics*, 148(24):241708, 2018.

[61] Jonathan Vandermause, Steven B Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(1):1–11, 2020.

[62] Ercheng Wang, Huiyong Sun, Junmei Wang, Zhe Wang, Hui Liu, John Z. H. Zhang, and Tingjun Hou. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem. Rev.*, 119(16):9478–9508, August 2019.

[63] Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E Charron, Gianni De Fabritiis, Frank Noé, and Cecilia Clementi. Machine learning of coarse-grained molecular dynamics force fields. *ACS central science*, 5(5):755–767, 2019.

[64] Wujie Wang, Simon Axelrod, and Rafael Gómez-Bombarelli. Differentiable molecular simulations for control and learning. *arXiv preprint arXiv:2003.00868*, 2020.

[65] Wujie Wang, Zhenghao Wu, and Rafael Gómez-Bombarelli. Learning pair potentials using differentiable simulations. *arXiv preprint arXiv:2209.07679*, 2022.

[66] Wujie Wang, Tzuhsiung Yang, William H Harris, and Rafael Gómez-Bombarelli. Active learning and neural network potentials accelerate molecular screening of ether-based solvate ionic liquids. *Chemical Communications*, 56(63):8920–8923, 2020.

[67] Michael A Webb, Yukyung Jung, Danielle M Pesko, Brett M Savoie, Umi Yamamoto, Geoffrey W Coates, Nitash P Balsara, Zhen-Gang Wang, and Thomas F Miller III. Systematic computational and experimental investigation of lithium-ion transport mechanisms in polyester-based polymer electrolytes. *ACS central science*, 1(4):198–205, 2015.

[68] Yujie Wu, Harald L. Tepper, and Gregory A. Voth. Flexible simple point-charge water model with improved liquid-state properties. *The Journal of Chemical Physics*, 124(2):024503, January 2006.

[69] Shuwen Yue, Maria Carolina Muniz, Marcos F. Calegari Andrade, Linfeng Zhang, Roberto Car, and Athanassios Z. Panagiotopoulos. When do short-range atomistic machine-learning models fall short? *J. Chem. Phys.*, 154(3):034111, January 2021.

[70] Yaoguang Zhai, Alessandro Caruso, Sigbjörn L Bore, and Francesco Paesani. A "short blanket" dilemma for a state-of-the-art neural network potential for water: Reproducing properties or learning the underlying physics? 2022.

[71] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and EJPRL Weinan. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters*, 120(14):143001, 2018.

[72] Linfeng Zhang, Jiequn Han, Han Wang, Wissam Saidi, Roberto Car, et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *Advances in Neural Information Processing Systems*, 31, 2018.

A Related Work

ML force fields learn the potential energy surface (PES) from the data by applying expressive regressors such as kernel methods [8] and neural networks on symmetry-preserving representations of atomic environments [5, 29, 50, 2, 60, 72, 71, 31, 54]. Recently, graph neural network architectures [18, 46, 17] have gained popularity as they provide a systematic strategy for building many-body correlation functions to capture highly complex PES. In particular, equivariant representations have been shown powerful in representing atomic environments [55, 40, 47, 4, 16, 35], leading to significant improvements in benchmarks such as MD17 and OC22/20. Some works presented simulation-based results [58, 38, 4, 37] but do not compare different models with simulation-based metrics.

Existing benchmarks for ML force fields [41, 8] mostly focus on force/energy prediction, with small molecules being the most typical systems. The catalyst-focused OC20 [6] and OC22 [56] benchmarks proposed the Initial Structure to Relaxed Structure/Energy tasks, which aim to predict the relaxed structure/energy through structural optimizations. These tasks do not characterize system properties under a structural ensemble which requires simulations. Several recent works [42] have also studied the utility of certain ML FFs in MD simulations. In particular, [51] uses GemNet [16] to simulate small molecules in the QM7-x [23] dataset, with a focus on simulation stability. [70] applies the DeepMD [71] architecture to simulate water and demonstrates its shortcoming in generalization across different phases. However, systematic benchmarks for simulation-based metrics are lacking in the existing literature, which obscures the challenges in applying ML FF for MD applications.

B Preliminaries

Training. An ML FF aims to learn the potential energy surface $\hat{E}(x) \in \mathbb{R}$ as a function of atomic coordinates $\boldsymbol{x} \in \mathbb{R}^{N\times 3}$ (N is the number of atoms), by fitting atom-wise forces $\hat{\boldsymbol{F}}(\boldsymbol{x})$ and energies from a training dataset:$\{\boldsymbol{x}_i, \boldsymbol{F}_i, E_i\}_{i=1}^{N_{\text{data}}}$, where $\boldsymbol{x}_i \in \mathbb{R}^{N\times 3}$, $\boldsymbol{F} \in \mathbb{R}^{N\times 3}$, $E \in \mathbb{R}$. For evaluation, the test force prediction accuracy is used as a proxy to quantify the quality of the learned PES. The force field learning protocol has been well established [59].

MD simulation. Simulating molecular behaviors requires integrating a Newtonian equation of motion with forces obtained by differentiating $\hat{E}(x)$: $\boldsymbol{F}(\boldsymbol{x}) = -\partial\hat{E}(x)/\partial\boldsymbol{x}$. To mimic desired thermodynamic conditions, an appropriate thermostat and barostat are chosen to augment the equation of motion with extended variables. The simulation produces a time series of positions: $\{x_t \in \mathbb{R}^{N\times 3}\}_{t=0}^{T}$, where t is the temporal order index, and T is the total simulation steps.

Observables. From the time series observations of positions and velocities, observables $O(x_t)$ can be computed to characterize the state of the system at different granularities. Under the ergodic hypothesis, the time averages of the simulation observables converge to distributional averages under the Gibbs measure: $\langle O\rangle = \frac{1}{T}\lim_{T\to\infty}\sum_t^T O(x_t) = \int dx\ p(x)\ O(x)$, where $p(x) \propto \exp(-\frac{\hat{E}(x)}{k_B T})$ with T as the bath temperature and k_B as the Boltzmann constant. Calculations of such observables require the system to reach equilibrium. Simulation observables connect simulations to experimental measurements and are predictive of macroscopic properties of matter. Common observables include radial distribution functions (RDFs), virial stress tensor, mean-squared displacement (MSD), etc. The metrics used in this work are all well-established observables in the respective types of systems. Definitions of benchmarked observables are in Appendix C.

C Evaluation Metrics

Distribution of interatomic distances is a low-dimensional description of the 3D structure and has been studied in previous work [71]. For a given configuration $\boldsymbol{x}$, the distribution of interatomic distances $h(r)$ is computed with:

$$h(r) = \frac{1}{N(N-1)}\sum_{i}^{N}\sum_{j\neq=i}^{N}\delta(r - ||\boldsymbol{x}_i - \boldsymbol{x}_j||) \tag{1}$$

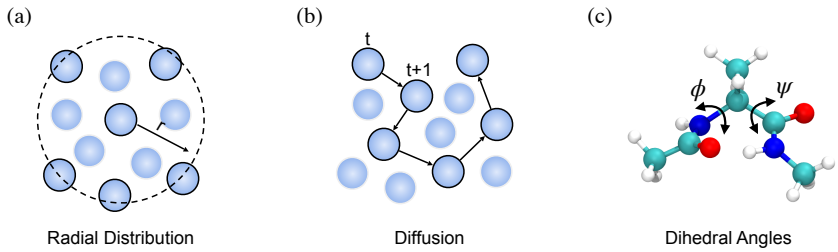

Figure 7: Illustrations regarding benchmarked metrics.

where r is the distance from a reference particle; N is the total number of particles; i, j indicates the pairs of atoms that contribute to the distance statistics; δ is the Dirac Delta function to extract value distributions. To calculate the ensemble average, $h(r)$ is calculated and averaged over frames from equilibrated trajectories. For learned simulations, we compute $h(r)$ using only the stable part of the simulation.

RDF. As one of the most informative simulation observables, the radial distribution function (RDF) describes the structural/thermodynamic properties of the system and is also experimentally measurable. By definition, the RDF describes how density varies as a function of distance from a particle (illustrated in Figure 7 (a)). For a given configuration $\boldsymbol{x}$, the RDF can be computed with the following formula:

$$\mathrm{RDF}(r) = \frac{1}{4\pi r^2} \frac{1}{N\rho} \sum_{i}^{N} \sum_{j \neq = i}^{N} \delta(r - ||\boldsymbol{x}_i - \boldsymbol{x}_j||) \tag{2}$$

where r is the distance from a reference particle; N is the total number of particles; i, j indicates the pairs of atoms that contribute to the distance statistics; ρ is the density of the system; δ is the Dirac Delta function to extract value distributions. To calculate the ensemble average, $\mathrm{RDF}(r)$ is calculated and averaged over frames from equilibrated trajectories. The final RDF MAE is then calculated by integrating r:

$$\mathrm{MAE}_{\mathrm{RDF}} = \int_{r=0}^{\infty} |\langle \mathrm{RDF}(r) \rangle - \langle \hat{\mathrm{RDF}}(r) \rangle| dr \tag{3}$$

where $\langle \cdot \rangle$ is the averaging operator, $\langle \mathrm{RDF}(r) \rangle$ is the reference equilibrium RDF, and $\langle \hat{\mathrm{RDF}}(r) \rangle$ is the model-predicted RDF.

Diffusivity coefficient. The diffusivity coeffcient D quantifies the time-correlation of the translational displacement (illustrated in Figure 7 (b)), and can be computed from the mean square displacement:

$$D = \lim_{t \to \infty} \frac{1}{6t} \frac{1}{N'} \sum_{i=1}^{N'} |\boldsymbol{x}_i(t) - \boldsymbol{x}_i(0)|^2 \tag{4}$$

where $x_i(t)$ is the coordinate of particle i at time t, and N' is the number of particles being tracked in the system. For the water system we monitor the liquid diffusivity coefficient and track all 64 oxygen atoms. For the LiPS system, we monitor Li-ion Diffusivity and track all 27 Li-ions. As the definition implies, D is a quantity that converges with longer simulation time. Accurate recovery of D requires sufficient long trajectories sampled from the Hamiltonian with ML FFs. In this paper, we only compute diffusivity for stable trajectories of at least 100 ps for water and 40 ps for LiPS. As we simulate multiple random runs for each system/model, we average the diffusivity coefficient extracted from each valid trajectory to obtain the final prediction.

Free energy surface. Given the probability distributions over configurations $p(\boldsymbol{x})$ and a chosen geometric coordinate $\boldsymbol{\xi}$ transformed from $\boldsymbol{x}$. Based on the marginalized density $p(\boldsymbol{\xi})$, the free energy can be calculated from the following:

$$F(\boldsymbol{\xi}) = -k_B T \ln p(\boldsymbol{\xi}) \tag{5}$$

In the specific case of alanine dipeptide, there are two main conformational DOF: dihedral angle ϕ of $\mathrm{C}-\mathrm{N}-\mathrm{C}_\alpha-\mathrm{C}$ and dihedral angle ψ of $\mathrm{N}-\mathrm{C}_\alpha-\mathrm{C}-\mathrm{N}$ (illustrated in Figure 7 (c)). Therefore, the FES w.r.t ϕ and ψ is the most physically informative. We propose our quantitative metric $\mathrm{MAE}_{F(\phi),F(\psi)}$ based on the absolute error in reconstructing the FES along the ϕ and ψ coordinates. We integrate the absolute difference between the reference free energy F and the model predicted $\hat{F}$ from $[-\pi, \pi)$:

$$\mathrm{MAE}_{F(\phi)} = \int_{\phi=-\pi}^{\pi} |F(\phi) - \hat{F}(\phi)| d\phi \tag{6}$$

$$\mathrm{MAE}_{F(\psi)} = \int_{\psi=-\pi}^{\pi} |F(\psi) - \hat{F}(\psi)| d\psi \tag{7}$$

Stability criterion. Abstractly speaking, stability is a notion of staying within physical (low-energy) configuration spaces. Since all MD systems studied in this paper are in equilibrium, practically we keep track of stability by closely monitoring equilibrium statistics. For systems with periodic boundary conditions, we monitor the RDF and say a simulation becomes "unstable" at time T when

$$\int_{r=0}^{\infty} ||\langle \mathrm{RDF}(r) \rangle - \langle \hat{\mathrm{RDF}}_t(r) \rangle_{t=T}^{T+\tau}|| dr > \Delta \tag{8}$$

where $\langle \cdot \rangle$ is the averaging operator, τ is a short time window, and Δ is the stability threshold. In this paper we use $\tau = 1$ ps, $\Delta = 3.0$ for water, and $\tau = 1$ ps, $\Delta = 1.0$ for LiPS. For water, we assert unstable if any of the three element-conditioned RDFs: $\mathrm{RDF}_{(O,O)}$, $\mathrm{RDF}_{(H,H)}$, $\mathrm{RDF}_{(H,O)}$ exceeds the threshold. For flexible molecules, we keep track of stability through the bond lengths and say a simulation becomes "unstable" at time T when:

$$\max_{(i,j)\in\mathcal{B}} |(||\boldsymbol{x}_i(T) - \boldsymbol{x}_j(T)|| - b_{i,j})| > \Delta \tag{9}$$

where $\mathcal{B}$ is the set of all bonds, i, j are the two endpoint atoms of the bond, and $b_{i,j}$ is the equilibrium bond length. We use $\Delta = 0.5$ Å for both MD17 molecules and alanine dipeptide.

We measure the stability of a learned model through the simulation time it remains stable. For each dataset, the threshold Δ we adopt is rather relaxed that an "unstable" equilibrium statistic usually indicates the system is already in a highly nonphysical configuration.

FPS. All frames per second (FPS) metrics are measured with an NVIDIA Tesla V100-PCIe GPU. We present FPS as a reference for models' computational efficiency but also note that code speed can be affected by many factors, and likely has room for improvement.

D Dataset Details

The MD17 dataset[3] [8] and the LiPS dataset[4] [4] are adapted from previous works and are publicly available. We refer interested readers to the respective papers for more details on the data generation process. The water dataset and alanine dipeptide dataset are generated by ourselves, and will be made publicly available.

MD17 [8] contains AIMD calculations for eight small organic molecules and is widely used as a force prediction benchmark for ML FFs. We adopt four molecules from MD17 and benchmark the simulation performance. In addition to force error, we evaluate the stability and the distribution of interatomic distances $h(r)$. For each molecule, we randomly sample 9,500 configurations for training and 500 for validation from the MD17 database. We randomly sample 10,000 configurations from the rest of the data for force error evaluation. We perform five simulations of 300 ps for each model/molecule by initializing from 5 randomly sampled testing configurations, with a time step of 0.5 fs, at 500 K temperature, under a Nosé–Hoover thermostat.

[3] http://www.sgdml.org/

[4] https://archive.materialscloud.org/record/2022.45

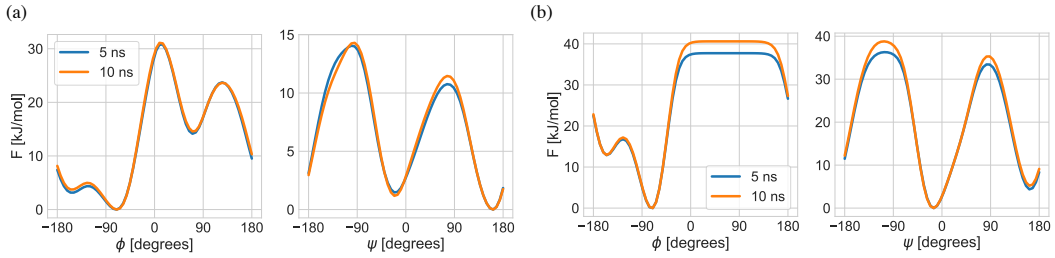

Figure 8: F_ϕ and F_ψ have converged for the reference force field (a) and NequIP (b) at time 5 ns under Metadynamics.

Water is arguably the most important molecular fluid in biological and chemical processes. Due to its complex thermodynamic and phase behavior, it poses great challenges for molecular simulations. In addition to force error, we evaluate simulation stability and recovery of both equilibrium statistics and dynamical statistics, namely the element-conditioned RDFs and liquid diffusion coefficient. Our dataset consists of 100,000 structures collected every 10 fs from a 1 ns trajectory sampled at equilibrium and a temperature of 300 K. We benchmarked all models with various training+validation dataset sizes (1k/10k/90k randomly sampled structures) and used the remaining 10,000 structures for testing. We performed 5 simulations of 500 ps by initializing from 5 randomly sampled testing configurations, with a time step of 1 fs, at 300 K temperature, under a Nosé–Hoover thermostat.

Alanine dipeptide features multiple metastable states, making it a classic benchmark for the development of MD sampling methods and force field development [22, 28]. Its geometric flexibility is well represented by the central dihedral (torsional) angles ϕ and ψ. Our reference data are obtained from simulations with explicit water molecules, with detailed protocols described in the next section. For faster simulation, we learn an implicitly solvated FF following a protocol similar to [7]. Our task is more challenging in that it aims to learn the implicitly solvated atomistic FF rather than the implicit solvation correction in [7]. To facilitate accelerated sampling, we apply metadynamics with ϕ and ψ as the collective variables. We evaluate force prediction, simulation stability, and free energy surface (FES) reconstruction $F(\phi, \psi)$. Our dataset consists of 50,000 structures dumped every 2 ps from a 100 ns trajectory at a temperature of 300 K. We used 38,000 randomly sampled structures for training, 2,000 for validation, and the rest as a test set. We performed 6 simulations of 5 ns by initializing from six local minima on the FES (Figure 5) with a time step of 2 fs at 300 K, and under a Langevin thermostat to mimic random noise from solvation effects.

LiPS is a crystalline superionic lithium conductor relevant to battery development and a representative system for MD simulation usage in studying kinetic properties in materials. We adopt this dataset from [4], and benchmark all models on their force error, stability, RDF recovery, and Li-ion diffusivity coefficient. The dataset has 25,000 structures in total, from which we use 19,000 randomly sampled structures for training, 1,000 structures for validation, and the rest for computing force error. We conduct 5 simulations of 50 ps by initializing from 5 randomly sampled testing configurations, with a time step of 0.25 fs, at 520 K temperature, under a Nosé–Hoover thermostat.

D.1 Data Generation and Simulation Protocols

Water. Our water dataset is generated from molecular dynamics simulations of a simple classical water model, namely, the flexible version of the Extended Simple Point Charge water model (SPC/E-fw) [68] at temperature $T = 300$ K and pressure $P = 1$ atm. This classical model has been well-studied in previous work [68, 69] and has shown reasonable predictions of the physical properties of liquid water. It provides a computationally inexpensive way to generate a large amount of training data. The experience and knowledge gained from the benchmark based on the simple model can be readily extended to systems with higher accuracy, such as the *ab-initio* models.

Alanine dipeptide. Our dataset is generated from the MD simulation of an alanine dipeptide molecule solvated in explicit water (1164 water molecules) performed in GROMACS [1] using the AMBER-03 [39] force-field. The NPT ensemble is applied in simulations, with hydrogen bond length constraints using LINear Constraint Solver (LINCS) and a time step of 2 fs. The temperature and pressure of the system are controlled at $T = 300$ K and $P = 1$ bar using a stochastic velocity

Table 7: Training-related hyperparameters for each dataset. *We adopt the original batch size from respective papers when available for MD17. DeepPot-SE: 4; SchNet: 100; DimeNet: 32; GemNet-T/dT: 1; NequIP: 5. We use a batch size of 8 for ForceNet.

Dataset	Training dataset size	Batch size	Max epoch	LR patience	Longest simulation time
MD17	9,500	1-100*	2,000	5 epochs	20 hours
Water-1k	950	1	10,000	50 epochs	28 hours
Water-10k	9,500	1	2,000	5 epochs	28 hours
Water-90k	85,500	1	400	3 epochs	28 hours
Alanine dipeptide	38,000	5	2,000	5 epochs	75 hours
LiPS	19,000	1	2,000	5 epochs	7 hours

rescaling thermostat with damping frequency $t_v = 0.1$ ps and Parrinello-Rahman barostat with coupling frequency $t_p = 2.0$ ps, respectively. The Particle Mesh Ewald approach is used to compute long-range electrostatics with periodic boundary conditions applied to the x, y, and z directions.

Implicit solvation. The explicit solvent of 1164 water molecules is not the subject of study but adds a significant computational burden. In this task, we attempt to learn an implicit solvent model (ISM) of the alanine dipeptide, in which the explicit solvent environment is incorporated in the learned FF. The ISM is commonly used in drug design [62] because it can speed up the computation by dramatically decreasing the number of particles required for simulation. In general, the mean-field estimation in ISM ignores the effect of solvent, thermal fluctuations, and solvent friction [13]. Thus, molecular kinetics is not directly comparable to the explicit solvation simulation. However, the equilibrium configurations can be explicitly compared, as conducted in [7].

Metadynamics simulation. Simulating energy barrier jump usually requires a long sampling of the trajectory in MD simulations. The conformational change of alanine dipeptide in water involves such a process, making it difficult to extract the complete free energy surface, i.e., the Ramachandran plot, in normal MD. In order to examine the learned ML FFs within a reasonable time limit, metadynamics [32] is employed to explore the learned FES of the solvated alanine dipeptide. Metadynamics is a widely used technique in atomistic simulations to accelerate the sampling of rare events and estimate the FES of a certain set of degrees of freedom. It is based on iteratively "filling" the potential energy using a sum of Gaussians deposited on a set of suitable collective variables (CVs) along the trajectory. At evaluation time, we perform metadynamics with dihedral angles ϕ and ψ as CVs[5], starting from the configurations located at one of the six energy minimums in the free energy surface indicated in Figure 5 (c). The Gaussians with height $h = 1.2$ and sigma $\sigma = 0.35$ are deposited every 1 ps centered on ψ and ϕ. As shown in Figure 8, the estimated FES of both ϕ and ψ do not significantly change after 5 ns. In addition, the height of the bias gaussian potential smoothly converges to ~ 0 in the time limit of 5 ns. Therefore, a simulation time of 5 ns is sufficient for the convergence of the metadynamics. This metadynamics simulation of alanine dipeptide with AMBER force-fields is carried out using GROMACS [1] integrated with the PLUMED library [57, 9] of version 2.8.

E Experimental details

The Open Catalyst Project[6] codebase and the official codebases of DeepPot-SE[7] and NequIP[8] are all publicly available. We build our MD simulation framework based on the Atomic Simulation Environment (ASE) library[9] [34]. Our code will be made publicly available after the reviewing process.

Hyperparameters. We adopt the original model hyperparameters in the respective papers and find they can produce good force prediction results that match the trend and numbers for MD17 reported in previous work. As we introduce new datasets, we set training hyperparameters such as the batch size and summarize them in Table 7. For water and LiPS, we use a batch size of 1 like in previous work [4]

[5] In practice, the selection of suitable collective variables can be a case-by-case challenge [49].

[6] https://github.com/Open-Catalyst-Project/ocp

[7] https://github.com/deepmodeling/deepmd-kit

[8] https://github.com/mir-group/nequip

[9] https://gitlab.com/ase/ase

Table 8: Results on Water-1k. Force MAE is reported in the unit of [meV/Å]; Stability is reported in the unit of [ps]; Diffusivity MAE is reported in the unit of [10^{-9} m^2/s]; RDF MAE and FPS are unitless.

	DeepPot-SE	SchNet	DimeNet	ForceNet	GemNet-T	GemNet-dT	NequIP
Force	6.7	13.1	3.5	13.6	5.0	29.2	1.4
Stability	$108_{(117)}$	$175_{(56)}$	$4_{(4)}$	$13_{(7)}$	$6_{(7)}$	$0_{(0)}$	$500_{(0)}$
$\mathrm{RDF}_{(O,O)}$	$0.17_{(0.10)}$	$0.52_{(0.05)}$	$0.46_{(0.22)}$	$0.86_{(0.09)}$	$0.62_{(0.48)}$	-	$0.07_{(0.02)}$
$\mathrm{RDF}_{(H,H)}$	$0.13_{(0.09)}$	$0.24_{(0.02)}$	$0.33_{(0.15)}$	$0.56_{(0.04)}$	$0.35_{(0.21)}$	-	$0.07_{(0.02)}$
$\mathrm{RDF}_{(H,O)}$	$0.28_{(0.15)}$	$0.54_{(0.01)}$	$0.43_{(0.17)}$	$1.44_{(0.09)}$	$0.71_{(0.65)}$	-	$0.26_{(0.07)}$
Diffusivity	0.24	1.79	-	-	-	-	0.37

Table 9: Results on Water-90k. Force MAE is reported in the unit of [meV/Å]; Stability is reported in the unit of [ps]; Diffusivity MAE is reported in the unit of [10^{-9} m^2/s]; RDF MAE and FPS are unitless.

	DeepPot-SE	SchNet	DimeNet	ForceNet	GemNet-T	GemNet-dT	NequIP
Force	5.9	8.4	1.7	8.6	0.7	1.1	1.4
Stability	$500_{(0)}$	$299_{(70)}$	$36_{(9)}$	$9_{(12)}$	$20_{(9)}$	$8_{(10)}$	$500_{(0)}$
$\mathrm{RDF}_{(O,O)}$	$0.07_{(0.02)}$	$0.67_{(0.03)}$	$0.21_{(0.03)}$	$1.31_{(0.49)}$	$0.35_{(0.23)}$	$0.20_{(0.01)}$	$0.06_{(0.01)}$
$\mathrm{RDF}_{(H,H)}$	$0.05_{(0.01)}$	$0.31_{(0.02)}$	$0.14_{(0.01)}$	$0.82_{(0.26)}$	$0.25_{(0.19)}$	$0.16_{(0.01)}$	$0.04_{(0.01)}$
$\mathrm{RDF}_{(H,O)}$	$0.29_{(0.08)}$	$0.67_{(0.04)}$	$0.18_{(0.02)}$	$2.05_{(0.60)}$	$0.24_{(0.06)}$	$0.26_{(0.02)}$	$0.25_{(0.06)}$
Diffusivity	0.35	1.97	-	-	-	-	0.18

as each structure already contains a reasonable number of atoms and interactions. Following previous work, we use an initial learning rate of 0.001 for all experiments except for NequIP, which uses 0.005 as the initial learning rate in the original paper. For models that minimize a mixture of force loss and energy loss, we set the force loss coefficient λ_F to be 1000 and the energy loss coefficient λ_E to be 1, if not specified in the original paper for a dataset. A higher force loss coefficient is common in previous work [71, 4] as simulations do not directly rely on the energy.

Notably, NequIP proposed several sets of hyperparameters for different datasets, including MD17, a water+ice dataset from [71], LiPS, etc. We follow the MD17 hyperparameters of NequIP for our MD17 and alanine dipeptide datasets; the water+ice hyperparameters of NequIP for our water dataset; and the LiPS hyperparameters of NequIP for the same LiPS dataset. The only architectural adjustment we made is because we observed training instability for ForceNet on water using the original hyperparameters. We resolve this issue by reducing the network width from 512 to 128 for ForceNet in our water experiments.

To facilitate fair benchmarking, we stop the training of a model if either of the following conditions is met: (1) a maximum training time of 7 days is reached on an NVIDIA Tesla V100-PCIe GPU; (2) a maximum number of epochs specified in Table 7 is reached; (3) The learning rate drops below 10^{-6} with a `ReduceLROnPlateau` scheduler with factor 0.8 and learning rate (LR) patience specified in Table 7. We also report the longest time for an ML model to finish our benchmark simulation in Table 7. All numbers are results of NequIP. The high computational cost for evaluating MD simulations has been a major consideration in designing our benchmark datasets and metrics.

Complete water results. We present results on water-1k in Table 8 and results on water-90k in Table 9. Results on water-10k is presented in Table 4 in the main text. All models generally achieve lower force error when trained with more data, but stability and estimation of ensemble statistics don't necessarily improve. Notably, DeepPot-SE shows clear improvement with more training data and becomes as good as NequIP on water-90k. SchNet demonstrates significant improvement in stability, but the estimation of ensemble statistics does not improve. This may be due to the limited accuracy of SchNet coming from the limited expressiveness of the invariant atomic representation.

Influence of model size. Table 10 shows an ablation study over the model size and radius cutoff of NequIP over water-1k. We observe that all models are highly stable and attain equally good performance in simulation-based metrics. Although a small radius cutoff of 4 leads to worse performance in force prediction, it is more computationally efficient and preserves the trajectory statistics. These results show that there exists a trade-off between accuracy and efficiency when

Table 10: Water-1k results on NequIP with various model sizes and radius cutoffs.

	Force	Stability	$\text{RDF}_{(O,O)}$	$\text{RDF}_{(H,H)}$	$\text{RDF}_{(H,O)}$	Diffusivity	FPS
Width=64, r=4	3.5	$500_{(0)}$	$0.07_{(0.02)}$	$0.05_{(0.01)}$	$0.27_{(0.06)}$	0.38	8.2
Width=32, r=6	1.5	$500_{(0)}$	$0.06_{(0.01)}$	$0.05_{(0.01)}$	$0.26_{(0.06)}$	0.25	5.2
Width=64, r=5	1.6	$500_{(0)}$	$0.07_{(0.02)}$	$0.05_{(0.01)}$	$0.27_{(0.06)}$	0.31	4.9
Width=64, r=6	1.4	$500_{(0)}$	$0.07_{(0.02)}$	$0.07_{(0.02)}$	$0.26_{(0.07)}$	0.37	3.9
Width=128, r=6	1.5	$500_{(0)}$	$0.07_{(0.02)}$	$0.05_{(0.01)}$	$0.29_{(0.07)}$	0.37	2.5

choosing the hyperparameters of an ML force field, and force error may not be the preferred criterion for model selection.

Distribution of interatomic distances for MD17. Figure 9 shows the $h(r)$ curves for all models and molecules benchmarked. We randomly selected one simulation out of the five simulations we conducted for each model and molecule. We observe that due to lack of stability, DeepPot-SE produces noisy $h(r)$ on Aspirin. ForceNet does not manage to learn the correct interatomic interactions and produces incorrect $h(r)$ curves. Most models are able to produce $h(r)$ that match well with the reference, with SchNet being less accurate on `Aspirin` and `Ethanol`.

RDFs for water. Selected RDF curves for water-1k/10k/90k are in Figure 10, Figure 11, and Figure 12. Most noisy curves are due to insufficient sampling time, which results in a small number of frames to be averaged in obtaining the RDF curves. We observe that SchNet and ForceNet produce inaccurate curves that are not very noisy, showing that their failure is not entirely due to lack of stability but because of inaccurate modeling of interactions caused by limited expressiveness and lower sample efficiency. Further, we note that the reference curves have zero values below a certain threshold, as any pair of atoms cannot get too close to each other. However, DimeNet and GemNet-T exhibit abnormal high values for very small distances, indicating the simulations have gone into nonphysical configurations and collapsed.

RDFs for LiPS. As shown in Figure 13, DeepPot-SE does not manage to stay stable on LiPS. ForceNet learns inaccurate interactions and produces inaccurate RDFs. All other models can produce highly accurate RDF and can reproduce Li-ion diffusivity relatively accurately, as demonstrated in Table 6.

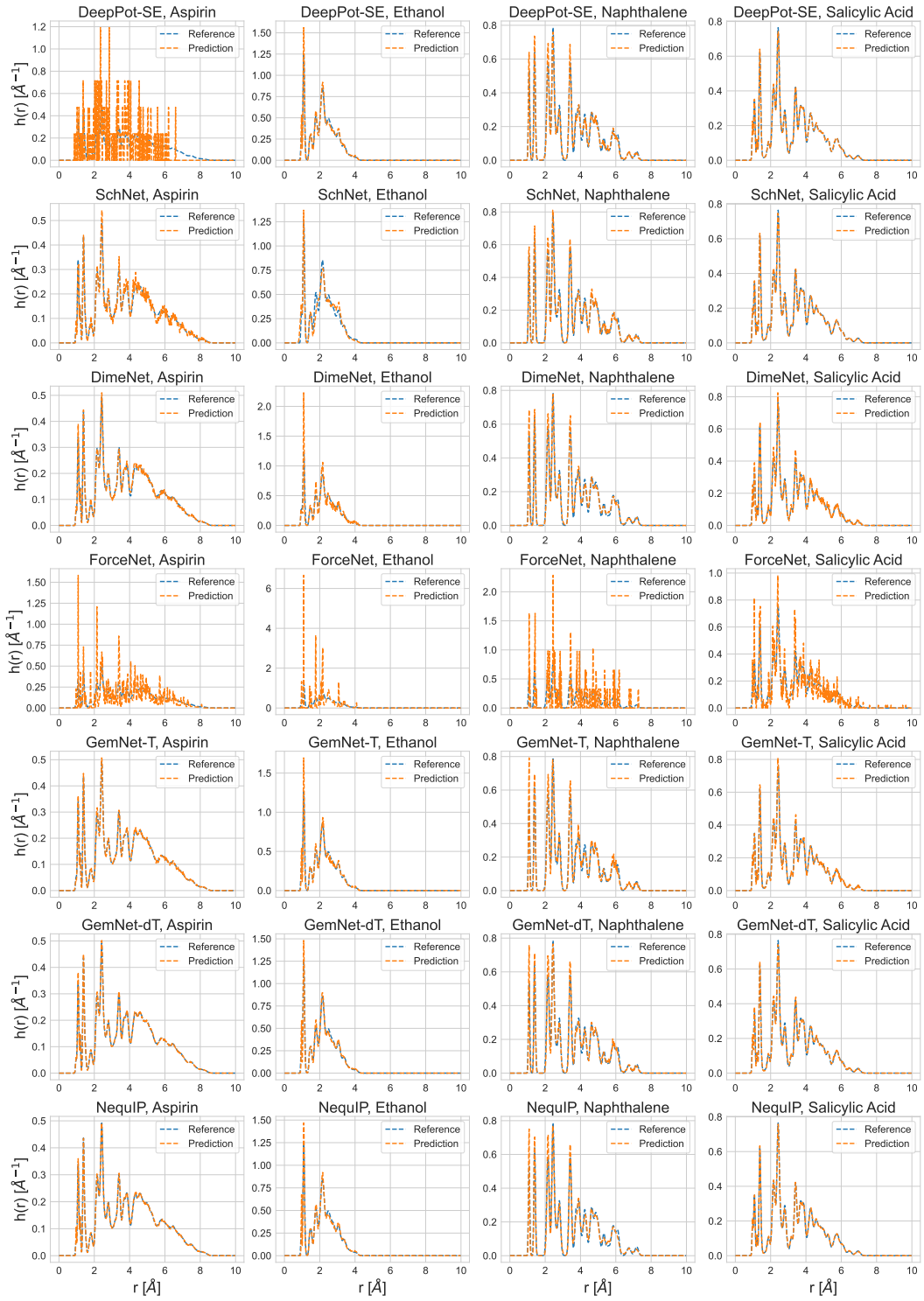

Figure 9: h(r) of md17

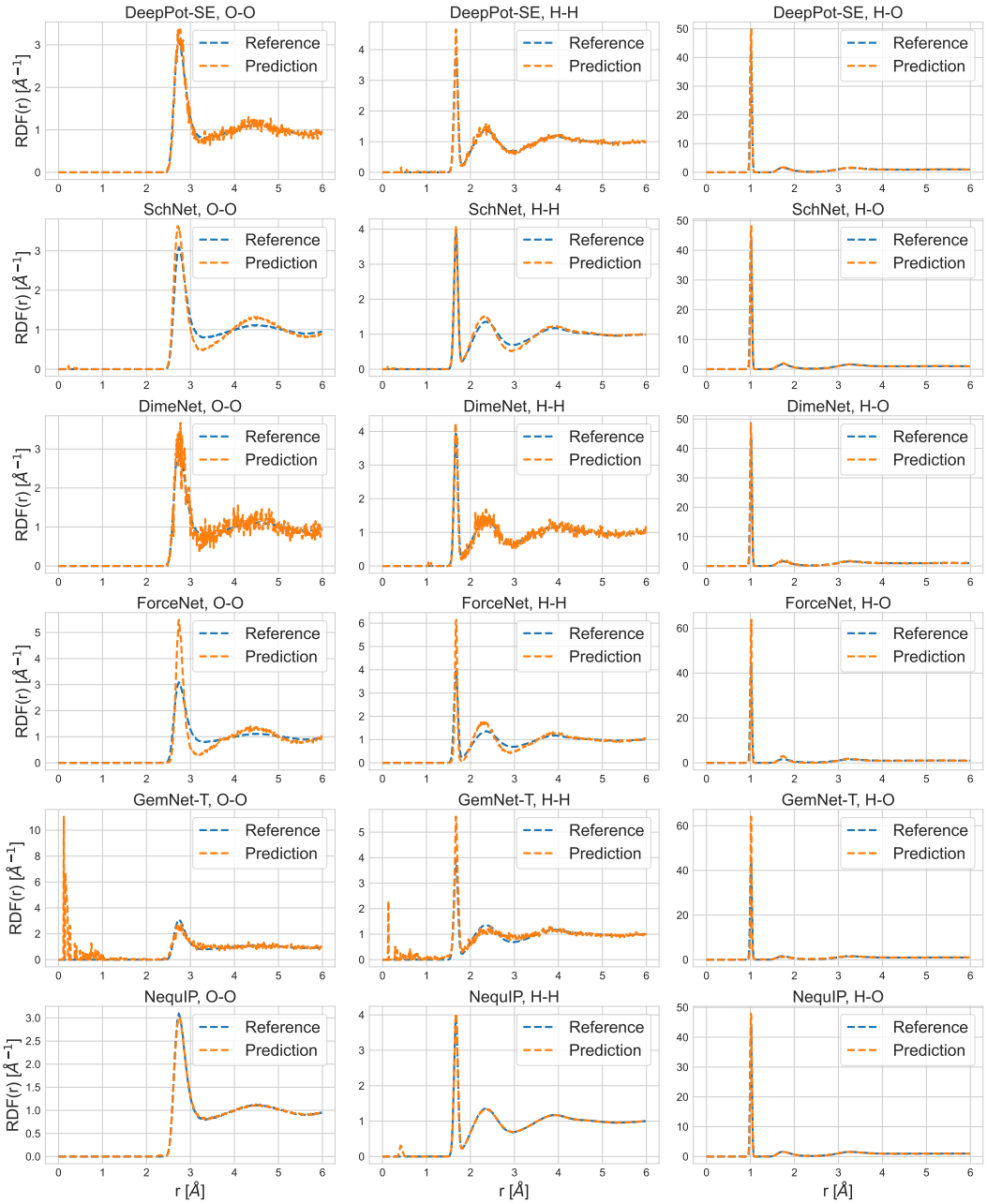

Figure 10: RDFs of Water-1k. GemNet-dT does not remain stable for more than 1 ps and is therefore not feasible for RDF computation.

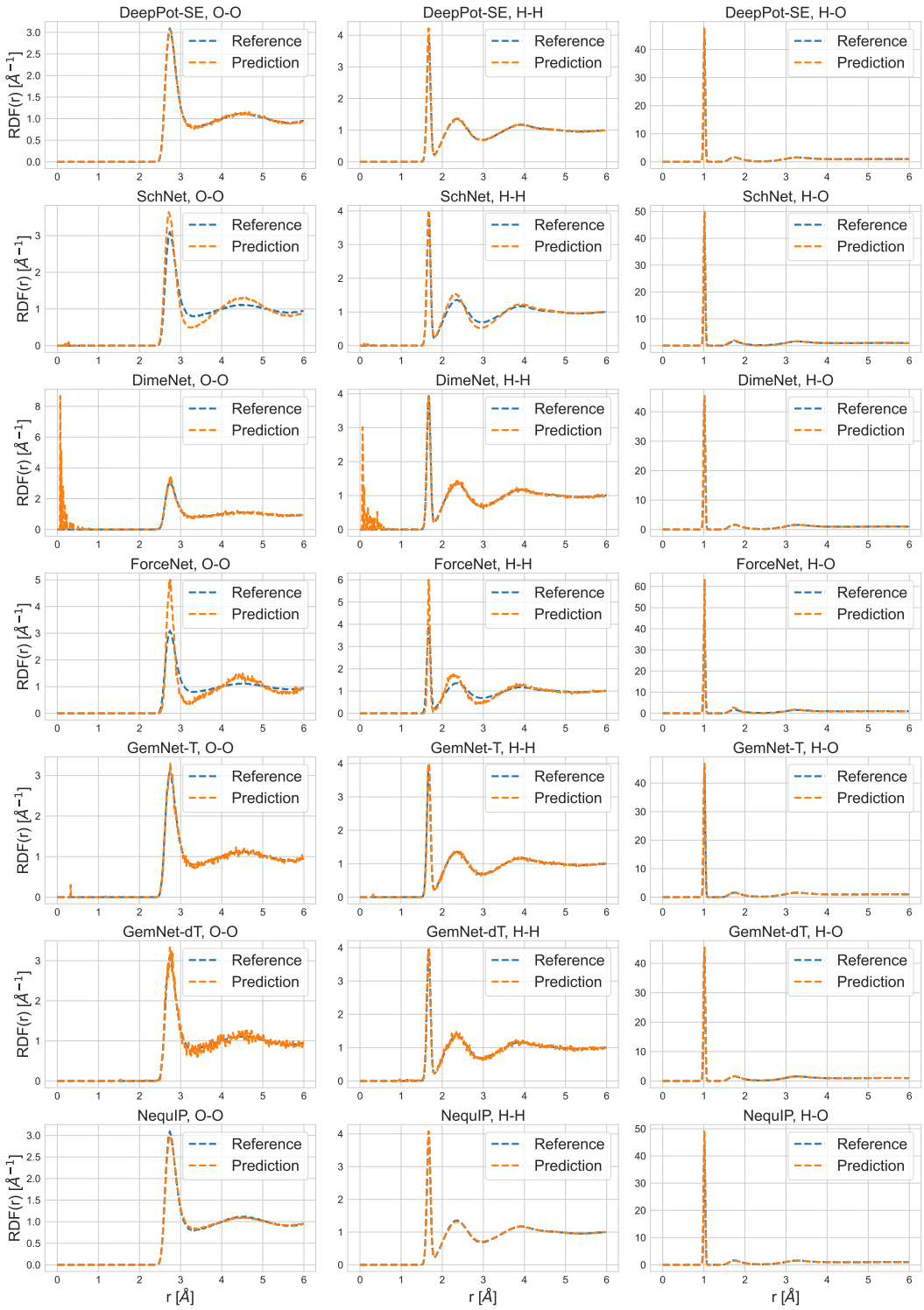

Figure 11: RDFs of Water-10k.

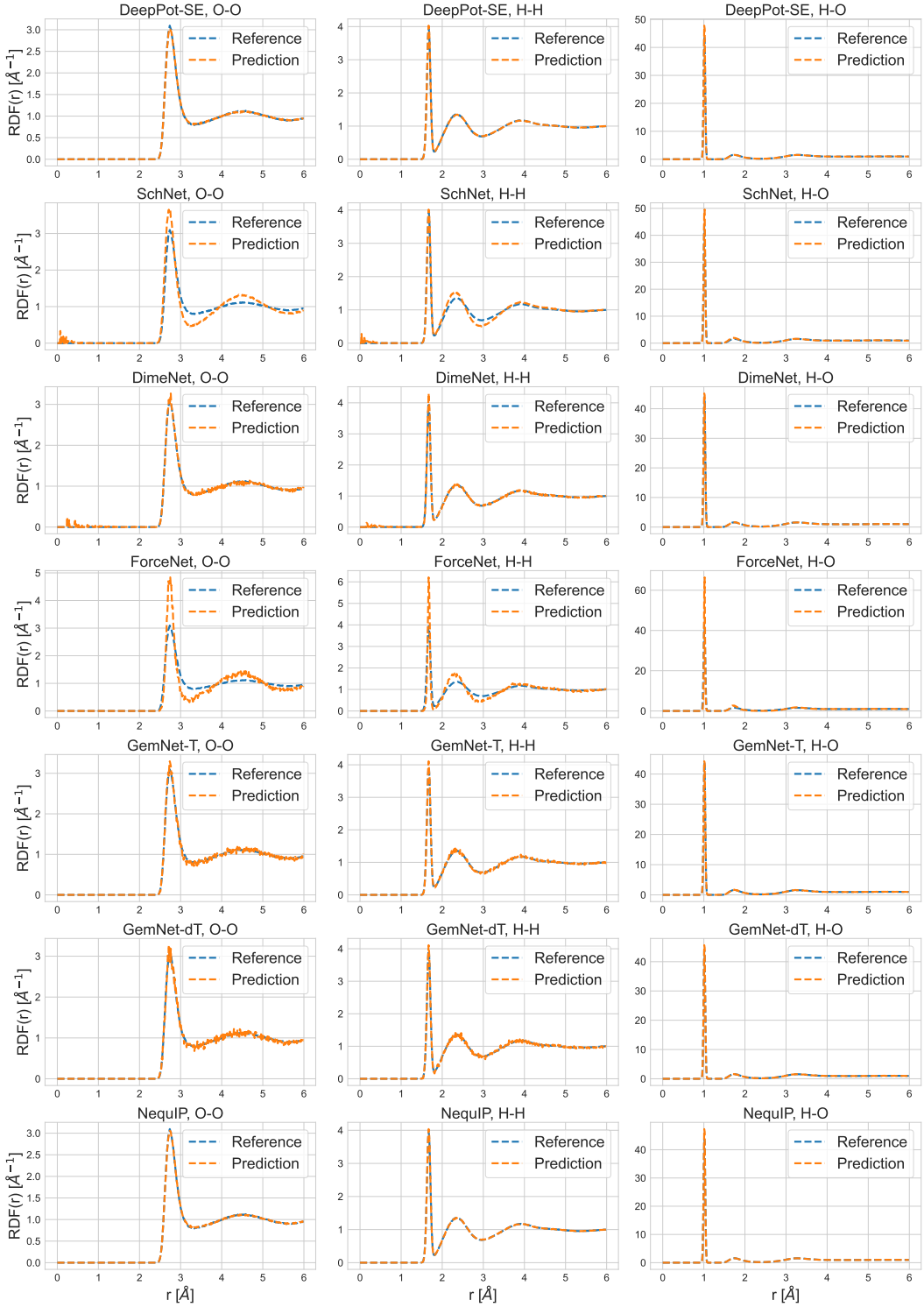

Figure 12: RDFs of Water-90k.

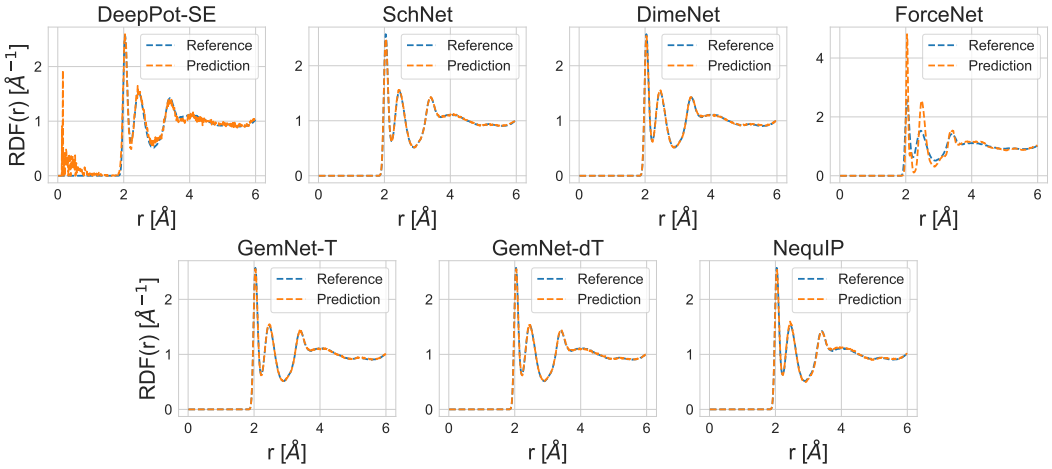

Figure 13: RDFs of LiPS.