MACHINE LEARNING OF FORCE FIELDS TOWARDS MOLECULAR DYNAMICS SIMULATIONS OF PROTEINS AT DFT ACCURACY

Christoph Brunken[†‡*] **Sebastien Boyer**[†‡*] **Mustafa Omar**[†] **Bakary Diallo**[†]

Karim Beguir[†] **Nicolas Lopez-Carranza**[†] **Oliver Bent**[†‡]

ABSTRACT

Deep learning model-based inference for molecular simulations offers a great speedup (orders of magnitude) when compared to reference quantum chemical methods such as density functional theory (DFT), along with evidence of increased accuracy compared to classical force field calculations on some systems. We demonstrate an entire atomistic simulation pipeline designed for protein systems to exploit the benefits of such models. The application of the MACE model architecture is combined with a physics-informed loss function inspired by PhysNet to improve the representation of molecular physics and account for long-range interactions explicitly. The model is trained on PhysNet's solvated fragments dataset. Our pipeline enables stable GPU-accelerated molecular dynamics (MD) simulations of small molecules within the same size as the molecules present in the dataset as well as generalisation towards larger peptides such as chignolin (175 atoms). Forces along the MD trajectories are assessed by comparison to a DFT reference. Furthermore, we present stable and accurate energy minimisations for a selection of six test molecules. Based on our results, we provide a discussion of the strengths and limitations of the approach including an outlook towards future improvements.

1 INTRODUCTION

Molecular dynamics (MD) simulations are an essential tool to determine macroscopic thermodynamic properties of molecular systems (e.g., proteins) computationally Hansson et al. (2002). In an MD simulation, the atomic coordinates are propagated through time by numerically solving Newton's equations of motion based on forces that are calculated by an interatomic potential or force field (FF), i.e., any function mapping a set of atomic coordinates to a set of atomic forces. Due to the quantum nature of molecular systems, an accurate description of their properties, including the forces, requires the application of quantum chemical (QC) methods Levine et al. (2009), such as Density-functional theory (DFT). However, these methods exhibit an unfavourable scaling with system size (typically, at least a cubic scaling) Ratcliff et al. (2017) and MD simulations can often require millions of individual force calculations on molecular systems of thousands of atoms. Therefore, pairing MD with DFT is unfeasible for simulating larger organic molecules such as proteins, and instead, forces are usually calculated with force fields based on classical physics (referred to as a classical FF), with reduced computational cost along with reduced physical accuracy for these simulations Monticelli & Tieleman (2013). Additionally, accurate atomic forces are required to run energy minimisations, which are an essential tool in computational chemistry and biology used to identify stationary points on potential energy surfaces relevant to the calculation of molecular properties, the understanding of chemical reactions, and vital to structural biology modelling.

In recent years, a tremendous advancement in the field of machine learning models as force fields (MLFF) Unke et al. (2021) has been made in an attempt to combine the accuracy of QC methods

[*]Equal Contributions

[†]InstaDeep Ltd, 5 Merchant Square, London, W2 1AY

[‡]Corresponding authors: {c.brunken, s.boyer, o.bent}@instadeep.com

such as DFT with a computational cost comparable to that of a classical FF. Among the most notable MLFF approaches are SchNet Schütt et al. (2017), BotNet Zhou et al. (2020), Nequip Batzner et al. (2022), sGDML Chmiela et al. (2019), ViSNet Wang et al. (2022), Allegro Musaelian et al. (2023) and MACE Batatia et al. (2022b). Furthermore, publicly available QC reference datasets are increasing in size and quality. Examples include QM9 Rupp et al. (2012); Blum & Reymond (2009), PhysNet's solvated fragments dataset Unke & Meuwly (2019) and the SPICE dataset Eastman et al. (2023). Despite these advances, MLFF models are still only rarely employed in applied computational studies of biosystems, most likely due to a lack of demonstrated robustness and quantitative accuracy in long-running MD simulations on large molecular systems, for which accurate QC reference data cannot be obtained directly.

In this work, we present a new ML-based atomic force prediction pipeline, combining the MACE model architecture with the physics-informed loss function and the solvated fragments dataset of PhysNet Unke & Meuwly (2019). Our approach adds an explicit treatment of long-range (LR) interactions to MACE and aims at increasing its understanding of molecular physics, creating a model that generalises well across various protein systems. Furthermore, we connect our force predictor to the JAX-MD simulation framework Schoenholz & Cubuk (2021) for GPU-accelerated protein simulations including MD and energy minimisations.

2 METHODOLOGY

2.1 MODEL AND TRAINING DATASET

We apply the MACE model architecture Batatia et al. (2022b;a) in this work and build upon a MACE implementation in JAX[1]. MACE is a state-of-the-art message passing GNN architecture predicting interatomic potentials. Its main innovation is the expansion of messages as an hierarchical body order expansion allowing for a decoupling between receptive field size, local geometry modelling and the number of message passing layers. Hence, even with a small number of model layers, higher order interactions can be captured. For more details on the original MACE model, we refer to the original work Batatia et al. (2022b).

Our model does not only predict site energies E_i for each node (i.e., atom) in the molecular graph like standard MACE, instead, we also output an atomic charge for each atom in accordance with the PhysNet Unke & Meuwly (2019) physics-informed loss function. Details on the loss function are provided in Appendix A.3.

We adopt PhysNet's main innovation, obtaining the total energy of the system as the sum of the site energies and the electrostatic correction term that explicitly models LR interactions:

$$E = \sum_{i=1}^{N} E_i + k_e \sum_{i=1}^{N} \sum_{i>j}^{N} \widetilde{q}_i \widetilde{q}_j r_{ij}^{-1} \quad . \tag{1}$$

The D3 correction term Grimme et al. (2010) that is added in the original PhysNet work is omitted as we observed numerical instability during gradient evaluation when including it within our setup. However, we plan to revisit this aspect in future work. In Eq. (1), $\widetilde{q}_i$ and $\widetilde{q}_j$ are the corrected atomic charges, i.e., the predicted atomic charges rescaled such that their sum matches the total system charge. Furthermore, k_e is the Coulomb constant and r_{ij} is the distance between atoms i and j. Moreover, to obtain more stable simulations for large systems, we apply a new message weighting strategy that differs from the one in the original MACE model (we refer to Appendix A.1 for details).

We employ MACE with two layers and 3-body correlations. The hidden layers are of type 128 scalars and 128 vectors, i.e., `128 0e 128 1o` (we refer to the original E3NN work for the notation Geiger & Smidt (2022)) and the output representation of each node is 2 scalars, i.e., `2 0e`, one representing the atomic site energy and one for the atomic charge. A constant bias for the energy per element type is added to each site energy (corresponds to the atomisation energy), however, we did not treat them as learnable but used tabulated values instead (see Table A1 in Appendix A.2). The mean and standard deviation scaling factor for the element type energy are set to 1 and not calculated.

[1] `https://github.com/ACEsuit/mace-jax` (accessed: 2023-08-11)

For building the atomic graphs, a distance cutoff of 5 Å per layer is applied (recommended by the original MACE work) and the electrostatic correction is applied for atom pairs with a distance up to 28 Å during training. Furthermore, we choose a batch size of 32 and a learning rate of 0.001.

In training our model, we use the solvated fragments dataset from PhysNet, which was generated by running semi-empirical MD simulations with the PM7 Stewart (2013) method on fragments of protein chains (up to 120 atoms) to sample geometries. Approximately 2.7×10^6 geometries are provided in total with energies, forces and dipoles calculated using DFT at the revPBE-D3(BJ)/def2-TZVP level of approximation. For our data split of training and validation sets, we ensure that geometries from the same fragments (i.e., same molecules) do not appear in the training and validation set, thus preventing any information 'leakage' into the training set. The split is performed in an 80:10:10 fashion for training, validation and testing, respectively. For speed-up, the model training is run on multiple GPUs by using data parallelism across batches.

2.2 MD SIMULATION AND ENERGY MINIMISATION

We utilise the JAX-MD Python library Schoenholz & Cubuk (2021) that provides an implementation of the necessary components to build atomistic simulations and energy minimisations based on JAX Bradbury et al. (2018) along with its just-in-time (JIT) compilation and automatic differentiation frameworks. JAX-MD is agnostic to the type of deep learning model architecture. We apply the NVT-Langevin integration algorithm Davidchack et al. (2009) for MD simulations and the Fast Inertial Relaxation Engine (FIRE) algorithm Guénolé et al. (2020) for energy minimisations, which are both available in JAX-MD. We do not apply periodic boundary conditions in this work as we work with finite systems, however, the framework allows to include them. To run efficient MD simulations with MLFF models depends on the ability to leverage JIT compilation on GPUs or TPUs. By splitting a simulation run into N_{ep} episodes of N_{s} steps each, resulting in $N = N_{\text{ep}} \cdot N_{\text{s}}$ total steps of the simulation, we can JIT-compile the N_{s} steps of one episode and perform any tasks with side-effects, e.g., logging to a remote file system or updating the dimensions of the neighbour lists, in between episodes.

3 RESULTS AND DISCUSSION

3.1 ACCURACY ON THE SOLVATED FRAGMENTS DATASET

In this section, we summarize the key metrics of training our MLFF model, presenting validation and test set performance. Overall, our predictions are correlated strongly with the ground truth, i.e., a Spearman rank correlation coefficient of 0.99 was obtained for both the total energy of the system and the norm of forces. We provide an overview of the absolute errors obtained for the validation and test set in Appendix A.4. It demonstrates that our model achieves mean and median absolute force errors of 5.1 and 3.1 kcal/(mol·Å), respectively. However, it also exhibits very large maximum errors, significantly effecting the mean error. Due to this observation, we conducted an analysis of the dataset, which is reported in Appendix A.5 and ideas on improving this dataset in future work are discussed in section A.11 of the Appendix.

3.2 MD SIMULATIONS

To assess the generalisation of the developed MLFF model to unseen structures beyond the test set, we conduct MD simulations using the complete MD pipeline and validate (1) qualitatively the stability of the simulations (i.e., molecules remain intact) and (2) quantitatively the accuracy of the predicted forces compared to DFT, for a selection of snapshots from the trajectory. As test systems, we select (i) two fragments from the original test set, (ii) a selection of small non-peptide molecules, and (iii) additional unseen peptide structures. The complete list of systems can be found in Table 1, in which the quantitative MD results are presented. We ran 100 ps simulations with a timestep of 1 fs and at a temperature of 300 K, and subsequently extracted 20 equidistant snapshots from the simulations to calculate DFT reference forces with the PySCF Sun et al. (2018) program. For each snapshot, a MAE value was obtained and the mean, standard deviation, and maximum of these values are presented in Table 1. For the peptide chignolin with 175 atoms, we restricted the DFT reference calculation to the initial structure of the simulation for which we obtained an MAE in forces of 2.4 kcal/(mol·Å).

For the qualitative analysis of all simulated molecular systems, we observed stable simulations. Visual inspection of the trajectories mostly revealed physically reasonable simulations, however, a few untypical proton transfers were observed. In particular, protonated amino groups ($-NH_3^+$) were found to be strong proton donors and deprotonated carboxyl groups ($-CO_2^-$) strong proton acceptors. In Figure A3 in the Appendix, we depict an example of an unphysical proton transfer that was observed during the MD simulation of chignolin. A proton was transferred from a nitrogen atom to the deprotonated carboxyl group. As this leaves the nitrogen atom behind with a formally negative charge, this proton transfer is unexpected and unphysical. Furthermore, we performed simulations at elevated temperatures of 500 K as a stress test and observed that the simulations of most systems remain stable. Only for the decaalanine system, the very high temperature leads to a decomposition of the molecule after about five picoseconds of stable simulation time.

The obtained force errors with respect to the DFT reference are between 0.8 and 4.0 kcal/(mol·Å) on average, i.e., lower than the metrics reported in section 3.1. We suspect this mainly to be a consequence of the lower MD temperature that was used to sample the structures (300 K instead of 1000 K). Furthermore, we obtain more accurate forces for the peptide-like structures than for the generic organic molecules, which can be explained by the peptide-derived nature of the training set. We also compared our forces to a GFN2-xTB reference, resulting in MAE values between 4.5 and 7.5 kcal/(mol·Å), which underlines (a) the spread that exists even between DFT and semi-empirical forces as well as (b) how our model is trained to reproduce the ones from accurate DFT calculations.

Moreover, we selected the molecules chignolin and decaalanine to perform long-running 5 ns MD simulations, (1) to verify the stability of the simulations and (2) generate Time-Independent Component Analysis (TICA) and Root-Mean-Squared-Deviation (RMSD) plots (see Appendix A.7). We also plot the velocity autocorrelation function for the first picosecond of the decaalanine simulation (see Appendix A.10).

Table 1: Mean absolute error (MAE) of the MLFF forces for snapshots along an MD trajectory with respect to revPBE-D3(BJ)/def2-TZVP. For each molecule, we provide the mean MAE of the snaphots, the standard deviation σ and the maximum MAE (all values in kcal/(mol·Å)).

MOLECULE	MEAN ERROR	σ	MAXIMUM ERROR
TEST SET FRAGMENT 1	0.809	0.094	1.101
TEST SET FRAGMENT 2	0.953	0.099	1.256
TRIPEPTIDE ACY	2.120	0.517	4.241
DECAALANINE (FOLDED)	1.854	0.225	2.693
DECAALANINE (UNFOLDED)	1.806	0.075	1.944
PROPYL-ANTHRACENE	2.943	0.188	3.263
ASPIRIN	3.017	0.272	3.470
CURCUMIN	3.951	0.352	4.738

For chignolin, the execution time is approximately 30 seconds (see Figure A5 in the Appendix) for each picosecond of simulation time with our MLFF/MD pipeline on a single NVIDIA Tesla V100 SXM3 GPU. Compared to a generic force field such as GFN-FF Spicher & Grimme (2020), this is around ten times slower, however, compared to the semi-empirical method GFN2-xTB, our pipeline provides more than a 20-times speed-up (reference simulations are performed with the xTB program Bannwarth et al. (2021)). We emphasize that there exists potential to further accelerate our pipeline, for example, by parallelising inference on multiple GPUs, running it on TPUs, or applying general code optimisations. In order for such improvement strategies to have a significant impact on the execution speed of the simulations, it is essential that the scaling of the model with system size follows a linear relationship (such as in classical force fields with an LR interaction cutoff distance) rather than a quadratic or cubic one (such as for semi-empirical QC methods). To confirm that our model exhibits this linear scaling, we visualise the execution times with respect to the number of atoms for our test molecules in Figure A5. As we expect the bottleneck of the model inference to be attributed to the GNN and not to the LR interaction correction term, the 5 Å graph cutoff should be small enough such that we observe a linear scaling already for these test molecules comprised of less than 200 atoms. The results presented in Figure A5 confirm this assumption.

3.3 ENERGY MINIMISATIONS

In this section, we analyse the performance of the MLFF model for accurate energy minimisations (structure optimisations). As mentioned in section 2.2, we have applied the FIRE algorithm Bitzek et al. (2006) to six test molecules, namely; tripeptide ACY; propyl-anthracene; curcumin; aspirin; and two fragments from the test set included in section 3.2. The FIRE algorithm has been run with an initial timestep of 0.5 fs, a maximum timestep of 2 fs, and all other parameters set to the defaults of the JAX-MD engine. In a successful and stable energy minimisation, the atomic forces should converge at zero, hence, resulting in the atom positions to remain unchanged after convergence.

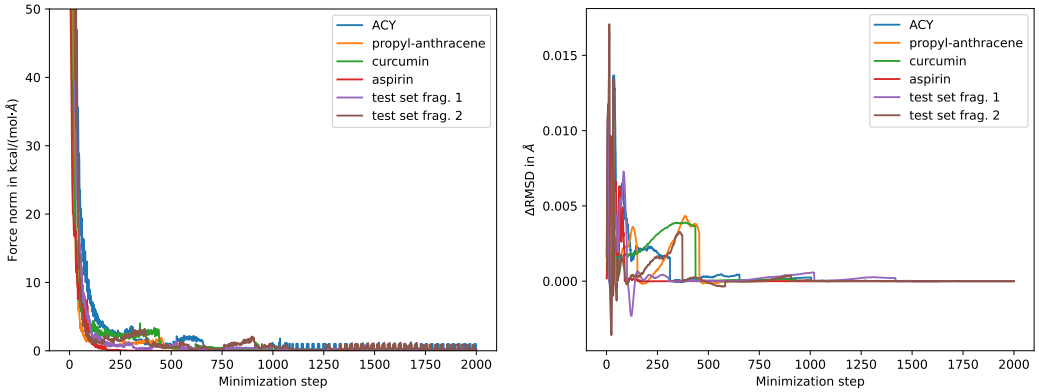

Figure 1: Norm of forces (left) and change in RMSD (right) during an MLFF energy minimisation.

In Figure 1, we demonstrate that the developed MLFF model fulfills these two criteria. After 500 minimisation steps, we observe a ΔRMSD between steps of less than 0.001 Å for all molecules and observe full convergence after at most 1000 steps for all but one of the test set fragments. Note that both test set fragments consist of two non-covalently bonded substructures that are typically prone to more difficult convergence. Furthermore, we observe a stable convergence of the forces to zero. For two molecules, the tripeptide and the second test set fragment, we observe small fluctuations in forces after 1000 steps, which appear to not result in significant fluctuations in the positions. We contribute those to numerical instabilities that can be potentially resolved by fine-tuning the parameters of the FIRE algorithm.

To assess the quality of the minimisation result, we optimised all six test structures with DFT employing the PSI4 Parrish et al. (2017) quantum chemistry program with its default settings. To these ground truth results, we compare (1) the initial structure, (2) a structure optimised with the GFN2-xTB method using the xTB program Bannwarth et al. (2021), and (3) the final structure after energy minimisation with the MLFF model. We compute the MAE of bond distances and bond angles to assess local structural agreement and the RMSD to assess global conformational alignment. As presented in Table A3 in the Appendix, we observe that the errors on bond distances are smaller for the MLFF method as compared with GFN2-xTB for all tested molecules. The bond angle errors of GFN2-xTB and the MLFF are similar. For RMSD, the MLFF exhibits similar errors as GFN2-xTB as well. In particular, we observe more accurate results for the test set fragments and the tripeptide than for the other systems, which can be explained by the peptide-based training set of the MLFF model.

In Figure A4 in the Appendix, we present a visual comparison of the initial structure to the final structures obtained from energy minimisations with the MLFF model and the DFT reference. In the two selected examples, it becomes clear that the optimised structures differ significantly from the initial structures, however, the structures of the MLFF model compared to the DFT reference, are similar, underlining the MLFF model's ability to run stable and accurate energy minimisations for the systems investigated in this work.

4 CONCLUSION

In this work, we introduced a new approach to combine the MACE model architecture and PhysNet physics-informed loss function, to build and train an MLFF model on the solvated fragments

dataset, and run energy minimisations and MD simulations based on the JAX-MD framework. We demonstrated that we are able to run long and stable MD simulations for small molecules as well as generalise to larger systems such as chignolin. In Appendix A.11 we elaborate further the outlook of this work and propose strategies for improvement towards an accelerated MD pipeline for simulating biomolecules at DFT accuracy.

REFERENCES

Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 11(2):e1493, 2021.

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*, 2022a.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35:11423–11436, 2022b.

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), may 2022.

Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical review letters*, 97(17):170201, 2006.

L. C. Blum and J.-L. Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, 131:8732, 2009.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Stefan Chmiela, Huziel E Sauceda, Igor Poltavsky, Klaus-Robert Müller, and Alexandre Tkatchenko. sgdml: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications*, 240:38–45, 2019.

Ruslan L. Davidchack, Richard Handel, and M. V. Tretyakov. Langevin thermostat for rigid body dynamics. *The Journal of Chemical Physics*, 130(23), jun 2009.

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.

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, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of chemical physics*, 132(15), 2010.

Julien Guénolé, Wolfram G. Nöhring, Aviral Vaid, Frédéric Houllé, Zhuocheng Xie, Aruna Prakash, and Erik Bitzek. Assessment and optimization of the fast inertial relaxation engine (fire) for energy minimization in atomistic simulations and its implementation in lammps. *Computational Materials Science*, 175:109584, apr 2020.

Tomas Hansson, Chris Oostenbrink, and Wilfred F. van Gunsteren. Molecular dynamics simulations. *Current opinion in structural biology*, 12(2):190–196, 2002.

Arthur Kosmala, Johannes Gasteiger, Nicholas Gao, and Stephan Günnemann. Ewald-based long-range message passing for molecular graphs. *arXiv preprint arXiv:2303.04791*, 2023.

Ira N. Levine, Daryle H. Busch, and Harrison Shull. *Quantum chemistry*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ, 2009.

Luca Monticelli and D. Peter Tieleman. Force fields for classical molecular dynamics. *Biomolecular simulations: Methods and protocols*, pp. 197–213, 2013.

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. *Nature Communications*, 14(1):579, 2023.

Robert M. Parrish, Lori A. Burns, Daniel G. A. Smith, Andrew C. Simmonett, A. Eugene DePrince III, Edward G. Hohenstein, Ugur Bozkaya, Alexander Yu Sokolov, Roberto Di Remigio, Ryan M. Richard, et al. Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability. *Journal of chemical theory and computation*, 13(7): 3185–3197, 2017.

Laura E. Ratcliff, Stephan Mohr, Georg Huhs, Thierry Deutsch, Michel Masella, and Luigi Genovese. Challenges in large scale quantum mechanical calculations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(1):e1290, 2017.

M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108:058301, 2012.

Samuel S. Schoenholz and Ekin D. Cubuk. Jax, md a framework for differentiable physics. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124016, 2021.

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.

Sebastian Spicher and Stefan Grimme. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angewandte Chemie International Edition*, 59(36):15665–15673, 2020.

James J. P. Stewart. Optimization of parameters for semiempirical methods vi: more modifications to the nddo approximations and re-optimization of parameters. *Journal of molecular modeling*, 19: 1–32, 2013.

Qiming Sun, Timothy C Berkelbach, Nick S Blunt, George H Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D McClain, Elvira R Sayfutyarova, Sandeep Sharma, et al. Pyscf: the python-based simulations of chemistry framework. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1340, 2018.

Oliver T. Unke and Markus Meuwly. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of Chemical Theory and Computation*, 15(6): 3678–3693, may 2019.

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

Yusong Wang, Shaoning Li, Xinheng He, Mingyu Li, Zun Wang, Nanning Zheng, Bin Shao, Tong Wang, and Tie-Yan Liu. Visnet: a scalable and accurate geometric deep learning potential for molecular dynamics simulation. *arXiv preprint arXiv:2210.16518*, 2022.

Jiawei Zhou, Zhiying Xu, Alexander M Rush, and Minlan Yu. Automating botnet detection with graph neural networks. *arXiv preprint arXiv:2003.06344*, 2020.

A APPENDIX

A.1 MESSAGE WEIGHTING STRATEGY

As mentioned in the main text, we apply a new message weighting strategy that differs from the one in the original MACE model. We observed that in our setup it is necessary to obtain stable simulations for large molecular systems of more than 100 atoms. Instead of weighting the messages with a learned or fixed average number of neighbours (original MACE), we employ,

$$m_i = \sum_{j \in \mathfrak{N}(i)} \frac{m_{i,j}}{\sqrt{d_i * d_j}} \quad , \tag{2}$$

where m_i is the aggregated message at node i. $\mathfrak{N}(i)$ is the set of nodes in the neighbourhood of i and $m_{i,j}$ the contribution of node j via edge $e_{i,j}$ to the final message at node i. Finally, d_i and d_j are the degree of nodes i and j, respectively.

A.2 ELEMENT-SPECIFIC ENERGY BIAS

As mentioned in the main text, a constant bias for the energy per element type (corresponds to the atomisation energy) is added to each site energy, however, we did not treat them as learnable but used tabulated values instead. These values are presented in Table A1. The mean and standard deviation scaling factor for the element type energy are set to 1 and not calculated.

Table A1: Element-specific energy bias added to the atomic site energies. These correspond to the element-specific atomisation energies.

ELEMENT	ATOMIC SITE ENERGY BIAS (EV)
HYDROGEN	-13.587
CARBON	-1029.489
NITROGEN	-1484.981
OXYGEN	-2041.982
SULFUR	-10831.265

A.3 PHYSICS-INFORMED LOSS FUNCTION

In our work, we adopt the physics-informed loss function introduced by PhysNet. in Eq. (**??**). The loss L can be understood as a sum of individual contributions which are the energy loss L_E, force loss L_F, charge loss L_Q, and dipole loss L_P.

$$L = \omega_E \, L_E + \frac{\omega_F}{3N} \, L_F + \omega_Q \, L_Q + \frac{\omega_P}{3} \, L_P \quad . \tag{3}$$

In Eq. (3), N is the number of atoms. In our setup, we use the values $\omega_E = \omega_Q = \omega_p = 1$ and $\omega_F = 100$ for the weights. Note that we put a larger weight on L_F due to the importance of accurate forces for atomistic simulations. The contributions L_E and L_F are defined as,

$$\begin{aligned} L_E &= \left| E - E^{\text{ref}} \right| \quad , \\ L_F &= \sum_{i=1}^{N} \sum_{\alpha=1}^{3} \left| -\frac{\partial E}{\partial r_{i,\alpha}} - F_{i,\alpha}^{\text{ref}} \right| \quad . \end{aligned} \tag{4}$$

The energy consistency term L_E ensures that the predicted energy E matches the reference energy E^{ref}. Note that the total energy E of the system is a sum of site energies E_i (for the i-th atom) in our MACE model and the forces are calculated as the negative gradient of the energy with respect to the atom coordinates r_i. $F_{i,\alpha}^{\text{ref}}$ is the reference force on atom i in Cartesian direction α (i.e., $\alpha = x, y, z$).

The terms L_Q and L_P of the loss function are defined as,

$$L_Q = \left| \sum_{i=1}^{N} q_i - Q^{\text{ref}} \right| \quad ,$$
$$L_P = \sum_{\alpha=1}^{3} \left| \sum_{i=1}^{N} q_i r_{i,\alpha} - P_\alpha^{\text{ref}} \right| \quad , \tag{5}$$

and enforce that the sum of atomic charges q_i matches the reference charge of the full system Q^{ref} and that the predicted dipole moments (calculated from the predicted atomic charges) align with the reference P_α^{ref} ($\alpha = x, y, z$). The PhysNet loss function additionally includes a non-hierarchical correction term to the loss function, however, this targets a particular concern within the PhysNet architecture and is thus omitted from our loss function L.

A.4 Results on validation and test set

We provide an overview of the absolute errors obtained for the validation and test set in Table A2.

Table A2: Analysis of absolute errors (AE) for the MLFF predictions on energies, forces, and dipole moments. Total charge is not included as the charge correction scheme renders the associated metrics nontrivial to interpret.

	Energy in kcal/mol		Forces in kcal/(mol·Å)		Dipole moment in Debye	
	validation	test	validation	test	validation	test
Mean AE	22.46	22.60	5.46	5.08	0.12	0.13
Maximum AE	255.14	1051.60	699.88	804.53	7.17	14.87
25th percentile AE	6.13	6.37	1.84	1.80	0.06	0.07
Median AE	13.74	13.57	3.18	3.06	0.10	0.10
95th percentile AE	68.44	75.41	16.65	14.87	0.28	0.28

A.5 Analysis of solvated fragments dataset

As mentioned in section 3.1, we observe that for some structures, the MLFF model exhibits low correlation and large average errors. Furthermore, we found that generic physics-based methods such as GFN-FF and GFN2-xTB also exhibit these errors compared to the reference values given in the dataset. A closer investigation of these structures demonstrate that they contain unusual bond patterns that are either (a) not found in typical protein structures (e.g., isolated molecular hydrogen H_2) or (b) not chemically reasonable (i.e., geometries with enormously high energies). Figure A1 provides four example structures of this kind from the solvated fragment dataset.

We conjecture that these structures arise from the fact that a quantum chemical method (i.e., PM7) combined with a high temperature of 1000 K was applied in the MD to sample the fragment configurations. This setup can lead to a variety of chemical reactions to occur as part of the MD, as well as reaching high energy conformations. To some extent this is desired such that a more diverse space of conformations is sampled (also far away from equilibrium), however, it results in two issues, namely that (i) DFT reference calculations may not converge properly, which can be hard to detect when monitoring millions of calculations, and (ii) chemical reactions may lead to structures that are not representative anymore of the dataset's original purpose (e.g., isolated H_2 molecules as part of protein fragments). Furthermore, these structures can hamper the MLFF's training procedure. In future work, we propose to add additional quality checks for such datasets to avoid these structures, either at dataset generation time or as a postprocessing step.

Furthermore, we take a look at the distribution of system sizes present in the solvated fragments dataset. It is presented in the Appendix in Figure A2 and shows that most structures in the dataset consist of 15 to 40 atoms. Relatively few structures contain more than 60 atoms. Implications of this and further perspectives on improvement are discussed in section A.11.

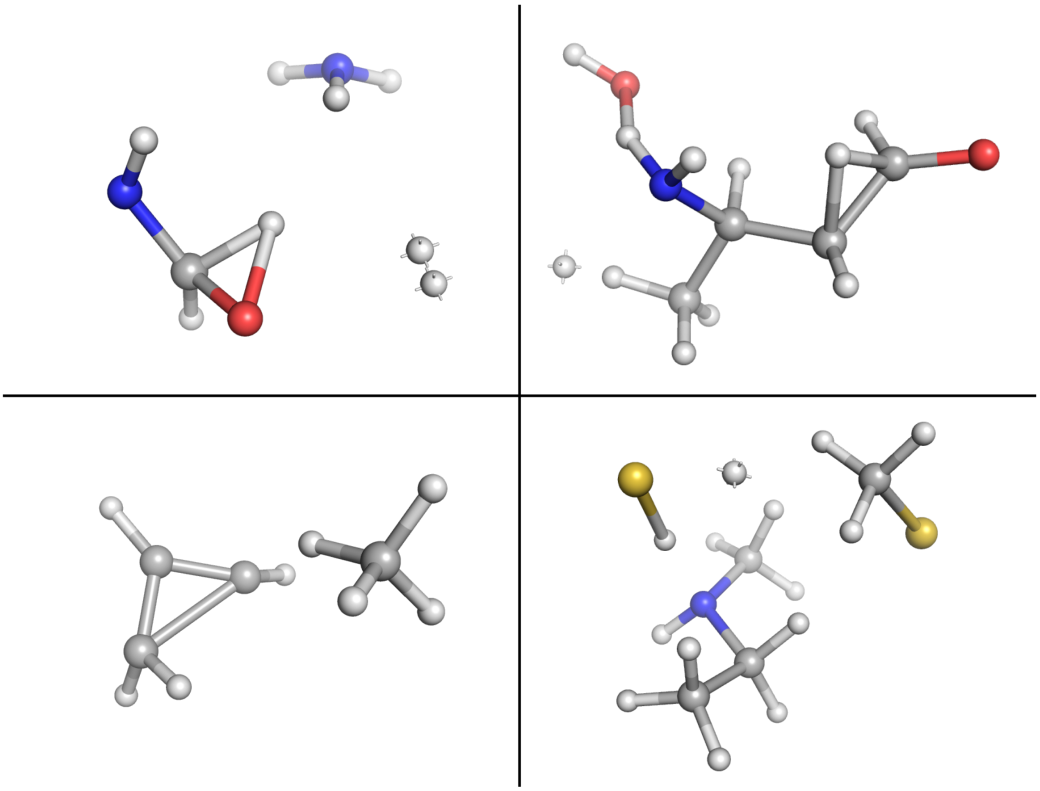

Figure A1: Four example structures from our reference dataset Unke & Meuwly (2019), which show large errors in GFN-FF, GFN2-xTB and MLFF inferred forces compared to the DFT reference forces provided. The discrepancy may be explained by a significant deviation from equilibrium atom positions and bonding patterns, which are highly uncommon for biological systems. These visualised molecular have the indices 2753, 9963, 6538, and 9957 (from top left proceeding clockwise), in the dataset.

A.6 Observed proton transfers

In Figure A3, we depict an example of an unphysical proton transfer that was observed during the MD simulation of chignolin. A proton was transferred from a nitrogen atom to the deprotonated carboxyl group. As this leaves the nitrogen atom behind with a formally negative charge, this proton transfer is unexpected and unphysical.

A.7 Long-running MD simulations of large systems

To evaluate the qualitative stability of long-running simulations with an MLFF model, we conduct simulations with a simulation time of 5 ns for chignolin and decaalanine. As for all other simulations in this work, the timestep is 1 fs. The temperature was set to 350 K, because we expect an increased temperature to result in enhanced conformational sampling. Furthermore, it allows us to directly compare our results on chignolin to Ref. Doerr et al. (2021). We also simulated the same molecules at 300 K, however, we did not observe a qualitatively different behaviour. Based on the trajectories, we generate TICA and RMSD plots for each simulation. These plots are presented in the Appendix in Figure A6 to A9. The results demonstrate that we are able to run stable long-running MD simulations for chignolin and decaalanine with the MLFF model. However, based on the TICA and RMSD plots, we infer that the energy landscape of the developed MLFF model clearly favours one molecular conformation and a diverse sampling of three different states, as seen in Ref. Doerr et al. (2021), is neither observed for chignolin nor for decaalanine. This may be attributed (i) to the significantly

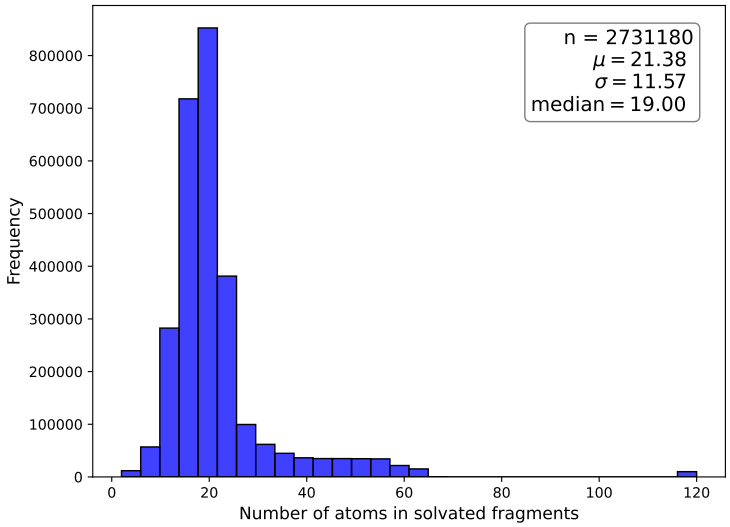

Figure A2: Size distribution of structures in the solvated fragments dataset. The total number of structures n, mean μ, standard deviation σ, and the median are presented.

shorter simulation time compared to the reference for chignolin or (ii) to differences in the potential energy landscape. Note that the long execution times for chignolin simulations with our current implementation prevented us from running significantly longer simulations as part of this work, however, we plan to deliver such results in future work. For decaalanine, we observe a similar behaviour. Comparing to the RMSD plot given in Ref. Unke & Meuwly (2019), we again observe a smaller number of different conformations visited during our MLFF−MD run. Furthermore, we observe a conformation in between the completely unfolded and folded structures to be favoured, while the original PhysNet model favours the folded conformation Unke & Meuwly (2019). To gain a solid understanding of why the aforementioned differences are observed, further experiments and direct comparison to reference methods will be necessary.

A.8 Energy Minimisations

The results of the energy minimisation results are presented in Table A3 and Figure A4.

Table A3: Molecular structures optimised with DFT (revPBE-D3(BJ)/def2-TZVP) compared to the unoptimised ones, and the ones optimised with GFN2-xTB and the MLFF model. We present the RMSD of positions in Å, MAE of bond distances Δr in Å, and MAE of bond angles $\Delta\alpha$ in degrees.

Molecule	DFT − initial structure			DFT − GFN2-xTB			DFT − MLFF		
	RMSD	Δr	$\Delta\alpha$	RMSD	Δr	$\Delta\alpha$	RMSD	Δr	$\Delta\alpha$
test set fragment 1	1.211	0.035	4.309	1.149	0.013	0.388	0.962	0.001	0.219
test set fragment 2	2.391	0.042	5.348	1.710	0.013	0.331	2.331	0.002	0.511
tripeptide acy	1.145	0.019	2.788	0.413	0.012	0.612	0.194	0.003	0.652
propyl-anthracene	1.028	0.015	1.206	0.951	0.013	0.318	1.604	0.005	0.457
aspirin	0.322	0.006	1.656	0.135	0.014	0.551	0.231	0.006	0.705
curcumin	1.149	0.009	1.297	0.371	0.012	0.431	0.591	0.005	0.575

A.9 Execution times

We present the execution times for 100 ps MD simulations for a variety of test systems in Figure A5.

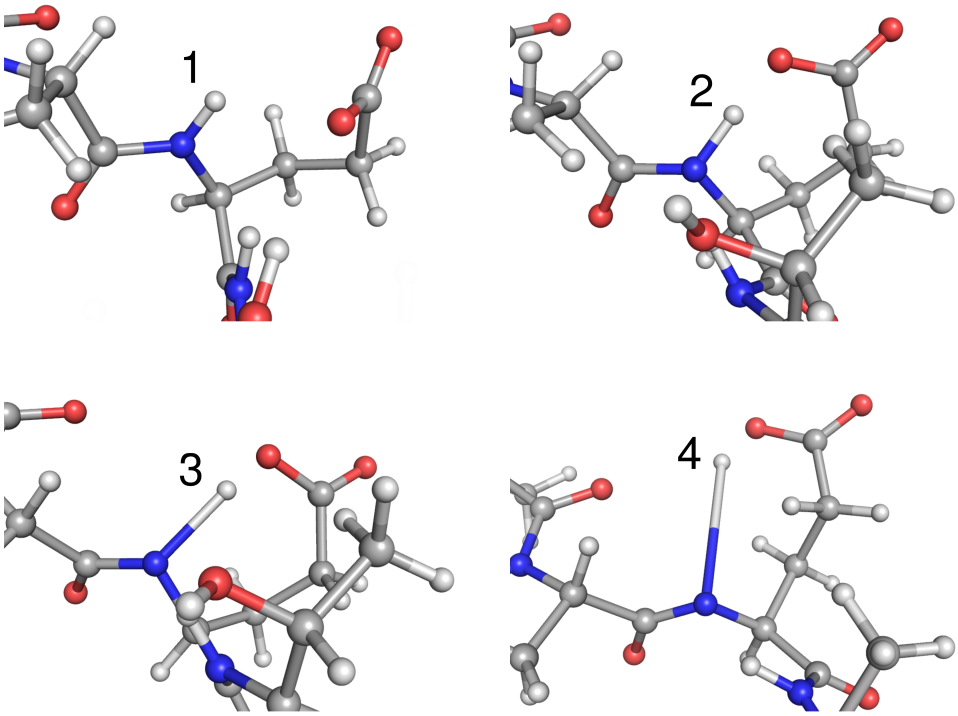

Figure A3: Example of an unphysical proton transfer observed during an MD simulation of chignolin. The depicted snapshots correspond to (1) the initial structure, (2) 18 fs, (3) 24 fs, and (4) 250 fs into the simulation.

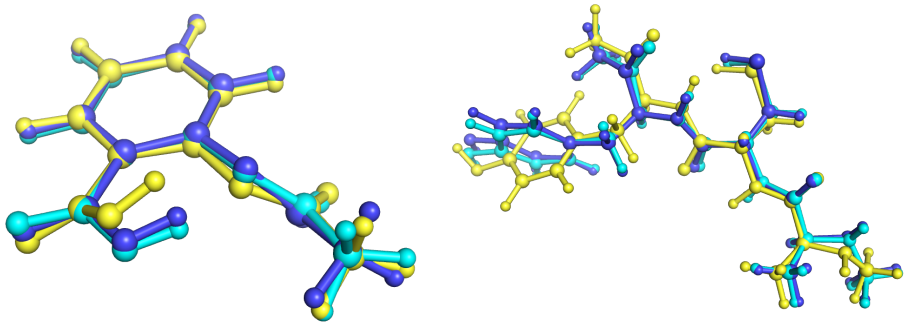

Figure A4: Comparison of optimised to initial structures for energy minimisations with the MLFF model and a DFT reference. The initial structures are depicted in yellow, the MLFF structure in cyan, and the DFT reference structure in blue. The two examples presented are aspirin (left) and the tripeptide ACY (right).

A.10 VELOCITY AUTOCORRELATION FUNCTION

For the first picosecond of the decaalanine simulation, we provide the autocorrelation function of the velocities VACF(t),

$$\text{VACF}(t) = \frac{1}{N} \sum_{i=1}^{N} v_i(t=0) \cdot v_i(t) \quad , \tag{6}$$

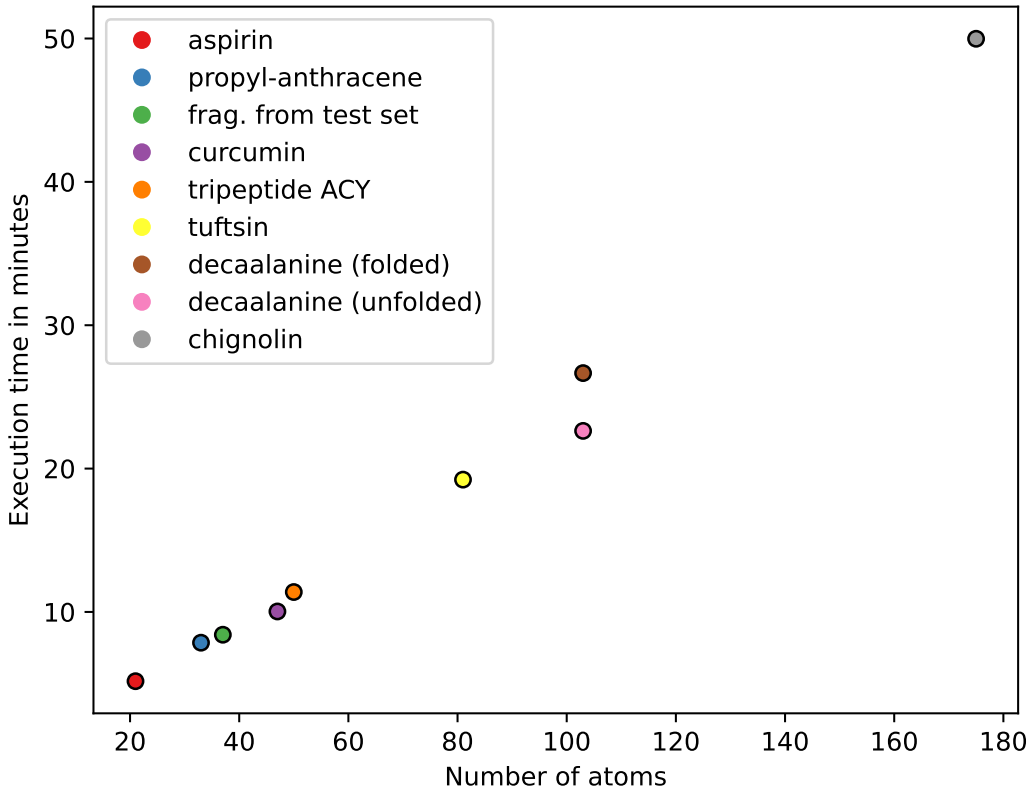

Figure A5: Execution time for running a 100 ps MD simulation with respect to system size. The data is presented for those systems for which the quantitative accuracy along the same trajectory is provided in Table 1. The test set fragments 1 and 2 from Table 1 have the same number of atoms and a very similar execution time, hence, we only present test set fragment 1 in this figure for clarity. Based on this plot, we demonstrate the linear scaling capabilities of the MLFF model.

with N being the number of atoms and $v_i(t)$ as the velocity vector of atom i at time t. The VACF for the decaalanine MD is depicted in Figure A10.

A.11 Outlook

First, the training dataset can be improved. The solvated fragments dataset has the drawback that it mostly consists of fragments between 15 and 40 atoms in size (see Appendix A2) and the average number of neighbours of atoms in the dataset is thus also in this range. However, in larger systems, in particular in the center of proteins, the average number of neighbours is much larger and we expect an increased ability to generalise by having a more balanced dataset in terms of fragment size, up to at least 120 atoms. Moreover, the training of the atomic charges that are inputs to the electrostatic interactions could as well benefit from a dataset that includes more structures with large interatomic distances. Pre-training of the model on a dataset obtained from semi-empirical QC calculations (e.g., with GFN2-xTB) would allow to include even larger molecular structures, i.e., up to several hundred atoms. Additionally, we recommend a thorough quality assessment and filtering of the dataset to avoid unreasonable structures, e.g., by analysing the energies for all geometries of a given fragment, similar to the filtering proposed by the SPICE dataset for strained molecules Eastman et al. (2023). Also, one could apply general-purpose methods with fixed bond topology like GFN-FF instead of PM7 to obtain physically reasonable MD sampling while avoiding that chemical reactions can occur as part of the sampling process.

Second, we value PhysNet's approach to add other physical properties (charges, dipole) to the loss function and suggest to extend it. For example, the atomic charges predicted by the model are trained

only indirectly via the electrostatic contribution to the energy and the total charge consistency. Instead one could compare these charges directly to charges obtained from QC calculations, which would likely guide the model's charge predictions towards physically reasonable values more efficiently and reduce the overall risk of over-fitting. Likewise, one may add other physical properties calculated by DFT calculations, for example, those derived from the electron density, to the model's predictions to obtain a more physics-informed model. Accurate treatment of LR interactions could also be achieved by architectural changes such as incorporating Ewald message passing layers Kosmala et al. (2023).

Lastly, we point out that this work is targeted towards the simulation of protein systems which is reflected by the employed dataset. However, the presented setup can, in principle, be extended to systems with different chemical compositions or to models that calculate properties of chemical reactions (as MLFF models should not be limited by defining a fixed set of chemical bonds). Finally, generically trained MLFF models have the ability to be fine-tuned in a system-focused manner, if additional system-specific reference data is available.

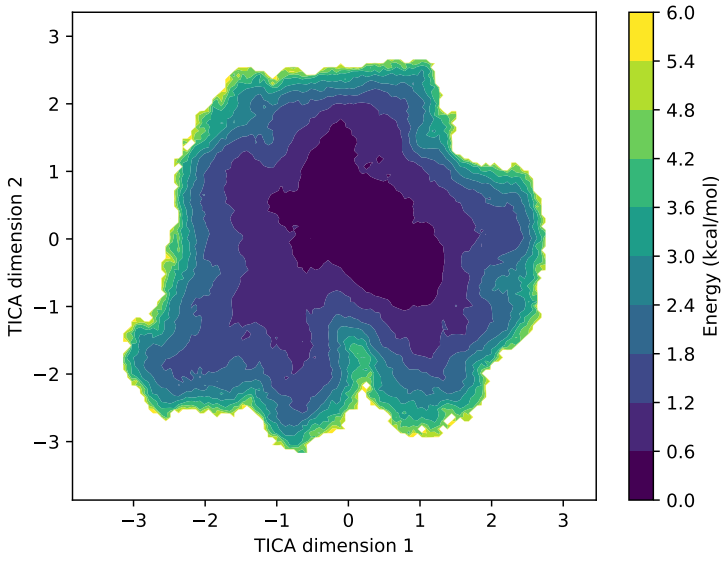

Figure A6: TICA plot generated from a 5 ns MD simulation of chignolin at 350 K.

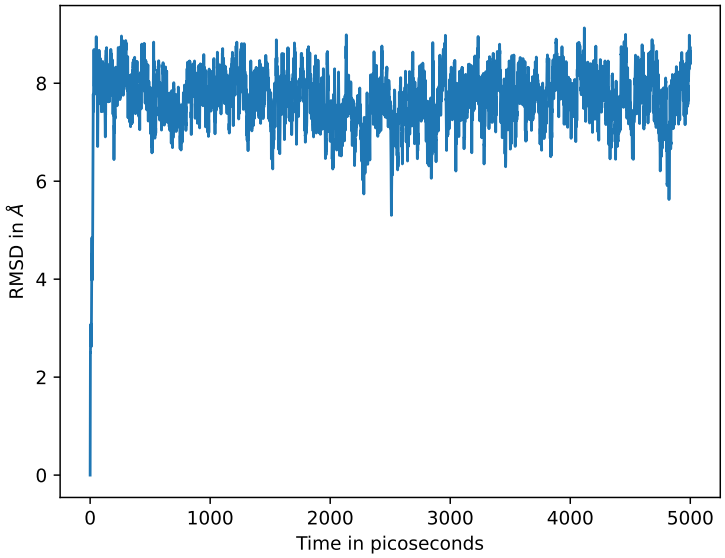

Figure A7: RMSD plot for a 5 ns MD simulation of chignolin at 350 K.

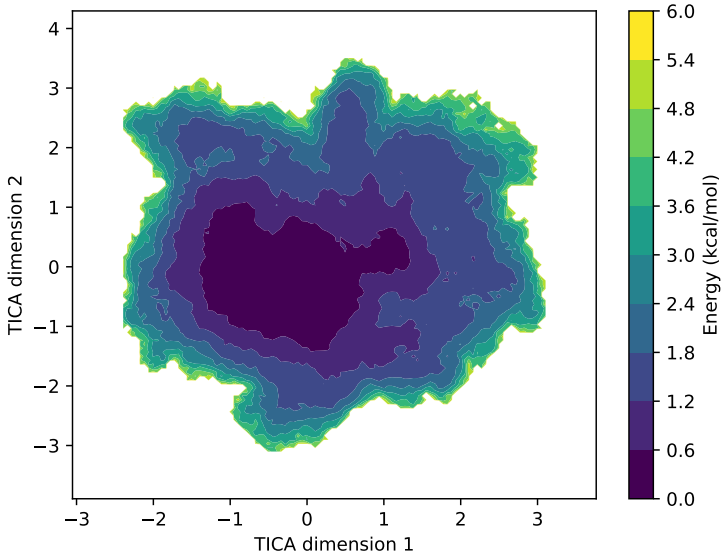

Figure A8: TICA plot generated from a 5 ns MD simulation of decaalanine at 350 K.

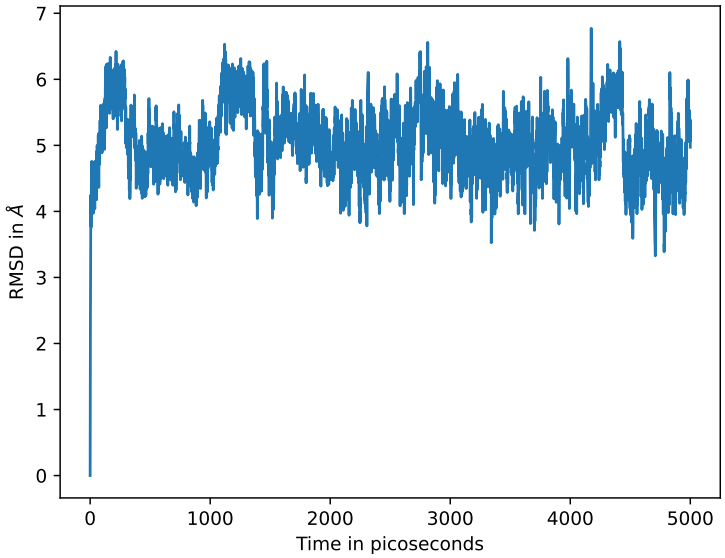

Figure A9: RMSD plot for a 5 ns MD simulation of decaalanine at 350 K.

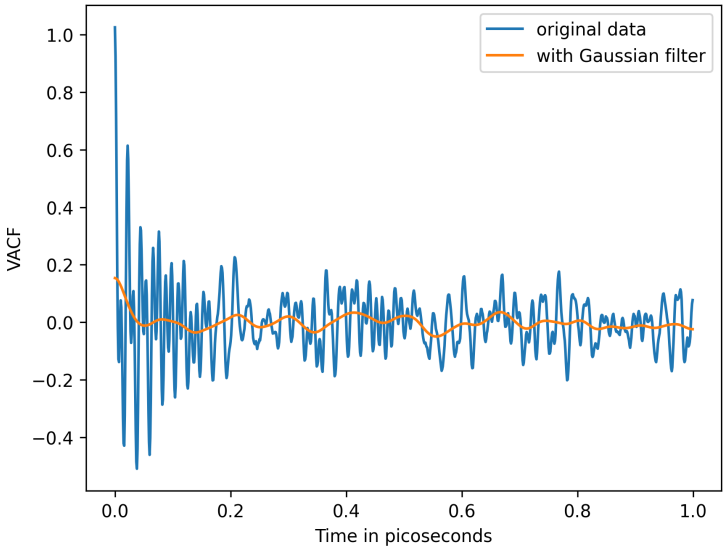

Figure A10: Velocity autocorrelation function (VACF) for the first picosecond of an MD simulation of decaalanine at 300 K. In addition to the original data, we also provide a smoothed out version using a Gaussian filter with a standard deviation of 15.