Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields

Adil Kabylda[a], **J. Thorben Frank**[b,c], **Sergio Suarez Dou**[a], **Almaz Khabibrakhmanov**[a], **Leonardo Medrano Sandonas**[d], **Oliver T. Unke**[e], **Stefan Chmiela**[b,c], **Klaus-Robert Müller**[b,c,e,f,g], **Alexandre Tkatchenko**[a]

[a] *Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg*

[b] *Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany*

[c] *Berlin Institute for the Foundations of Learning and Data -- BIFOLD, Germany*

[d] *Institute for Materials Science and Max Bergmann Center of Biomaterials, TU Dresden, 01062 Dresden, Germany*

[e] *Google DeepMind, Berlin, Germany*

[f] *Max Planck Institute for Informatics, Stuhlsatzenhausweg, 66123 Saarbrücken, Germany*

[g] *Department of Artificial Intelligence, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea*

1. Introduction

The longstanding goal of performing quantitative molecular dynamics simulations based solely on nuclear charges and electron numbers remains unfulfilled despite significant scientific progress. Most existing methods involve trade-offs in efficiency, accuracy, scalability, or transferability (EAST) [1]. While recent developments in machine-learned force fields (MLFFs) offer promising solutions, challenges such as limited scalability, incomplete treatment of long-range interactions, and insufficient dataset coverage persist [2].

In this work, we introduce SO3LR [3], a pretrained MLFF that integrates semi-local many-body interactions from SO3krates [4] with physically grounded short-range repulsion, electrostatics, and a universal van der Waals dispersion potential [5] (Fig. 1A). We demonstrate the applicability of SO3LR in nanosecond-long simulations of units of four major biomolecule types, polyalanine systems, bulk water, crambin protein, N-linked glycoprotein, and a lipid bilayer (Fig. 1B). SO3LR can be scaled to simulations involving up to ~200k atoms, with a latency of ~3 μs/atom/step on a single H100 GPU, thus approaching sizes and timescales relevant for realistic biomolecules.

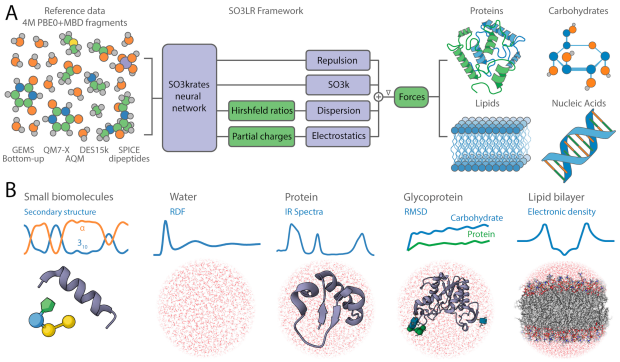

Fig. 1: Overview of the SO3LR model and simulation results. (A) SO3LR combines the SO3krates neural network with physically inspired interactions that interact directly with the neural network model. All building blocks are jointly trained on a carefully curated dataset of 4M fragments that covers a broad range of chemical space and interaction classes. (B) SO3LR enables simulations of small biomolecular units of all four major types of biomolecules, and large-scale simulations of three types.

2. Results

The dataset, architecture, and training details, as well as extended static benchmarks and simulation results are detailed in Ref. [3]. In this abstract, we focus on evaluating the performance of the SO3LR long-range modules and the folding of extended polyalanine $AcAla_{15}NMe$ in the gas phase.

2.1 Performance of long-range modules

To assess electrostatic interaction quality, we benchmarked partial charge prediction using the QM7b and AlphaML datasets, computed at the LR-CCSD/d-aug-cc-pVDZ level [6]. SO3LR predicts dipole moments with mean absolute errors (MAEs) of 0.13 D in magnitude and 5.1° in angle (Fig. 2A). AlphaML is a more challenging benchmark that spans diverse chemistries, including nucleobases, amino acids, carbohydrates, drugs, and hydrocarbons. SO3LR achieves a MAE of 0.14 D on this dataset (Fig. 2A), demonstrating accurate and transferable dipole moment predictions.

To evaluate overall noncovalent interaction energy predictions, we used the SAPT10k benchmark, computed at the SAPT2+(3)(CCD)/aug-cc-pVTZ level [7]. It consists of 70 subsets, featuring challenging dimer binding motifs dominated by electrostatics and/or dispersion interactions, offering substantial diversity across chemical space. The model achieves sub-chemical accuracy with a MAE of 0.89 kcal/mol (Fig. 2B). Rare outliers include ClO_4^-–π, NO_3^-–π, and SO_2–π complexes. This is an impressive performance overall, particularly given that part of the error arises from the PBE0+MBD reference data.

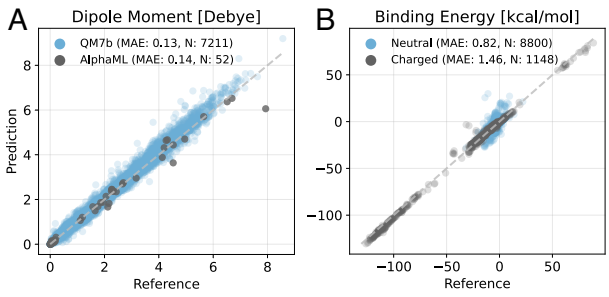

Fig. 2: (A) Evaluation of SO3LR on dipole moment prediction for 7k QM7b molecules and AlphaML showcase database [6]. (B) Performance of the model evaluated on the unseen SAPT10k dataset [7], separated into neutral and charged subsets.

Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields

Adil Kabylda[a], **J. Thorben Frank**[b,c], **Sergio Suarez Dou**[a], **Almaz Khabibrakhmanov**[a], **Leonardo Medrano Sandonas**[d], **Oliver T. Unke**[e], **Stefan Chmiela**[b,c], **Klaus-Robert Müller**[b,c,e,f,g], **Alexandre Tkatchenko**[a]

[a] *Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg*

[b] *Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany*

[c] *Berlin Institute for the Foundations of Learning and Data -- BIFOLD, Germany*

[d] *Institute for Materials Science and Max Bergmann Center of Biomaterials, TU Dresden, 01062 Dresden, Germany*

[e] *Google DeepMind, Berlin, Germany*

[f] *Max Planck Institute for Informatics, Stuhlsatzenhausweg, 66123 Saarbrücken, Germany*

[g] *Department of Artificial Intelligence, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea*

2.1 Folding of polyalanine

The folding of polyalanine presents a significant challenge due to the delicate balance of hydrogen bonding, polarization, and dispersion interactions. Previous attempts to simulate them without incorporating top-down fragments using MLFFs either failed to correctly fold $AcAla_{15}NMe$ or overstabilized the α-helix [8, 9].

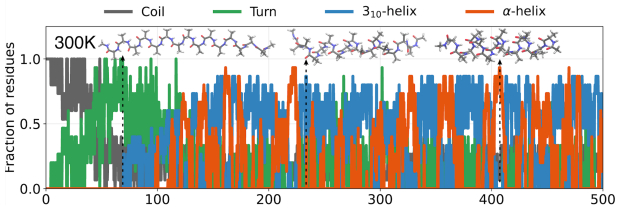

Fig. 3: Secondary structural motifs observed along a typical folding trajectory of $AcAla_{15}NMe$ at 300 K in gas phase.

We performed four 500-ps runs. The extended $AcAla_{15}NMe$ structure folded in all cases (Fig. 3). Initially, the peptide primarily consists of turns, then passes through a 'wavy' intermediate, and finally folds into a helical form with dynamic transitions between α- and 3_{10}-helices. The latter is particularly noteworthy, as empirical force fields tend to overestimate the stability of α-helices [10]. This result underscores the capability of our approach to accurately capture complex conformational dynamics driven by subtle long-range interactions.

3. Outlook

Key areas for enhancing the SO3LR model include: (i) expanding the DFT+MBD training sets to encompass a broader spectrum of (bio)chemical entities, such as ions, sugars, lipids, DNA, supramolecules, and a variety of solvents, (ii) generating higher-level coupled cluster or quantum Monte Carlo reference data for small fragments, (iii) refining long-range interaction modules to effectively account for anisotropic many-body interactions, (iv) optimizing SO3LR for multi-GPU architectures, and (v) extending simulations to treat nuclear quantum effects beyond classical Newtonian molecular dynamics. This is a non-exhaustive list of research directions, all of which are subject of ongoing efforts in the community.

Acknowledgments

The authors express their gratitude to Mirela Puleva for support with data generation during the initial stages of the project, Marcel F. Langer for helpful discussions, and Igor Poltavsky for valuable comments. The simulations were performed on the Luxembourg national supercomputer MeluXina. The authors gratefully acknowledge the LuxProvide teams for their expert support. A.K. acknowledges financial support from the Luxembourg National Research Fund (FNR AFR Ph.D. Grant 15720828). J.T.F. and S.C. acknowledge support by the German Ministry of Education and Research (BMBF) for BIFOLD (01IS18037A). K.R.M. was in part supported by the German Ministry for Education and Research (BMBF) for BIFOLD (01IS18037A) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation). A.T. acknowledges the Luxembourg National Research Fund under grant FNR-CORE MBD-in-BMD and the European Research Council under ERC-AdG grant FITMOL.

References

[1] G.F. Huang, G.F. von Rudorff, O.A. von Lilienfeld. The central role of density functional theory in the AI age. *Science* 381, 170–175 (2023).

[2] Unke et al. Machine Learning Force Fields. *Chem. Rev.* 121, 10142 (2021).

[3] Kabylda et al. Molecular simulations with a pretrained neural network and universal pairwise force fields. *ChemRxiv* 10.26434/chemrxiv-2024-bdfr0-v2 (2025).

[4] J. T. Frank, O. T. Unke, K.-R. Muller, and S. Chmiela. A Euclidean transformer for fast and stable machine learned force fields. *Nat. Commun.* 15, 6539 (2024).

[5] A. Khabibrakhmanov, D. V. Fedorov, and A. Tkatchenko. Universal pairwise interatomic van der waals potentials based on quantum drude oscillators. *J. Chem. Theory Comput.* 19, 7895 (2023).

Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields

Adil Kabylda[a], **J. Thorben Frank**[b,c], **Sergio Suarez Dou**[a], **Almaz Khabibrakhmanov**[a], **Leonardo Medrano Sandonas**[d], **Oliver T. Unke**[e], **Stefan Chmiela**[b,c], **Klaus-Robert Müller**[b,c,e,f,g], **Alexandre Tkatchenko**[a]

[a] *Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg*

[b] *Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany*

[c] *Berlin Institute for the Foundations of Learning and Data -- BIFOLD, Germany*

[d] *Institute for Materials Science and Max Bergmann Center of Biomaterials, TU Dresden, 01062 Dresden, Germany*

[e] *Google DeepMind, Berlin, Germany*

[f] *Max Planck Institute for Informatics, Stuhlsatzenhausweg, 66123 Saarbrücken, Germany*

[g] *Department of Artificial Intelligence, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea*

[6] Y. Yang, K. U. Lao, D. M. Wilkins, A. Grisafi, M. Ceriotti, and R. A. DiStasio Jr. Quantum mechanical static dipole polarizabilities in the QM7b and AlphaML showcase databases. *Sci. Data* 6, 152 (2019).

[7] C. Villot and K. U. Lao. Ab initio dispersion potentials based on physics-based functional forms with machine learning*. J. Chem. Phys.* 160, 184103 (2024).

[8] Unke et al. Biomolecular dynamics with machine-learned quantum mechanical force fields trained on diverse chemical fragments. *Sci. Adv.* 10, eadn4397 (2024).

[9] Kovacs et al. MACE-OFF23: Transferable machine learning force fields for organic molecules. *arXiv* 10.48550/arXiv.2312.15211 (2023).

[10] K. A. Bolin and G. L. Millhauser. α and 3_{10}: the split personality of polypeptide helices. *Acc. Chem. Res.* 32, 1027 (1999).