Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2025, **16**, 3738

All publication charges for this article have been paid for by the Royal Society of Chemistry

Crash testing machine learning force fields for molecules, materials, and interfaces: molecular dynamics in the TEA challenge 2023†

Igor Poltavsky, *[a] Mirela Puleva,‡[ab] Anton Charkin-Gorbulin,‡[ac] Grégory Fonseca,‡[a] Ilyes Batatia,[d] Nicholas J. Browning,[e] Stefan Chmiela,[fg] Mengnan Cui,[h] J. Thorben Frank, [fg] Stefan Heinen,[i] Bing Huang,[j] Silvan Käser, [k] Adil Kabylda, [a] Danish Khan,[il] Carolin Müller, [m] Alastair J. A. Price,[no] Kai Riedmiller, [q] Kai Töpfer, [k] Tsz Wai Ko,[p] Markus Meuwly, [k] Matthias Rupp, [r] Gábor Csányi, [d] O. Anatole von Lilienfeld, [fginost] Johannes T. Margraf,[u] Klaus-Robert Müller [fguvwx] and Alexandre Tkatchenko [*ab]

We present the second part of the rigorous evaluation of modern machine learning force fields (MLFFs) within the TEA Challenge 2023. This study provides an in-depth analysis of the performance of MACE, SO3krates, sGDML, SOAP/GAP, and FCHL19* in modeling molecules, molecule-surface interfaces, and periodic materials. We compare observables obtained from molecular dynamics (MD) simulations using different MLFFs under identical conditions. Where applicable, density-functional theory (DFT) or experiment serves as a reference to reliably assess the performance of the ML models. In the absence of DFT benchmarks, we conduct a comparative analysis based on results from various MLFF architectures. Our findings indicate that, at the current stage of MLFF development, the choice of ML model is in the hands of the practitioner. When a problem falls within the scope of a given MLFF architecture, the resulting simulations exhibit weak dependency on the specific architecture used. Instead, emphasis should be placed on developing complete, reliable, and representative training datasets. Nonetheless, long-range noncovalent interactions remain challenging for all MLFF models, necessitating special caution in simulations of physical systems where such interactions are prominent, such as molecule-surface interfaces. The findings presented here reflect the state of MLFF models as of October 2023.

Received 26th September 2024
Accepted 25th December 2024

DOI: 10.1039/d4sc06530a

rsc.li/chemical-science

[a]*Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg, Luxembourg. E-mail: alexandre.tkatchenko@uni.lu; igor.poltavskyi@uni.lu*

[b]*Institute for Advanced Studies, University of Luxembourg, Campus Belval, L-4365 Esch-sur-Alzette, Luxembourg*

[c]*Laboratory for Chemistry of Novel Materials, University of Mons, B-7000 Mons, Belgium*

[d]*Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK*

[e]*Swiss National Supercomputing Centre (CSCS), 6900 Lugano, Switzerland*

[f]*Machine Learning Group, Technical University Berlin, Berlin, Germany*

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

[h]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany*

[i]*Vector Institute for Artificial Intelligence, Toronto, ON, M5S 1M1, Canada*

[j]*Wuhan University, Department of Chemistry and Molecular Sciences, 430072 Wuhan, China*

[k]*Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland*

[l]*Chemical Physics Theory Group, Department of Chemistry, University of Toronto, St. George Campus, Toronto, ON, Canada*

[m]*Friedrich-Alexander-Universität Erlangen-Nürnberg, Computer-Chemistry-Center, Nägelsbachstraße 25, 91052 Erlangen, Germany*

[n]*Department of Chemistry, University of Toronto, St. George campus, Toronto, ON, Canada*

[o]*Acceleration Consortium, University of Toronto, 80 St George St, Toronto, ON M5S 3H6, Canada*

[p]*Department of NanoEngineering, University of California San Diego, 9500 Gilman Dr, Mail Code 0448, La Jolla, CA 92093-0448, USA*

[q]*Heidelberg Institute for Theoretical Studies, Heidelberg, Germany*

[r]*Luxembourg Institute of Science and Technology (LIST), L-4362 Esch-sur-Alzette, Luxembourg*

[s]*Department of Materials Science and Engineering, University of Toronto, St. George campus, Toronto, ON, Canada*

[t]*Department of Physics, University of Toronto, St. George campus, Toronto, ON, Canada*

[u]*University of Bayreuth, Bavarian Center for Battery Technology (BayBatt), Bayreuth, Germany*

[v]*Department of Artificial Intelligence, Korea University, Seoul, South Korea*

[w]*Max Planck Institut für Informatik, Saarbrücken, Germany*

[x]*Google DeepMind, Berlin, Germany*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4sc06530a

‡ These authors contributed equally to the work.

1 Introduction

The practical application of machine learning force fields (MLFF) aims to enhance the capabilities of computational chemistry reference methods, enabling dynamical simulations that would otherwise be unfeasible. Achieving this goal requires a high degree of trust in simulation results, allowing MLFF models to become standard tools in research and industry pipelines. While architecture development is greatly facilitated by easy "pointwise" testing of models on standardised train/test splits of toy problems, and this approach has been standard in the machine learning (ML) community, creating models that are actually useful for materials and molecular science research requires complicated system-specific evaluation. Even the earliest successful models in the materials field that were targeting specific systems already did this, demonstrating high accuracy in computing observables such as phonon spectra, phase transitions, defect formation energies, *etc.*, as well as pointwise accuracy of reproducing the potential-energy surfaces and atomic forces.[1–9] Later, as the computer science community engaged with the problem of molecular modeling, their practices also came to be prominent and, especially in works that compared different ML architectures, developers assessed the accuracy of models mostly by evaluating errors in energies and forces relative to the ground truth they were targeting.[10–29] There is a widely held view that we need to return to assessing observables.[30] In the meantime, sophisticated MLFF accuracy measures[31–35] and visualization tools[29,36] have been developed to address ML models' performance on local and global measures. It was suggested in particular that long molecular dynamics (MD) simulations[19,37–49] provide a robust test of MLFF reliability as predictors of physical behavior when mean absolute error (MAE) or root mean square error (RMSE) may be insufficient or even misleading when considered on their own.[50–57]

In this study, we evaluate the quality of modern MLFF architectures by comparing the outcomes of MD simulations performed with MACE,[13,14] SO3krates,[22,23] sGDML,[17,18,38,58] SOAP/GAP,[2,6] and FCHL19*[11,12,59,60] models. MACE and SO3krates are equivariant message-passing graph neural networks (NNs), representing many-body information about the geometric atomic configuration employing spherical harmonics and radial distributions function learned through multilayer perceptrons. SO3krates also relies on an equivariant attention mechanism to enhance the model's efficiency. FCHL19*, sGDML, and SOAP/GAP are kernel-based ML architectures. FCHL19* and SOAP/GAP are based on local atom-centered representations, while sGDML employs a global descriptor. Full details of the MLFFs are available in the ESI of ref. 61. We would like to highlight that the MLFF architectures included in the TEA Challenge 2023 were limited to those whose developers could participate in the benchmark. This approach was adopted to minimize the risk of misinterpretation or drawing misleading conclusions due to potential mistraining. The complete list of modern MLFFs is considerably broader. Prominent MLFFs that did not participate include ANI,[62] Alegro,[24] ACE,[63] ALIGNN-FF,[64] AIMNet2,[65] DeepMD,[66,67] Elemental-SDNNFF,[68] FIREANN,[69] FLARE,[70] GMBNN,[71] GPTFF,[72] MTP,[73] NequIP,[25] PIP,[74,75] SevenNet,[76] SNAP,[77] and M3GNet,[78] among others. To ensure that the current benchmark can be extended to include any additional MLFF model in the future, all the data and scripts necessary to train other ML architectures and replicate the simulations and analyses are available online and upon request.

Given the computational expense of explicit electronic structure methods, often the only plausible test for the performance of an ML model (apart from comparing to experimental observables, which brings its own complexities) is another model trained and used under identical conditions. Achieving consistency in results across different ML architectures would mark a significant milestone in the development of MLFFs. This work builds upon our previous manuscript, "Crash Testing Machine Learning Force Fields for Molecules, Materials, and Interfaces: Model Analysis in the TEA Challenge 2023," which provided a detailed analysis of force and energy predictions on test datasets.[61] In contrast, this study focuses on analyzing various observables derived from MD simulations. Only trajectories that provide sufficient statistical data under the specified simulation conditions are considered. Each challenge is independently analyzed to assess the ability of different MLFF architectures to simulate specific types of systems. We focus on evaluating the capability of MLFFs to perform classical MD simulations under ambient and near-ambient conditions. While these simulations cover a broad range of potential MLFF applications, they do not address all possible scenarios. For example, simulations that capture nuclear quantum effects, such as imaginary-time path-integral MD,[79] or those involving chemical bond breaking, fall outside the scope of this study. Tackling such challenges would require additional components, including advanced sampling techniques to capture low-probability or classically forbidden system geometries, as well as costly multi-reference *ab initio* calculations to accurately describe dissociation processes. Each of these represents an open challenge in its own right.

This article is divided into four sections as follows. In Section 2, we present a comprehensive analysis of the classical MD simulations conducted for each system studied in the TEA Challenge 2023. These systems include two biomolecular systems (alanine tetrapeptide and *N*-acetylphenylalanyl-pentaalanyl-lysine), a 1,8-naphthyridine/graphene interface, and a methylammonium lead iodide perovskite. The starting point of each of the 12 MD trajectories simulated in each experiment and the corresponding MLFF models can be found in the Zenodo archive **https://doi.org/10.5281/zenodo.13832724** with trajectories available upon request. From their analyses, we extract key insights into the applicability and reliability of MLFFs, and identify opportunities for further development and improvement. Section 3 contains guidelines for the development, training, and application of MLFFs. Section 4 represents the overall conclusions.

2 Results and discussion

2.1 Alanine tetrapeptide

We begin our analysis of the MD trajectories provided by different MLFF models with the simplest case among the TEA

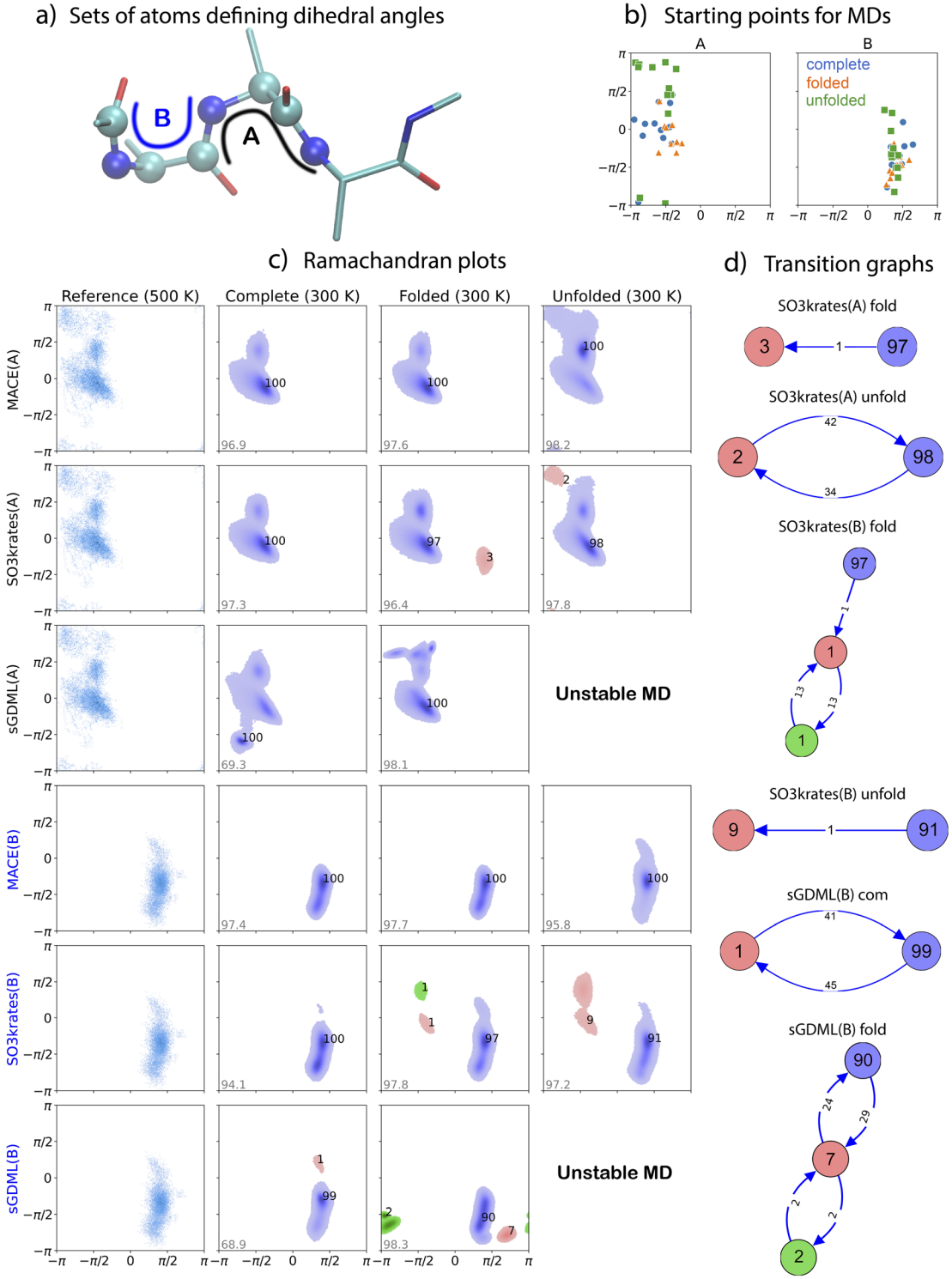

Fig. 1 Challenge I, Alanine tetrapeptide: analysis and Ramachandran plots for MD simulations at 300 K. (a) Diagram of the peptide depicts the peptide with atom sets A and B forming two consecutive pairs of dihedral angles (left four atoms – *x*-axis, right four atoms – *y*-axis). (b) Initial points for 12 MD trajectories. (c) Ramachandran plots for the reference systems at 500 K and for MD simulations at 300 K using MACE, SO3krates, and sGDML MLFFs trained on com (complete), fold (folded), and unfold (unfolded) datasets. The numbers near the clusters indicate their relative population (in percent), while the grey number in the lower left corner of each plot shows the percentage of configurations from the MD trajectories identified as belonging to one of the clusters. (d) Graphical representation of the transitions between different (meta)stable domains. The values on the arrows show the number of transitions identified in the dynamics.

2023 challenges: the alanine tetrapeptide molecule in Challenge I. For a decade, organic molecules with a few dozen atoms in the gas phase have been routinely treated with different MLFF architectures. Additionally, such molecules are well within the capabilities of DFT codes, which can routinely compute hundreds of thousands of geometries without excessive computational effort. In this particular case, the reference dataset is comprised of 85 109 molecular geometries representing an *ab initio* NVT MD trajectory generated at 500 K (taken from MD22 benchmark,[58] details available in Section 2 of ref. 61). This comprehensive dataset facilitated the generation of three distinct training sets, categorized by the distance between the farthest non-hydrogen atoms, which served as a measure of molecular compactness: (1) the complete set, encompassing representative samples of both extended and compact Ac-Ala3-NHMe structures; (2) the folded set, consisting of the 70% most compact structures extracted from the MD trajectory; and (3) the unfolded set, comprising the remaining 30% of less compact structures from the same trajectory. Details of the training procedures for different MLFFs for this and the other Challenges are available in ESI of ref. 61.

Analysis of the MD trajectories and a comparison between the outputs for the different MLFFs are presented here *via* Ramachandran plots, as shown in Fig. 1. The Ramachandran plots are useful for visualizing the allowed conformational space of the peptide backbone.[80] For the Ac-Ala3-NHMe tetrapeptide, two selected pairs of dihedral angles in Fig. 1a, A and B, are conventionally referred to as ϕ_2/ψ_2 and ϕ_1/ψ_1, respectively.[80–82] The initial points for MD simulations are depicted in Fig. 1b.

In order to obtain an informative picture of the results, analysis involving a clustering algorithm was carried out to identify the high density regions of population during the MD and filter out noise and unrepresentative low density areas.[83,84] The full step-by-step algorithm description is available in the ESI.† Our analysis algorithm identifies different (meta)stable domains in the Ramachandran plots, illustrated in various colors, Fig. 1c. The transitions between these (meta)stable domains obtained with SO3krates folded/unfolded and sGDML complete/folded models are represented in the graph form in Fig. 1d. It is important to note that the benchmark reference trajectory was obtained at 500 K to allow for more extensive sampling of the conformational space and bond lengths. To ensure statistically significant results, only the Ramachandran plots for ML models that produced stable 1 ns dynamics are presented. Consequently, the analyses are based on MLFF MD trajectories obtained at 300 K, as most MLFF models failed to produce stable 1 ns dynamics at 500 K. Nevertheless, the Ramachandran plots for the reference *ab inito* MD at 500 K still provide a qualitative guideline for the 300 K MLFF MD results.

The MACE models trained on both complete and folded datasets exhibit excellent mutual agreement. However, they undersample the upper left corner of the Ramachandran plot for A and the upper part of the dihedral cluster for B compared to the reference data. Both of these areas correspond to the highly unfolded conformations of the tetrapeptide. Training the MACE model on the unfolded dataset results in better qualitative agreement with the reference MD trajectory. Quantitative agreement estimation is challenging due to the short length of the *ab initio* MD trajectory and the difference in temperatures.

The SO3krates models display distinct Ramachandran profiles depending on whether they are trained on complete, folded, or unfolded datasets. Firstly, when trained on both complete and folded datasets, SO3krates model also undersample regions of highly unfolded conformations, similar to MACE. This indicates that this might be due to lower simulation temperature compared to the reference. Notably, the SO3krates model trained on the folded dataset exhibits additional metastable states with low populations (3% for A and 1% for B) and low transition probabilities. For dihedral B, the SO3krates model trained on the unfolded dataset identifies an extra metastable region with a 9% cluster population, though the transition probability into this state is low, with only one transition observed in 12 ns of total dynamics (12×1 ns). The seeds of this cluster also appear as two small clusters with a 1% population and relatively high mutual transition rates in the SO3krates model trained on the folded dataset. These molecular geometries were not observed in MD simulations using the SO3krates model trained on the complete dataset or by any MACE or sGDML models. However, exploration to these regions have been studied previously in the original SO3krates article.[23] It is worth noting that the two small clusters in the *SO3 fold* plot, dihedrals B, and the two clusters in the *SO3 unfold* plot, dihedrals A, can be merged due to their relatively high transition rates. The clustering algorithm employed here uses a predefined fixed number of transitions to merge clusters chosen to suit the data in general. It does not account for their population, providing suboptimal results when at least one of the clusters is small.

The sGDML model trained on a complete dataset also demonstrates acceptable results. The total number of transitions between the two clusters identified for B is 86, slightly below the manually selected threshold of 100 for merging the clusters into one. However, *removing parts of the reference geometries in folded or unfolded datasets leads to significant differences in the Ramachandran profiles or even MD instability*. This sensitivity is attributed to the sGDML model's interpolation in the space of a global system descriptor of inverse distances, making the model highly dependent on the quality and completeness of the training dataset compared to MLFFs employing local descriptors.

In ESI, a comprehensive Table SI 1† lists the chemical bonds responsible for the instability of all MLFF architectures trained on different datasets at 300, 500, and 700 K. Most broken bonds involve carbon atoms connected to other elements. Additionally, there are notable differences in bond-breaking patterns between kernel-based and equivariant NN-based ML models. For sGDML, SOAP/GAP, and FCHL19*, the specific bond causing molecular instability was readily identifiable. By contrast, for SO3krates, bond breaking exhibited an explosion-like behavior, with a large part of the molecule decomposing into atoms within a few dozen steps, making it challenging to pinpoint the exact bond responsible for the instability.

In summary, when trained on a complete dataset, we observed strong mutual agreement in the molecular dynamics

of the alanine tetrapeptide generated by MACE and SO3krates models. Discrepancies arise when the training datasets lack either folded or unfolded geometries. Both models reasonably explored the unfolded potential energy surface (PES) basin when trained on the unfolded dataset. The MACE folded model is consistent with the complete one. The SO3krates folded and unfolded models were mostly consistent, while the sGDML MLFFs demonstrated increased sensitivity to variations in the training data.

2.2 *N*-Acetylphenylalanyl-pentaalanyl-lysine

The second challenge in the TEA Challenge 2023 was performed on a larger organic system, namely the protonated Ac-Phe-Ala5-Lys peptide. The dataset, comprising 100 000 reference geometries with energies and forces calculated using the PBE0 (ref. 85) exchange-correlation functional and the nonlocal MBD (MBD-NL)[86] method for modeling dispersion interactions, was specifically generated for the TEA Challenge 2023. Full details can be found in Section 2 of ref. 61. The main aim of the

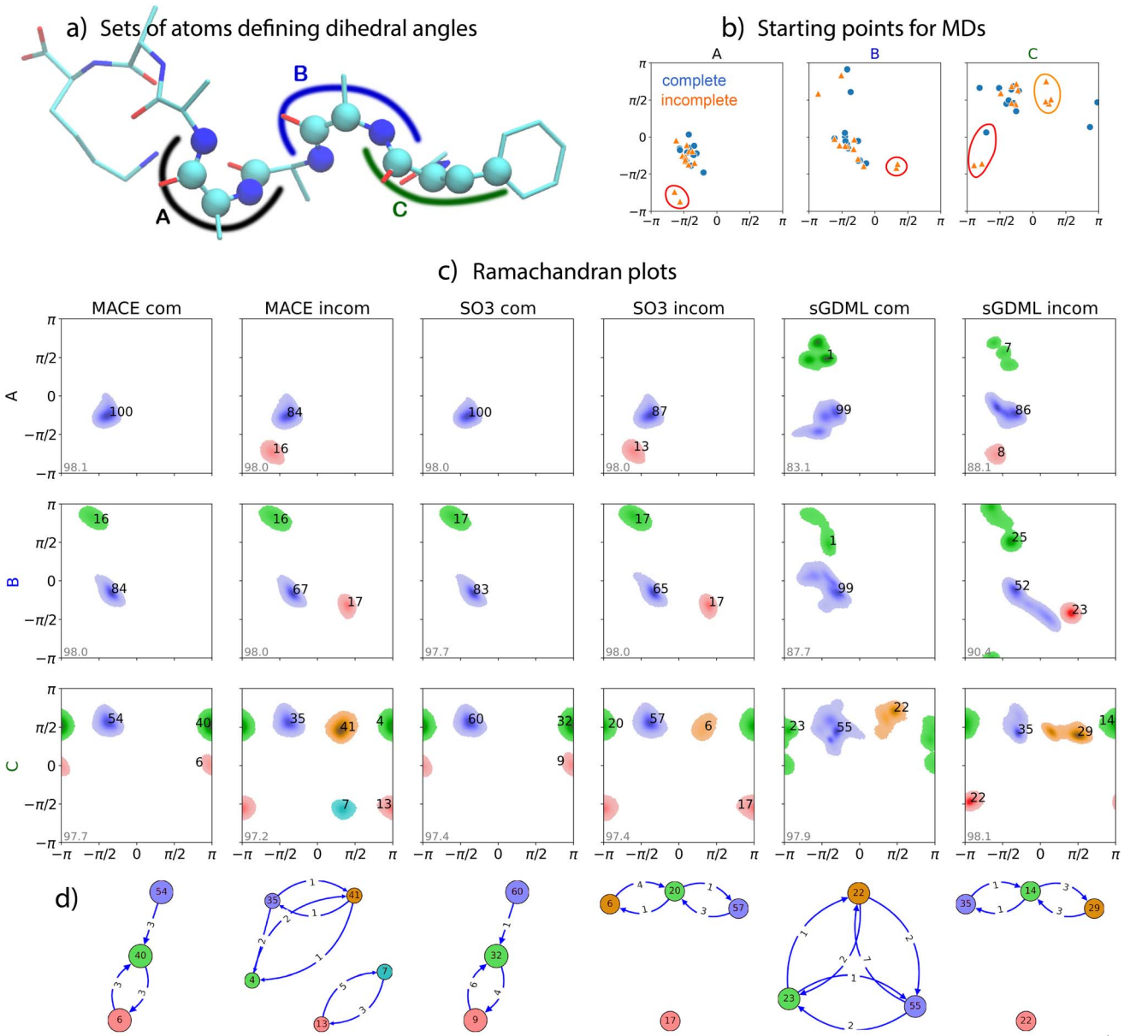

Fig. 2 Challenge II, *N*-acetylphenylalanyl-pentaalanyl-lysine peptide: Analysis and Ramachandran plots for MD simulations at 300 K. (a) Diagram of the peptide with sets of atoms A, B, and C forming three consecutive pairs of dihedral angles (right four atoms – *x*-axis, left four atoms – *y*-axis). (b) Initial points for 12 MD trajectories. (c) Ramachandran plots for MD simulations employed different MLFF models: MACE com (complete), MACE incom (incomplete), SO3 com, SO3 incom, sGDML com, and sGDML incom. The numbers near the clusters indicate their relative population (in percent). The grey number in the bottom left corner of each plot shows the relative number of configurations (in percent) from the MD trajectories identified as belonging to one of the clusters. (d) Graphical representation of transitions between different (meta)stable domains for dihedrals C. The values on the arrows show the number of transitions identified in the dynamics.

challenge was to assess the ability of MLFFs to handle different types of reference data incompleteness. To this end, two training datasets were created for the Ac-Phe-Ala5-Lys peptide system based on MD simulations around the 200 lowest energy conformers. The "complete" dataset consisted of 20 randomly selected molecular configurations from each of these trajectories, while the "incomplete" dataset contained 32 randomly selected configurations from only 125 trajectories. The clustering algorithm used here is the same one as for Challenge I (full details in ESI†). Fig. 2 presents the Ramachandran plots for three selected pairs of neighboring dihedral angles along the peptide. The dihedral angle pairs, labeled as A, B, and C (see Fig. 2a), represent key cases for comparing different MLFFs. The initial points for MD simulations are shown in Fig. 2b. Only the Ramachandran plots for ML models that produced stable 1 ns long dynamics are displayed. The FCHL19* and SOAP/GAP MLFFs did not achieve the required 1 ns to perform statistically-significant analysis. Our analysis identifies distinct (meta)stable domains in the Ramachandran plots, represented in different colors, Fig. 2c. Since the original dataset was not derived from an MD trajectory, reference density plots are unavailable. Transition graphs illustrate the transitions between (meta)stable domains for dihedrals C across different ML models, Fig. 2d.

We begin by comparing the MD results of the MACE and SO3krates models, each trained on complete (com) and incomplete (incom) datasets, Fig. 2c. The most noticeable difference is the increased number of clusters in the dynamics generated by the models trained on the incomplete dataset, primarily reflecting variations in the starting points for the MD simulations. Additional red clusters appear for the A and B dihedrals due to distinct starting points in the incomplete trajectories (highlighted by red circles in Fig. 2b). These starting points significantly differ from other initial configurations, leading to divergences in the Ramachandran plots. Both models, however, accurately reproduce the close-to-equilibrium regions of the PES for A and B, providing similar positions of the minima and shapes of the probability distributions around them (within 3% agreement in population distributions between clusters).

For the C dihedrals, a block of four starting points (highlighted by an orange circle in Fig. 2b) appears like an orange cluster. Additionally, a shift in the red cluster position from 0 to $-\pi/2$ along the *Y*-axis is attributed to differences in starting configurations between the complete and incomplete datasets (emphasized by a red circle in Fig. 2b). Differences in the MLFFs from the MACE and SO3krates architectures significantly influence the distribution of the C dihedrals, as supported by transition graphs in Fig. 2d, differing especially in the case of training on an incomplete dataset. The MACE MLFF shows a 54% relative population for the complete C dihedral case for the largest cluster $(-\pi/2, \pi/2)$ compared to 60% with the SO3krates MLFF. This difference stems from a higher transition probability within the MACE PES from this state to the green cluster $(-\pi, \pi/2)$. Conversely, the green cluster population is higher with the MACE PES at 40%, compared to 32% for the SO3krates PES, due to a nearly twice higher transition probability between the green and red $(\pi, 0)$ metastable states with the SO3krates MLFF trained on the complete dataset.

These results suggest that *the primary difference between MACE and SO3krates MLFFs lies in the description of out-of-equilibrium regions of the PES*, which are responsible for rare transitions or large geometry fluctuations. This is also underpinned by the PES analyses provided in the first part of the manuscript, see Fig. 5, of ref. 61 or the MD results of Challenge III presented later in this article. Further comparison of the MACE and SO3krates models trained on the incomplete dataset for the C dihedrals supports this. Noticeably different transition patterns emerge, as well as differing cluster populations and even the appearance of an extra metastable state (blue cluster) within the PES reconstructed with the MACE MLFF (incomplete dataset). It is also likely that such differences would emerge between different MLFFs trained on the same data with the same architecture, just using a different set of initial weights. We want to state, though, that the statistically converged analyses of the transition patterns would require more extended MD simulations or the employment of enhanced sampling techniques which is beyond the scope of the current work. Simultaneously, the MACE and SO3krates architectures, when trained on the complete dataset, demonstrate remarkable mutual agreement in modeling the dynamics of the Ac-Phe-Ala5-Lys peptide. This consistency underscores the capability of modern MLFFs based on equivariant NN with different architectures to handle relatively large and complex organic molecules and model their dynamics and transitions.

For the sGDML architecture, we observe only qualitative agreement with the results from both NNs in Challenge II. The shape and population of clusters differ significantly due to the global nature of the sGDML model, which requires a reliable and comprehensive set of representative query configurations for effective interpolation between system states. Achieving such a representative set is challenging for large molecules undergoing complex structural transformations, especially with a relatively small training dataset of 4000 configurations. Consequently, the *sGDML model operates in a low-data regime, leading to a notable depreciation in performance*.

In ESI, Table SI 2† details the broken chemical bonds responsible for the instability of MD simulations across all MLFF architectures. We observe behavior similar to that reported for Challenge I. Notably, the kernel-based MLFFs utilizing local descriptors, namely SOAP/GAP and FCHL19*, could not sustain stable dynamics over a 1 ns duration for both peptides evaluated in this study. This finding indicates that MD simulations employing global sGDML models or NN architectures, incorporating nonlocality through message-passing elements, exhibit greater stability than MLFFs based solely on local ML models.

In summary, our observations indicate that the results for MACE, SO3krates, and sGDML MLFFs in this test were consistent with those from Challenge I. Overall, we can conclude that equivariant NN MLFFs can reliably reproduce the dynamics of organic molecules (at least up to 100 atoms). The main discrepancies between MD results occur when the training datasets lack representative reference geometries. This suggests

that active learning or similar iterative approaches[87–89] should become an integral part of MLFF training procedures to ensure the completeness and representativeness of training datasets.

2.3 1,8-Naphthyridine molecule on graphene

For modern MLFFs, dealing with either a 1,8-Naphthyridine molecule or a pristine graphene sheet separately is easy. However, combining the two systems introduces additional complexity, namely the molecule–substrate interaction responsible for multiple key properties of the system. MLFFs have to learn this interaction simultaneously with the much stronger covalent bonds in 1,8-naphthyridine and graphene. In our analyses, the focus is on four measures that capture the molecule–substrate interaction. Firstly, the distribution of the *distance D* between the graphene sheet and the molecule, defined as a distance between the averaged graphene plane (AGP) and the center of mass of the molecule, is considered. Secondly, the *incline* angle between the molecule and the surface, defined as the angle between the averaged plane of the molecule (H atoms excluded) and the AGP is analysed. Lastly, the *tilt* and *slant* angles are investigated. See plots of all the interaction's aspects in Fig. 3. Note that the molecule is symmetrical for the tilt rotation (the tilt angle is defined in the range $[0, \pi/2]$), while it is asymmetric for the slant rotation (the slant angle is defined in the range $[-\pi/2, \pi/2]$). Fig. 3 illustrates the results for MLFF MD simulations conducted at 300 and 500 K. Notably, all ML models encounter angles and distances that extend significantly beyond those observed in shorter reference MD simulations at 500 K, indicating that the MLFF models operate in an extrapolation regime.

At 300 K, sGDML, SO3krates, and MACE produced similar patterns for all four examined dependencies. This includes subtle features found in agreement between models observed in the plot of incline angle probability as a function of surface-to-molecule distance. A notable split for angles close to $\pi/2$ (molecular orientation perpendicular to the graphene surface) is visible for sGDML, SO3krates, and MACE. The split arises from the asymmetry in the molecule's slant rotation, where the side containing N atoms can approach the surface closer than the sides with H atoms attached to C atoms. The molecular asymmetry also explains the trend to negative slant values in the distance *versus* slant angle plots. Furthermore, it is the reason for the asymmetric large fluctuations of the slant angle in the tilt *versus* slant angle plots – a negative slant angle indicates that the side of the molecule with N atoms is closer to the surface. The main disagreement between the models is a tendency for the molecule to desorb from the surface when SO3krates models is employed. This tendency is absent in the MD simulations produced with the MACE and sGDML MLFFs.

At 500 K, the predictions between the MLFFs begin to diverge. The MD simulations obtained with MACE and SO3krates models differ drastically from those produced by sGDML. Both the MACE and SO3krates models predict the molecule's desorption from the surface, while the sGDML model keeps the molecule within an 8 Å distance. This suggests that the PES profile for the molecule in the direction perpendicular to the graphene surface predicted by sGDML differs from that predicted by the NNs. To verify this conjecture, the molecule-surface interaction energy is computed as a function of the distance, see Fig. 3b. The computation starts from the relaxed structure (obtained using the reference DFT setup), and confirmations are produced, for which the z-coordinates of the atoms belonging to the molecule are moved further from or closer to the surface in 0.1 Å steps. This provides us with the $\Delta U(D)$ dependence computed at the reference level of accuracy. The same calculations are repeated using the MACE, SO3krates, sGDML, SOAP/GAP, and FCHL19* models. Notably, some MLFFs failed to provide reasonable energy predictions at large molecule-to-surface distances. Therefore, the minimum energy within the distance range of 3 to 4 Å is chosen for each method as the zero energy level. The thin black line in Fig. 3b corresponds to the DFT energy at an infinite molecule-to-surface separation.

In complete agreement with the MD simulations, the *SO3krates and MACE models significantly underestimate the potential energy minima compared to the DFT reference*. Fig. 3b shows the difference between the DFT and MLFFs energy predictions at an 8 Å distance as small numbers on the top. Notably, this distance is beyond the cutoff for both NNs, leading them to predict zero interaction between the molecule and the surface. The MACE model underestimates the molecule-surface interaction energy at 8 Å by 2.9 kcal mol^{-1} and SO3krates by 4.4 kcal mol^{-1}. For the adsorption minimum, when aligning the DFT and MLFFs energies at infinite distance, the MACE and SO3krates models underestimate the value by 3.5 and 4.9 kcal mol^{-1}, respectively.

Interestingly, *this underestimation* of the molecule-surface interaction minimum *aligns with the high accuracy in predicting the force acting on the molecule as a whole in the test set*. The reference molecule-to-surface distances are limited to an interval of 3 to 5 Å. Within this range, the curvature of the $\Delta U(D)$ function obtained from DFT calculations and those predicted by MACE and SO3krates models agree. Moreover, these distances fall within the cutoff radii for MACE and SO3krates, which are 6 Å and 5 Å, respectively. This results in minor MAEs in forces, 0.12 kcal (mol^{-1} $Å^{-1}$) for MACE and 0.58 kcal (mol^{-1} $Å^{-1}$) for SO3krates, despite a significant underestimation of the barrier.

Fig. 3c and d illustrate the differences of free energies of adsorption calculated employing different MLFFs. The free energies derived from MD trajectories *via* the thermodynamic integration method,[90] are significantly smaller than the adsorption energy estimations calculated as the difference between the energy of the DFT-optimized structure and the state, for which the molecule and the surface are at infinite separation. This discrepancy arises due to the substantial rotational freedom of the molecule's plane relative to the surface plane. Specifically, the free energy minima predicted by the MACE and SO3krates models at 500 K are only 5.8 and 4.4 kcal mol^{-1}, respectively, whereas the adsorption energies predicted by these models are 11.3 and 9.8 kcal mol^{-1}.

The kernel-based models also face significant challenges in reproducing the molecule-surface interaction. The sGDML model over-stabilizes the system due to an artificial barrier at

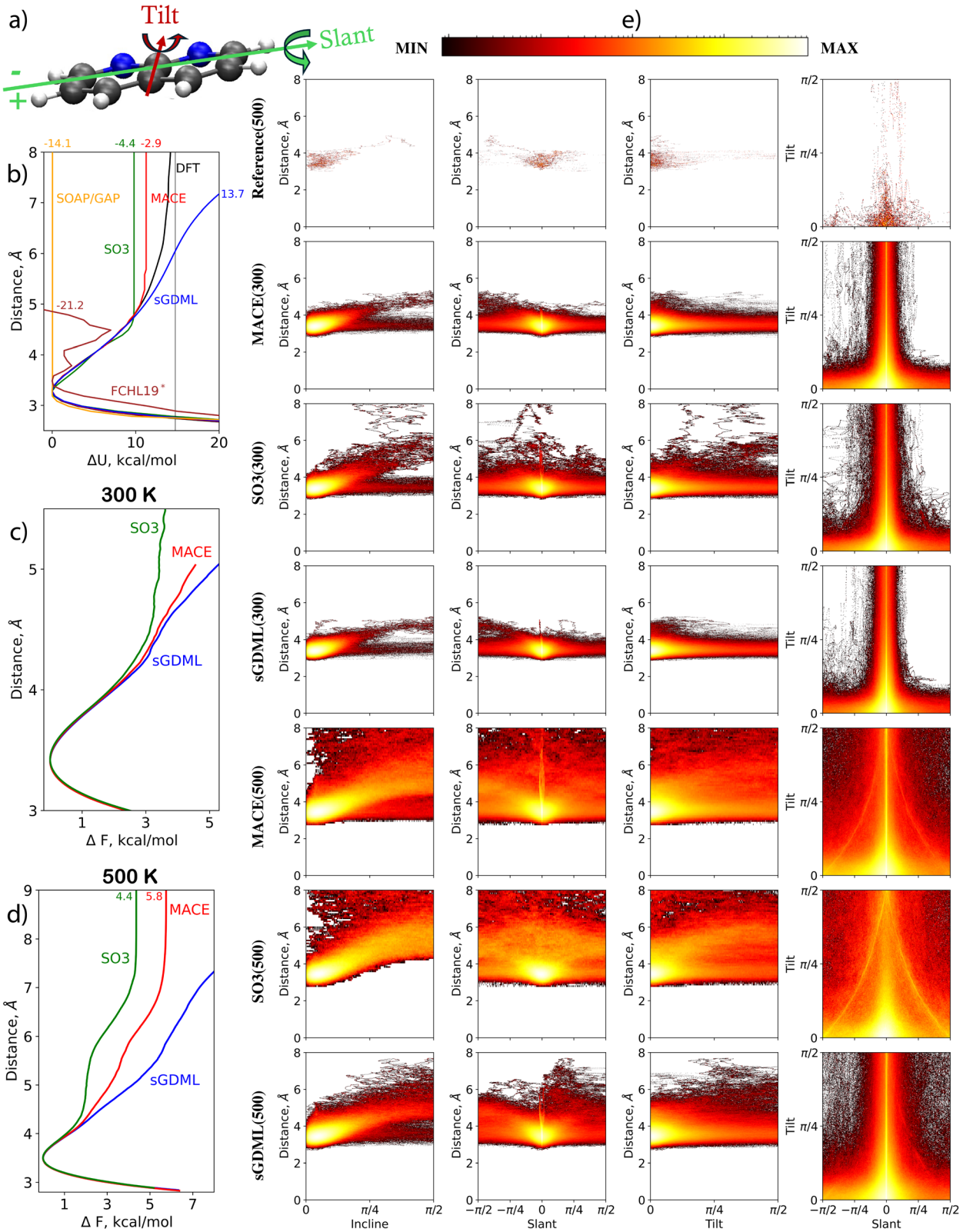

Fig. 3 Analysis of MDs for 1,8-Naphthyridine molecule adsorbed on graphene. (a) Schematic representation of the molecule and definitions of tilt and slant angles. (b) Interaction energy profile (ΔU) as a function of the distance between the centers of mass of the molecule and graphene. (c) and (d) Free energy change profile from MD simulations at 300 and 500 K, respectively. (e) Distributions of incline, tilt, and slant angles at different molecule-to-surface distances, and the tilt-slant angle distribution in MD simulations at 300 and 500 K using MACE, SO3krates, and sGDML models compared to the reference dataset.

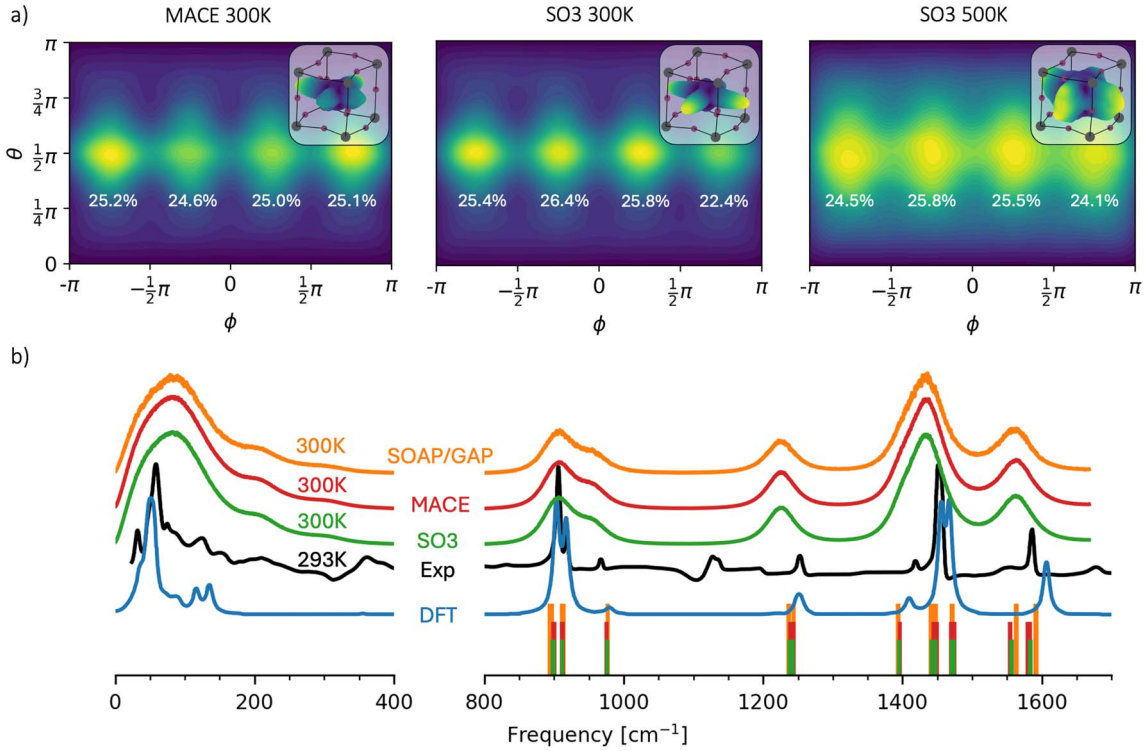

Fig. 4 Observables extracted from the MD simulations of the orthorhombic phase of $MAPbI_3$. (a) Distribution of C–N bond orientation in spherical coordinates. The numbers near the maxima indicate the participation ratios of the bond orientation for four different regions of ϕ. 3D spatial representations of the distribution within the PbI framework on a logarithmic scale are shown in the upper right corners. (b) Spectral analysis of the velocity autocorrelation function (solid lines) obtained from MD simulations using SO3krates (green), MACE (red), and SOAP/GAP (orange) models at 300 K, compared to the experimental infrared spectrum[93] (black) and DFT within the harmonic approximation[93] (blue). The vertical bars indicate phonon frequencies calculated within the harmonic approximation for each MLFF model, spanning a frequency range of 800 cm^{-1} to 1800 cm^{-1}.

intermediate distances. For instance, at 8 Å, the model overestimates the DFT energy by 13.6 kcal mol^{-1}. The artificial barrier only disappears at much larger distances, above 30 Å. The SOAP/GAP model fails to reproduce any molecule-to-surface attraction, providing a flat interaction profile immediately after the expected physisorption minima around 3.2 Å. The behavior of $\Delta U(D)$ within the FCHL19* MLFF lies somewhere between SOAP/GAP and the NNs. Qualitatively, the model reproduces the repulsion and a part of the attraction regions but with additional peculiar minima and the largest error at 8 Å distance of 21.2 kcal mol^{-1} among all MLFFs participating in the TEA Challenge 2023.

A significant point is that adding more training points with a better sampling of molecule-to-surface distances can only help the global sGDML MLFF to improve the description of the molecule-surface interaction. The finite cutoffs intrinsic to all other MLFFs participating in TEA 2023 would prevent them from correctly describing the molecule adsorption process or long-term dynamics unless additional elements targeting long-range interactions are incorporated into these MLFFs. We would like to emphasize that the above statement pertains specifically to the versions of the MLFF models that participated in the TEA Challenge 2023. The updated version of SO3krates (the pre-trained SO3LR model[91] for bio-molecular simulations includes long-range electrostatic and universal dispersion interactions) includes the option to incorporate long-range interactions. With the appropriate reference data, SO3LR could accurately describe the adsorption process.

In ESI, Table SI 3† details the broken chemical bonds for the instability of MD simulations across all MLFF architectures.

In summary, molecule–surface interfaces need careful attention when using MLFFs for simulations. Modern MLFFs achieve high accuracy, capturing detailed system behavior with MAEs and RMSEs in the fraction of one kcal mol^{-1} and kcal (mol^{-1} Å^{-1}). However, they require embedding long-range interaction models into MLFF architectures to handle system elements beyond the intrinsic cutoff distances, enhancing simulation accuracy and reliability.

2.4 Methylammonium lead iodide perovskite

The periodic structure of $MAPbI_3$ features an organic methylammonium (MA) cation at the center of the inorganic eight corner-sharing $[PbI_6]_4$-octahedral. The stability of the simulations was assessed by applying two different thresholds for the organic and inorganic parts of the structure: 2 and 4.3 Å, respectively. The organic threshold was defined using the covalent radius. In contrast, the inorganic threshold was based on experimental results from various scattering techniques for the radial distribution function of the Pb–I ionic bond.[92]

MA cations within the PbI framework are characterized by ionic interactions with the rest of the lattice. To investigate how ML models reproduce these weak interactions, we computed the distribution of the orientation of the C–N bond of MA cations. The results in spherical coordinates are shown in Fig. 4a for the MACE and SO3krates models in cases where MD simulations produced sufficiently long (over 10 ns) trajectories. The polar angle θ is the angle between the z-axis and the C–N bond, while the azimuthal angle ϕ is the angle between the orthogonal projection of the bond on the x–y plane and the x-axis. The upper right corner of the distribution plot for each set of MDs presents a 3-dimensional representation of the C–N orientation on a log scale for peak visibility. The numbers near the maxima of the distributions indicate the probability of the bond orientation for four regions: $-\pi < \phi < -\pi/2$, $-\pi/2 < \phi < 0$, $0 < \phi < \pi/2$ and $\pi/2 < \phi < \pi$. The bond tends to orient into cavities in the PbI framework along the x and y axes for all presented MDs. As expected, the distribution for SO3krates at 500 K is more smeared than that for both NNs at 300 K. The unlikely orientation of the bond along the z-axis relates to the alignment of this axis with the c-axis of the orthorhombic structure of $MAPbI_3$.

We would like to briefly address the stability issues encountered with the MACE model during MD simulations at 500 K. The MACE model permitted significant fluctuations in the positions of Pb atoms, which ultimately compromised the structural integrity of the system when these atoms approached the MA molecule. This instability can be attributed to the inadequate sampling of large atomic fluctuations in the reference dataset, which was derived from relatively short *ab initio* MD simulations.[94] Additionally, the training dataset lacked supplementary data necessary for the ML models to capture these large-scale atomic movements accurately. Notably, the escape events involving Pb atoms in our simulations occurred after significantly longer times compared to the reference dynamics. Despite this, the SO3krates model demonstrated a markedly improved stability. In our tests, SO3krates resulted in only 2 out of 12 failed 1 ns long trajectories, compared to 12 out of 12 failures with MACE. However, it is important to note that SO3krates also encountered similar issues twice, underscoring the current limitations in accurately describing interatomic interactions at distances beyond the typical covalent bond range in a sparse training data regime. A comparative analysis of atomic radial distribution functions from MACE and SO3krates MD simulations at 500 K, alongside 1 ps *ab initio* MD results at the same temperature, is provided in ESI.†

Fig. 4b presents the spectral analysis of the $MAPbI_3$ system for MACE, SO3krates, and SOAP/GAP models. This analysis yields the vibrational frequencies of the system, derived through the Fourier transform of the velocity autocorrelation function (VAF), which provides insights into vibrational dynamics. The experimental infrared (IR) spectrum and a DFT-simulated IR spectrum (within the harmonic approximation) are also shown for reference. Since IR intensities are governed by changes in dipole moments, while VAF spectra are based on velocity correlations, the resulting peaks from the VAF method are generally broader, with different intensity ratios. Nevertheless, the VAF-based frequency positions can still be compared meaningfully to the IR spectra. Unlike the harmonic approximation assumed in DFT, the VAF approach can recover anharmonic vibrational modes. For instance, a notable feature at 200 cm^{-1} is observed in both the MD-derived and experimental IR spectra but is absent from the DFT predictions. Overall, *all MLFF models produce nearly identical spectra, showing good agreement with experimental data and DFT* in terms of peak positions.

At the same time, *the experimental peaks* observed at 360 and 1140 cm^{-1} *were not detected by either MLFF*. We attribute this discrepancy to the *limitations of the reference DFT calculations*. Notably, these peaks are also absent in the DFT spectra presented in Fig. 4. However, it is important to note that the DFT functionals used in this study differ from those in ref. 93. Unfortunately, computing the IR spectra for $MAPbI_3$ with the unit cell size used for the TEA Challenge 2023 at the level of accuracy PBE + MBD-NL is not feasible.

Additionally, phonon frequencies, depicted as vertical bars on Fig. 4b, were calculated within the harmonic approximation for each model, covering the frequency range from 800 cm^{-1} to 1800 cm^{-1}. To obtain these frequencies, the $2 \times 2 \times 2$ $MAPbI_3$ system was first optimized separately by each MLFF model using the BFGS algorithm implemented in ASE, with a convergence criterion set to a maximum force component of 0.005 eV $Å^{-1}$.[95] The phonon frequencies were then derived from these optimized structures using the finite difference method for each respective model, employing a 0.01 Å displacement. These frequencies are consistent across the models and align well with the IR spectra, demonstrating the MLFFs' ability to accurately capture the region of the PES near the system's equilibrium state. However, differences between the MACE, SO3krates, and SOAP/GAP models are slightly more pronounced in the harmonic spectra compared to the VAF spectra. This highlights the greater sensitivity of physical properties that depend on specific regions of the PES to the accuracy of the MLFFs, in contrast to more "global" statistical properties like VAFs, which reflect a broader portion of the PES.

In ESI, Table SI 4† details the broken chemical bonds manifesting in the instability of MD simulations across all MLFF architectures. The stability of MD simulations of $MAPbI_3$ was primarily compromised by broken covalent bonds within MA cations, although C–N bond breaks were less frequent. Interestingly, the instabilities observed in the MD trajectories generated using the MACE model were triggered by significant fluctuations of the Pb atoms rather than covalent bond breakage.

In summary, kernel-based MLFFs failed to provide stable MD trajectories, with only the SOAP/GAP model successfully generating 2 out of 12 one ns-long trajectories without loss of structural integrity. In contrast, equivariant NNs demonstrated reliable stability and efficiency. Spectral analysis showed that MACE, SO3krates, and SOAP/GAP models aligned well with experimental and DFT-derived spectra, despite missing some peaks, likely due to limitations in the reference DFT calculations. Therefore, the main bottleneck for atomistic simulations in this class of systems (similar to organic molecules) is

obtaining high-quality and representative reference data rather than the MLFF architectures themselves.

3 Guidelines for MLFF

The last two decades of developing MLFFs can be characterized as a rapidly growing research activity to create efficient, accurate, scalable, and transferable ML architectures. While this work continues and no architecture design has been universally accepted yet, other factors have become equally important. At the current level of accuracy achieved by modern MLFFs, the quality and completeness of training data and the training process have become defining factors. Below, we present a list of guidelines to follow in the development, training, and application of MLFFs.

(1) Cross-validation: even the most advanced single MLFF architecture should not be blindly trusted. Cross-checking results between different MLFF models can help to increase the reliability of simulations, particularly where reference data (computational or experimental) is sparse or unavailable.

(2) Detailed performance analysis: comparing MLFFs' performance based on overall MAE, RMSE, or similar aggregate measures are only sensible for simple and small systems with comprehensive datasets. In more complex cases, a detailed analysis of MLFF performance (per atom, per chemical element, per environment type) is crucial.

(3) Reducing heterogeneity of atomic errors: reducing the heterogeneity of atomistic MAEs while maintaining acceptable overall accuracy leads to more reliable MLFFs than those trained solely to minimize aggregated errors.

(4) Training dataset quality: the completeness and composition of training datasets significantly impacts MLFF performance. Using datasets that over-represent certain types of states can decrease overall MAE and RMSE but might lead to incorrect simulation results.

(5) Active learning: active learning and similar iterative techniques for correcting the training set should be intrinsic elements of the MLFF training process. Additionally, complementing the training dataset with structures corresponding to very small and very large interatomic distances, even if such situations are unlikely in expected application conditions, can improve MLFF quality, by enforcing the proper asymptotic behaviours.

(6) Incorporating multiscale forces: for systems with a multiscale composition (*e.g.*, atoms forming molecules, molecules forming molecular clusters), adding corresponding force terms into the MLFF loss function during training (with appropriate weights) can improve the reliability of system behavior during simulations.[96] Minimizing only total atomistic errors might be insufficient and could lead to incorrect behavior of larger-scale system components.

(7) Appropriate accuracy levels: depending on the application, MLFFs with MAEs of, for instance, 0.5 or 0.1 kcal (mol^{-1} $Å^{-1}$) might produce the same results in MD simulations. A more accurate model requires more computationally demanding reference data and is slower in production and training, without providing any significant practical benefits. Even within the same MLFF architecture, modellers should explore the tradeoff between model size, accuracy, and computational efficiency.

(8) Saving training information: it is crucial to document the complete training settings (hyperparameters), MLFF software version, and details of the training and validation datasets to ensure future applicability and potential retraining of an ML model. Ideally, this information should be automatically embedded in the MLFF model files, enabling the exact reproduction of the training process if the initial dataset is available.

(9) Transparency: developers of MLFFs should provide comprehensive details about modifications between different software and ML model versions, optimal preprocessing of training data beyond the intrinsic MLFF procedures, and any related offsets present in the outputs.

By adhering to these guidelines, the development and application of MLFFs can achieve greater reliability, ensuring more robust and trustworthy simulations.

4 Conclusions

The TEA Challenge 2023 extensively examined contemporary MLFF architectures, starting with error and stability assessments in the initial paper, "Crash Testing Machine Learning Force Fields for Molecules, Materials, and Interfaces: Model Analysis in the TEA Challenge 2023". In this paper, we advance to a comprehensive comparative analysis of MD simulations conducted under identical conditions. Our objective is not to single out the best MLFF model but to present a current snapshot of the field, identifying reliable application areas and those requiring further improvement. This study focuses on three types of physical systems: organic molecules, molecule-surface interfaces, and 3D periodic systems.

For organic molecules, we observed excellent agreement between MD results obtained using MACE and SO3krates MLFFs when trained on comprehensive datasets. Discrepancies were primarily in the transition regions between (meta)stable states or large atomic fluctuations, likely due to the incompleteness of the training dataset rather than the ML architecture itself. The sGDML model also performed well for the smaller peptide, providing reliable MD trajectories. In contrast, the other two kernel-based ML models, SOAP/GAP and FCHL19*, exhibited insufficient stability, rendering them unsuitable for extended MD simulations.

Despite the success with MLFFs trained on comprehensive datasets, the dynamics of alanine tetrapeptide and *N*-acetylphenylalanyl-pentaalanyl-lysine molecules in Challenges I and II revealed noticeable artifacts when MLFFs were trained on incomplete datasets. This issue affected both kernel-based models and neural networks, underscoring the importance of reliable, high-quality, and comprehensive training data as a major bottleneck in developing effective MLFFs for organic molecules. Incorporating active learning or similar iterative approaches into MLFF training procedures is crucial to ensure thorough and representative datasets.

In Challenge III, which focused on studying the molecule-surface interface of a 1,8-Naphthyridine molecule on graphene, we identified significant limitations across all MLFFs

participating in the TEA Challenge 2023. Although MLFFs have demonstrated strong performance in modeling covalent bonds, they currently lack the mechanisms needed to effectively capture long-range interactions. Consequently, none of the machine learning models were able to accurately reproduce the molecule desorption process, which can occur during extended MD simulations at virtually any temperature. Enhancing the reference dataset with more configurations that include larger molecule-to-surface distances would necessitate incorporating MLFF components that account for long-range non-covalent interactions. These mechanisms were absent in all but the sGDML MLFFs that participated in the TEA Challenge 2023. Nonetheless, addressing long-range non-covalent interactions remains a major focus for development, and by the time of the manuscript's publication, corresponding architectural elements had been proposed and implemented in some of the MLFFs. One should check the description of the relevant version of an MLFF software package.

Lastly, our evaluation of the 3D periodic system $MAPbI_3$, kernel-based MLFFs struggled to maintain stable molecular dynamics trajectories without structural integrity loss. Conversely, MACE and SO3krates architectures provided stable and similar MD trajectories at 300 K, effectively sampling the part of the PES well-represented in the training dataset. Spectral analysis indicated good alignment of MACE, SO3krates, and SOAP/GAP models with experimental and DFT-derived spectra. Therefore, periodic systems like $MAPbI_3$ can be considered within the reliable application range of modern MLFFs. The primary challenge for accurate atomistic simulations is once again obtaining high-quality and representative reference data. Additionally, optimizing models with respect to their model size to avoid unnecessary computational overhead is essential, especially in long-duration MD simulations of large systems, where achieving significant speedups is crucial.

Abbreviations

SO3	SO3krates
sGDML	Symmetric Gradient Domain Machine Learning
SOAP/ GAP	Smooth Overlap of Atomic Position/Gaussian Approximation Potential
FCHL19*	Faber–Christensen–Huang–Lilienfeld 19
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
(in)com	(in)complete
(un)fold	(un)folded

Data availability

All the TEA 2023 datasets are in the Zenodo archive **https://doi.org/10.5281/zenodo.14138387**.

Author contributions

Igor Poltavsky – conceptualization, data curation, formal analysis, funding acquisition, methodology, project administration, software, supervision, validation, visualization, writing – original draft, writing – review & editing. Anton Charkin-Gorbulin – data curation, formal analysis, investigation, software, validation, visualization, writing – original draft, writing – review & editing. Mirela Puleva – data curation, formal analysis, investigation, project administration, software, validation, visualization, writing – original draft, writing – review & editing. Grégory Fonseca – data curation, formal analysis, investigation, software, validation, visualization, writing – original draft, writing – review & editing. Ilyes Batatia – data curation, formal analysis, methodology, software, writing – review & editing. Nicholas J. Browning – data curation, formal analysis, writing – review & editing. Stefan Chmiela – funding acquisition, supervision, writing – review & editing. Mengnan Cui – data curation, formal analysis, methodology, funding acquisition, writing – review & editing. J. Thorben Frank – data curation, formal analysis, funding acquisition, methodology, software, writing – review & editing. Stefan Heinen – data curation, formal analysis, methodology, software, writing – review & editing. Bing Huang – data curation, formal analysis, methodology, software, writing – review & editing. Silvan Käser – data curation, formal analysis, methodology, software, writing – review & editing. Adil Kabylda – data curation, formal analysis, methodology, software, writing – review & editing. Danish Khan – data curation, formal analysis, methodology, writing – review & editing. Carolin Müller – data curation, funding acquisition, supervision, writing – review & editing. Alastair J. A. Price – data curation, formal analysis, methodology, software, writing – review & editing. Kai Riedmiller – data curation, formal analysis, funding acquisition, writing – review & editing. Kai Töpfer – data curation, formal analysis, methodology, software, writing – review & editing. Tsz Wai Ko – writing – review & editing. Markus Meuwly – funding acquisition, resources, supervision, writing – review & editing. Matthias Rupp – data curation, writing – review & editing. Gabor Csanyi – funding acquisition, resources, supervision, writing – review & editing. O. Anatole von Lilienfeld – funding acquisition, resources, supervision, writing – review & editing. Johannes T. Margraf – funding acquisition, resources, supervision, writing – review & editing. Klaus-Robert Müller – funding acquisition, resources, supervision, writing – review & editing. Alexandre Tkatchenko – conceptualization, funding acquisition, methodology, project administration, resources, supervision, writing – review & editing.

Conflicts of interest

GC has equity interest in Symmetric Group LLP that licenses force fields commercially and also in Ångström AI, Inc.

Acknowledgements

The simulations were performed on the Luxembourg national supercomputer MeluXina. The authors gratefully acknowledge the LuxProvide teams for their expert support. I. Poltavsky would like to acknowledge the financial support afforded by the Luxembourg National Research (FNR) (Grant C19/MS/13718694/QML-FLEX). A. Tkatchenko acknowledges support

from the European Research Council (ERC-AdG grant FITMOL) and Luxembourg National Research Fund (FNR-CORE Grant MBD-in-BMD). M. Puleva acknowledges the financial support from Institute of Advanced Studies, University of Luxembourg under the Young Academics project AQMA. The work of A. Charkin-Gorbulin was performed with the support of the Belgian National Fund for Scientific Research (F.R.S.-FNRS). Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI) funded FNRS under Grant 2.5020.11. We thank J. Weinreich for the helpful discussions. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), RGPIN-2023-04853. O. A. von Lilienfeld has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 772834). This research was undertaken thanks in part to funding provided to the University of Toronto's Acceleration Consortium from the Canada First Research Excellence Fund, grant number: CFREF-2022-00042. O.A. von Lilienfeld has received support as the Ed Clark Chair of Advanced Materials and as a Canada CIFAR AI Chair. I. Batatia acknowledges access to CSD3 GPU resources through a University of Cambridge EPSRC Core Equipment Award (EP/X034712/1). IB was supported by the Harding Distinguished Postgraduate Scholarship. S. Chmiela and J.T. Frank acknowledge support by the German Ministry of Education and Research (BMBF) for BIFOLD (01IS18037A). M. Cui acknowledges the Max Planck Computing and Data Facility (MPCDF) for computation time. J.T. Margraf acknowledges support by the Bavarian Center for Battery Technology (BayBatt) at the University of Bayreuth. The research at Uni Basel was financially supported (to MM) by the Swiss National Science Foundation through grants 200020_219779 and 200021_215088, which is gratefully acknowledged. A. Kabylda acknowledges financial support from the Luxembourg National Research Fund (FNR) (AFR PhD Grant 15720828). C. Müller acknowledges funding by a Feodor-Lynen fellowship of the Alexander von Humboldt foundation. K.-R. Müller was in part supported by the German Ministry for Education and Research (BMBF) under Grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A, 031L0207D, and 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). K. Riedmiller acknowledges financial support from the Klaus Tschira Foundation. We thank O. Unke and C. Quarti for their helpful comments.

References

1 J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, *Phys. Rev. Lett.*, 2007, **98**(14), 146401, DOI: **10.1103/physrevlett.98.146401**.

2 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, *Phys. Rev. Lett.*, 2010, **104**(13), 136403, DOI: **10.1103/physrevlett.104.136403**.

3 J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, *J. Chem. Phys.*, 2011, **134**(7), 074106, DOI: **10.1063/1.3553717**.

4 M. Rupp, A. Tkatchenko, K. R. Müller and O. A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, *Phys. Rev. Lett.*, 2012, **108**(5), 058301, DOI: **10.1103/physrevlett.108.058301**.

5 G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, Accelerating materials property predictions using machine learning, *Sci. Rep.*, 2013, **3**(1), 2810, DOI: **10.1038/srep02810**.

6 A. P. Bartók, R. Kondor and G. Csányi, On representing chemical environments, *Phys. Rev. B*, 2013, **87**(18), 184115, DOI: **10.1103/physrevb.87.184115**.

7 A. Bartók, M. Gillan, F. Manby and G. Csányi, Machine-learning approach for one-and two-body corrections to density functional theory: Applications to molecular and condensed water, *Phys. Rev. B*, 2013, **88**, 054104, DOI: **10.1103/physrevb.88.054104**.

8 W. Szlachta, A. Bartók and G. Csányi, Accuracy and transferability of Gaussian approximation potential models for tungsten, *Phys. Rev. B*, 2014, **90**, 104108, DOI: **10.1103/physrevb.90.104108**.

9 T. D. Huan, A. Mannodi-Kanakkithodi and R. Ramprasad, Accelerated materials property predictions and design using motif-based fingerprints, *Phys. Rev. B*, 2015, **92**, 014106, DOI: **10.1103/PhysRevB.92.014106**.

10 E. Slootman, I. Poltavsky, R. Shinde, J. Cocomello, S. Moroni, A. Tkatchenko, *et al.*, Accurate Quantum Monte Carlo Forces for Machine-Learned Force Fields: Ethanol as a Benchmark, *J. Chem. Theory Comput.*, 2024, **20**(14), 6020–6027, DOI: **10.1021/acs.jctc.4c00498**.

11 A. S. Christensen, L. A. Bratholm, F. A. Faber and O. A. von Lilienfeld, FCHL revisited: Faster and more accurate quantum machine learning, *J. Chem. Phys.*, 2020, **152**(4), 044107, DOI: **10.1063/1.5126701**.

12 N. J. Browning, F. A. Faber and O. Anatole von Lilienfeld, GPU-accelerated approximate kernel method for quantum machine learning, *J. Chem. Phys.*, 2022 12, **157**(21), 214801, DOI: **10.1063/5.0108967**.

13 I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csányi, MACE: Higher order equivariant message passing neural networks for fast and accurate force fields, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 11423–11436, DOI: **10.1007/978-3-030-40245-7_10**.

14 I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky and G. Csányi, The design space of E(3)-equivariant atom-centered interatomic potentials, *Nat. Mach. Intell.*, 2025, **7**, 56–67, DOI: **10.1038/s42256-024-00956-x**.

15 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, Quantum-chemical insights from deep

tensor neural networks, *Nat. Commun.*, 2017, **8**(1), 13890, DOI: 10.1038/ncomms13890.

16 K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko and K. R. Müller, SchNet–a deep learning architecture for molecules and materials, *J. Chem. Phys.*, 2018, **148**(24), 241722, DOI: 10.1063/1.5019779.

17 S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt and K. R. Müller, Machine learning of accurate energy-conserving molecular force fields, *Sci. Adv.*, 2017, **3**(5), e1603015, DOI: 10.1126/sciadv.1603015.

18 S. Chmiela, H. E. Sauceda, K. R. Müller and A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields, *Nat. Commun.*, 2018, **9**(1), 3887, DOI: 10.1038/s41467-018-06169-2.

19 O. T. Unke, M. Stöhr, S. Ganscha, T. Unterthiner, H. Maennel, S. Kashubin, *et al.*, Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments, *Sci. Adv.*, 2024, **10**(14), eadn4397, DOI: 10.1126/sciadv.adn4397.

20 K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, K. R. Müller, *Machine learning meets quantum physics, LNP*, 2020: DOI: 10.1063/pt.3.4164.

21 O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda and K. R. Müller, SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects, *Nat. Commun.*, 2021, **12**(1), 7273, DOI: 10.1038/s41467-021-27504-0.

22 J. T. Frank, O. T. Unke and K. R. Müller, So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 29400–29413, DOI: 10.1103/physrevresearch.2.032064.

23 J. T. Frank, O. T. Unke, K. R. Müller and S. Chmiela, A Euclidean transformer for fast and stable machine learned force fields, *Nat. Commun.*, 2024, **15**(1), 6539, DOI: 10.1038/s41467-024-50620-6.

24 A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, *et al.*, Learning local equivariant representations for large-scale atomistic dynamics, *Nat. Commun.*, 2023, **14**(1), 579, DOI: 10.1038/s41467-023-36329-y.

25 S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, *et al.*, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, *Nat. Commun.*, 2022, **13**(1), 2453, DOI: 10.1038/s41467-022-29939-5.

26 T. Plé, L. Lagardère and J. P. Piquemal, Force-field-enhanced neural network interactions: from local equivariant embedding to atom-in-molecule properties and long-range effects, *Chem. Sci.*, 2023, **14**(44), 12554–12569, DOI: 10.1039/d3sc02581k.

27 K. T. Schütt, O. T. Unke and M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, in *International Conference on Machine Learning. PMLR*, 2021, pp. 9377–9388, DOI: 10.1103/physrevb.105.165131.

28 V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csányi, Gaussian process regression for materials and molecules, *Chem. Rev.*, 2021, **121**(16), 10073–10141, DOI: 10.1021/acs.chemrev.1c00022.

29 G. Fonseca, I. Poltavsky and A. Tkatchenko, Force Field Analysis Software and Tools (FFAST): Assessing Machine Learning Force Fields under the Microscope, *J. Chem. Theory Comput.*, 2023, **19**(23), 8706–8717, DOI: 10.1021/acs.jctc.3c00985.

30 T. Bischoff, B. Jäckl and M. Rupp, Hydrogen under Pressure as a Benchmark for Machine-Learning Interatomic Potentials, *arXiv*, 2024, preprint, arXiv:240913390, DOI: 10.2139/ssrn.3787885.

31 S. Chong, F. Grasselli, M. C. Ben, J. D. Morrow, V. L. Deringer and M. Ceriotti, Robustness of local predictions in atomistic machine learning models, *J. Chem. Theory Comput.*, 2023, **19**(22), 8020–8031, DOI: 10.1021/acs.jctc.3c00704.s001.

32 E. Heid, C. J. McGill, F. H. Vermeire and W. H. Green, Characterizing uncertainty in machine learning for chemistry, *J. Chem. Inf. Model.*, 2023, **63**(13), 4012–4029, DOI: 10.1021/acs.jcim.3c00373.

33 G. Vishwakarma, A. Sonpal and J. Hachmann, Metrics for benchmarking and uncertainty quantification: Quality, applicability, and best practices for machine learning in chemistry, *Trends Chem.*, 2021, **3**(2), 146–156, DOI: 10.1016/j.trechm.2020.12.004.

34 P. Pernot and A. Savin, Probabilistic performance estimators for computational chemistry methods: The empirical cumulative distribution function of absolute errors, *J. Chem. Phys.*, 2018, **148**(24), 241707, DOI: 10.1063/1.5016248.

35 K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing and Z. W. Ulissi, Methods for comparing uncertainty quantifications for material property predictions, *Mach. Learn: Sci. Technol.*, 2020, **1**(2), 025006, DOI: 10.1088/2632-2153/ab7e1a.

36 M. Haghighatlari, G. Vishwakarma, D. Altarawy, R. Subramanian, B. U. Kota, A. Sonpal, *et al.*, ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2020, **10**(4), e1458, DOI: 10.1002/wcms.1458.

37 R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse and M. Bokdam, Phase Transitions of Hybrid Perovskites Simulated by Machine-Learning Force Fields Trained on the Fly with Bayesian Inference, *Phys. Rev. Lett.*, 2019, **122**, 225701, DOI: 10.1103/PhysRevLett.122.225701.

38 H. E. Sauceda, L. E. Gálvez-González, S. Chmiela, L. O. Paz-Borbón, K. R. Müller and A. Tkatchenko, BIGDML—Towards accurate quantum machine learning force fields for materials, *Nat. Commun.*, 2022, **13**(1), 3733, DOI: 10.1038/s41467-022-31093-x.

39 Z. A. H. Goodwin, M. B. Wenny, J. H. Yang, A. Cepellotti, J. Ding, K. Bystrom, *et al.*, Transferability and Accuracy of Ionic Liquid Simulations with Equivariant Machine Learning Interatomic Potentials, *J. Phys. Chem. Lett.*, 2024, **15**(30), 7539–7547, DOI: 10.1021/acs.jpclett.4c01942.

40 J. Vandermause, Y. Xie, J. S. Lim, C. J. Owen and B. Kozinsky, Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of H/Pt, *Nat. Commun.*, 2022, **13**(1), 5183, DOI: 10.1038/s41467-022-32294-0.

41 C. J. Owen, Y. Xie, A. Johansson, L. Sun and B. Kozinsky, Low-index mesoscopic surface reconstructions of Au surfaces using Bayesian force fields, *Nat. Commun.*, 2024, **15**(1), 3790, DOI: 10.1038/s41467-024-48192-6.

42 J. Weinreich, D. Lemm, G. F. von Rudorff and O. A. von Lilienfeld, Ab initio machine learning of phase space averages, *J. Chem. Phys.*, 2022 07, **157**(2), 024303, DOI: 10.1063/5.0095674.

43 J. Weinreich, N. J. Browning and O. A. von Lilienfeld, Machine learning of free energies in chemical compound space using ensemble representations: Reaching experimental uncertainty for solvation, *J. Chem. Phys.*, 2021 04, **154**(13), 134113, DOI: 10.1063/5.0041548.

44 I. B. Magdău, D. J. Arismendi-Arrieta, H. E. Smith, C. P. Grey, K. Hermansson and G. Csányi, Machine learning force fields for molecular liquids: Ethylene Carbonate/Ethyl Methyl Carbonate binary solvent, *npj Comput. Mater.*, 2023, **9**(1), 146, DOI: 10.1038/s41524-023-01100-w.

45 L. L. Schaaf, E. Fako, S. De, A. Schäfer and G. Csányi, Accurate energy barriers for catalytic reaction pathways: an automatic training protocol for machine learning force fields, *npj Comput. Mater.*, 2023, **9**(1), 180, DOI: 10.1038/s41524-023-01124-2.

46 W. J. Baldwin, X. Liang, J. Klarbring, M. Dubajic, D. Dell'Angelo, C. Sutton, *et al.*, Dynamic Local Structure in Caesium Lead Iodide: Spatial Correlation and Transient Domains, *Small*, 2024, **20**(3), 2303565, DOI: 10.1002/smll.202303565.

47 W. G. Stark, C. van der Oord, I. Batatia, Y. Zhang, B. Jiang, G. Csányi, *et al.*, Benchmarking of machine learning interatomic potentials for reactive hydrogen dynamics at metal surfaces, *Mach. Learn: Sci. Technol.*, 2024, **5**(3), 030501, DOI: 10.1088/2632-2153/ad5f11.

48 S. Stocker, H. Jung, G. Csányi, C. F. Goldsmith, K. Reuter and J. T. Margraf, Estimating Free Energy Barriers for Heterogeneous Catalytic Reactions with Machine Learning Potentials and Umbrella Integration, *J. Chem. Theory Comput.*, 2023, **19**, 6796–6804, DOI: 10.1021/acs.jctc.3c00541.

49 A. Musaelian, A. Johansson, S. Batzner and B. Kozinsky, Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2023, vol. 12, pp. 1–12, DOI: 10.1145/3581784.3627041.

50 D. P. Kovács, C. van der Oord, J. Kucera, A. E. Allen, D. J. Cole, C. Ortner, *et al.*, Linear atomic cluster expansion force fields for organic molecules: beyond RMSE, *J. Chem. Theory Comput.*, 2021, **17**(12), 7696–7711, DOI: 10.33774/chemrxiv-2021-7qlf5-v3.

51 Z. Wang, H. Wu, L. Sun, X. He, Z. Liu, B. Shao, *et al.*, Improving machine learning force fields for molecular dynamics simulations with fine-grained force metrics, *J. Chem. Phys.*, 2023, **159**(3), 035101, DOI: 10.1063/5.0147023.

52 X. Fu, Z. Wu, W. Wang, T. Xie, S. Keten, R. Gomez-Bombarelli, *et al.*, Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations, *arXiv*, 2022, preprint arXiv:221007237, DOI: 10.1063/5.0147023.

53 J. D. Morrow, J. L. Gardner and V. L. Deringer, How to validate machine-learned interatomic potentials, *J. Chem. Phys.*, 2023, **158**(12), 121501, DOI: 10.1063/5.0139611.

54 S. Stocker, J. Gasteiger, F. Becker, S. Günnemann and J. T. Margraf, How robust are modern graph neural network potentials in long and hot molecular dynamics simulations? Mach Learn, *Sci. Technol.*, 2022, **3**, 045010, DOI: 10.1088/2632-2153/ac9955.

55 N. Artrith, K. T. Butler, F. X. Coudert, S. Han, O. Isayev, A. Jain, *et al.*, Best practices in machine learning for chemistry, *Nat. Chem.*, 2021, **13**(6), 505–508, DOI: 10.1038/s41557-021-00716-z.

56 A. Bender, N. Schneider, M. Segler, W. W. Patrick, O. Engkvist and T. Rodrigues, Evaluation guidelines for machine learning tools in the chemical sciences, *Nat. Rev. Chem*, 2022, **6**(6), 428–442, DOI: 10.1038/s41570-022-00391-9.

57 P. Pernot, B. Huang and A. Savin, Impact of non-normal error distributions on the benchmarking and ranking of quantum machine learning models, *Mach. Learn: Sci. Technol.*, 2020, **1**(3), 035011, DOI: 10.1088/2632-2153/aba184.

58 S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, *et al.*, Accurate global machine learning force fields for molecules with hundreds of atoms, *Sci. Adv.*, 2023, **9**(2), eadf0873, DOI: 10.1126/sciadv.adf0873.

59 A. S. Christensen, F. A. Faber, B. Huang, L. A. Bratholm, A. Tkatchenko, K. R. Müller, *et al.*, *QML: A Python Toolkit for Quantum Machine Learning*, preprint, 2017, DOI: 10.1201/9781032669182-25.

60 A. Rahimi and B. Recht, Random features for large-scale kernel machines, *Adv. Neural Inf. Process. Syst.*, 2007, **20**, DOI: 10.1109/igarss.2015.7325686.

61 I. Poltavsky, A. Charkin-Gorbulin, M. Puleva, G. Cordeiro Fonseca, I. Batatia, N. J. Browning, S. Chmiela, M. Cui, J. T. Frank, S. Heinen, B. Huang, S. Käser, A. Kabylda, D. Khan, C. Müller, A. J. A. Price, K. Riedmiller, K. Töpfer, T. W. Ko, M. Meuwly, M. Rupp, G. Csanyi, O. Anatole von Lilienfeld, J. T. Margraf, K.-R. Müller and A. Tkatchenko, *Chem. Sci.*, 2025, DOI: 10.1039/d4sc06529h.

62 J. S. Smith, O. Isayev and A. E. Roitberg, ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost, *Chem. Sci.*, 2017, **8**, 3192–3203, DOI: 10.1039/C6SC05720A.

63 R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, *Phys. Rev. B*, 2019, **99**, 014104, DOI: 10.1103/PhysRevB.99.014104.

64 K. Choudhary, B. DeCost, L. Major, K. Butler, J. Thiyagalingam and F. Tavazza, Unified graph neural

network force-field for the periodic table: solid state applications, *Digital Discovery*, 2023, **2**, 346–355, DOI: 10.1039/D2DD00096B.

65 D. Anstine, R. Zubatyuk and O. Isayev, AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs, *ChemRxiv*, 2024, preprint, DOI: 10.26434/chemrxiv-2023-296ch-v2.

66 J. Zeng, D. Zhang, D. Lu, P. Mo, Z. Li, Y. Chen, *et al.*, DeePMD-kit v2: A software package for deep potential models, *J. Chem. Phys.*, 2023 08, **159**(5), 054801, DOI: 10.1063/5.0155600.

67 H. Wang, L. Zhang, J. Han and W. E, DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics, *Comput. Phys. Commun.*, 2018, **228**, 178–184. https://www.sciencedirect.com/science/article/pii/S0010465518300882.

68 A. Rodriguez, C. Lin, H. Yang, M. Al-Fahdi, C. Shen, K. Choudhary, *et al.*, Million-scale data integrated deep neural network for phonon properties of heuslers spanning the periodic table, *npj Comput. Mater.*, 2023, **9**(1), 20, DOI: 10.1038/s41524-023-00974-0.

69 Y. Zhang and B. Jiang, Universal machine learning for the response of atomistic systems to external fields, *Nat. Commun.*, 2023, **14**(1), 6424, DOI: 10.1038/s41467-023-42148-y.

70 J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, *et al.*, On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events, *npj Comput. Mater.*, 2020, **6**(1), 20, DOI: 10.1038/s41524-020-0283-z.

71 P. L. Kang, Z. X. Yang, C. Shang and Z. P. Liu, Global Neural Network Potential with Explicit Many-Body Functions for Improved Descriptions of Complex Potential Energy Surface, *J. Chem. Theory Comput.*, 2023, **19**(21), 7972–7981, DOI: 10.1021/acs.jctc.3c00873.

72 F. Xie, T. Lu, S. Meng and M. Liu, GPTFF: A high-accuracy out-of-the-box universal AI force field for arbitrary inorganic materials, *Sci. Bull.*, 2024, 3525–3532.

73 A. V. Shapeev, Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials, *Multiscale Model. Simul.*, 2016, **14**(3), 1153–1173, DOI: 10.1137/15M1054183.

74 P. L. Houston, C. Qu, A. Nandi, R. Conte, Q. Yu and J. M. Bowman, Permutationally invariant polynomial regression for energies and gradients, using reverse differentiation, achieves orders of magnitude speed-up with high precision compared to other machine learning methods, *J. Chem. Phys.*, 2022 01, **156**(4), 044120, DOI: 10.1063/5.0080506.

75 P. L. Houston, C. Qu, Q. Yu, P. Pandey, R. Conte, A. Nandi, *et al.*, No Headache for PIPs: A PIP Potential for Aspirin Runs Much Faster and with Similar Precision Than Other Machine-Learned Potentials, *J. Chem. Theory Comput.*, 2024, **20**(8), 3008–3018, DOI: 10.1021/acs.jctc.4c00054.

76 Y. Park, J. Kim, S. Hwang and S. Han, Scalable Parallel Algorithm for Graph Neural Network Interatomic Potentials in Molecular Dynamics Simulations, *J. Chem. Theory Comput.*, 2024 06, **20**(11), 4857–4868, DOI: 10.1021/acs.jctc.4c00190, DOI: 10.1021/acs.jctc.4c00190.

77 A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles and G. J. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, *J. Comput. Phys.*, 2015, **285**, 316–330, DOI: 10.1016/j.jcp.2014.12.018.

78 C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, *Nat. Comput. Sci.*, 2022, **2**(11), 718–728, DOI: 10.1038/s43588-022-00349-3.

79 T. E. Markland and M. Ceriotti, Nuclear quantum effects enter the mainstream, *Nat. Rev. Chem*, 2018, **2**(3), 0109, DOI: 10.1038/s41570-017-0109.

80 G. Ramachandran, C. Ramakrishnan and V. Sasisekharan, Stereochemistry of polypeptide chain configurations, *J. Mol. Biol.*, 1963, **7**, 95–99, DOI: 10.1016/s0022-2836(63)80023-6.

81 C. M. Venkatachalam, Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units, *Biopolymers*, 1968, **6**(10), 1425–1436, DOI: 10.1002/bip.1968.360061006.

82 G. N. Ramachandran, V. Sasisekharan. Conformation of Polypeptides and Proteins**The literature survey for this review was completed in September 1967, with the journals which were then available in Madras and the preprinta which the authors had received. †† By the authors' request, the publishers have left certain matters of usage and spelling in the form in which they wrote them, *Advances in Protein Chemistry*. Academic Press, 1968, vol. 23, pp. 283–437, DOI: 10.1007/978-3-662-69832-7_13.

83 M. Ester, H. P. Kriegel, J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96*, AAAI Press, 1996, pp. 226–231, DOI: 10.5120/739-1038.

84 E. Schubert, J. Sander, M. Ester, H. P. Kriegel and X. Xu, DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN, *ACM Trans. Database Syst.*, 2017, **42**(3), 1–21, DOI: 10.1145/3068335.

85 C. Adamo and V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model, *J. Chem. Phys.*, 1999 04, **110**(13), 6158–6170, DOI: 10.1063/1.478522.

86 J. Hermann and A. Tkatchenko, Density functional model for van der Waals interactions: Unifying many-body atomic approaches with nonlocal functionals, *Phys. Rev. Lett.*, 2020, **124**(14), 146401, DOI: 10.1103/physrevlett.124.146401.

87 P. Ren, Y. Xiao, X. Chang, P. Y. Huang, Z. Li, B. B. Gupta, *et al.*, A Survey of Deep Active Learning, *ACM Comput. Surv.*, 2021, **54**(9), 1–40, DOI: 10.1145/3472291.

88 L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay and C. W. Coley, Uncertainty Quantification Using Active Learning for Molecular Property Prediction, *J. Chem. Inf. Model.*, 2020, **60**(8), 3770–3780, DOI: 10.1021/acs.jcim.0c00502.s001.

89 A. R. Tan, S. Urata, S. Goldman, *et al.*, Active Learning for Machine Learning Force Fields in Material Science, *npj Comput. Mater.*, 2023, **9**, 225, DOI: **10.1038/s41524-023-01180-8**.

90 D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Academic Press, 2007, DOI: **10.1063/1.881812**.

91 A. Kabylda, J. T. Frank, S. S. Dou, A. Khabibrakhmanov, L. M. Sandonas, O. T. Unke, *et al.*, Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields, *ChemRxiv*, 2024, preprint, DOI: **10.26434/chemrxiv-2024-bdfr0**.

92 J. J. Choi, X. Yang, Z. M. Norman, S. J. L. Billinge and J. S. Owen, Structure of Methylammonium Lead Iodide Within Mesoporous Titanium Dioxide: Active Material in High-Performance Perovskite Solar Cells, *Nano Lett.*, 2014, **14**(1), 127–133, DOI: **10.1021/nl403514x**.

93 T. Ivanovska, C. Quarti, G. Grancini, A. Petrozza, F. De Angelis, A. Milani, *et al.*, Vibrational Response of Methylammonium Lead Iodide: From Cation Dynamics to Phonon–Phonon Interactions, *ChemSusChem*, 2016, **9**(20), 2994–3004. **https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cssc.201600932**.

94 V. DiezCabanes, S. Giannini, D. Beljonne and C. Quarti, On the Origin of Energetic Disorder in Mixed Halides Lead Perovskites, *Adv. Opt. Mater.*, 2024, **12**(8), 2301105, DOI: **10.1002/adom.202301105**.

95 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, *et al.*, The atomic simulation environment – a Python library for working with atoms, *J. Phys. Condens. Matter*, 2017, **29**(27), 273002, DOI: **10.1088/1361-648x/aa680e**.

96 S. Wengert, G. Csányi, K. Reuter and J. T. Margraf, Data-efficient machine learning for molecular crystal structure prediction, *Chem. Sci.*, 2021, **12**, 4536–4546, DOI: **10.1039/D0SC05765G**.