MLIPAudit: A benchmarking tool for Machine Learned Interatomic Potentials

Anonymous Author(s)
Affiliation
Address
email

Abstract

Machine-learned interatomic potentials (MLIPs) promise to significantly advance atomistic simulations by delivering quantum-level accuracy for large molecular systems at a fraction of the computational cost of traditional electronic structure methods. While model hubs and categorisation efforts have emerged in recent years, it remains difficult to consistently discover, compare, and apply these models across diverse scenarios. The field still lacks a standardised and comprehensive framework for evaluating MLIP performance. We introduce MLIPAudit, an open, curated and modular benchmarking suite designed to assess the accuracy of MLIP models across a variety of application tasks. MLIPAudit offers a diverse collection of benchmark systems, including small organic compounds, molecular liquids, proteins and flexible peptides, along with pre-computed results for a range of pre-trained and published models. MLIPAudit also provides tools for users to evaluate their models using the same standardised pipeline. A continuously updated leaderboard tracks performance across benchmarks, enabling direct comparison on downstream tasks. By offering a unified and transparent reference framework for model validation and comparison, MLIPAudit aims to foster reproducibility, transparency, and community-driven progress in the development of MLIPs for complex molecular systems. The library is available on GitHub and on PyPI 14 under the Apache license 2.0.

1 Introduction

The accurate prediction of molecular and material properties is a cornerstone of scientific progress across disciplines, including drug discovery, functional material design, and process chemistry [1–3]. Traditionally, this has been done using classical force fields, which enable efficient simulations of large systems relying on predefined functional forms and parameters derived from experiments or first-principles methods [4, 5]. Although computationally inexpensive, classical force fields often struggle to capture complex chemical interactions or generalise beyond the systems for which they were parametrised. At the other end of the spectrum, first-principles methods such as density functional theory (DFT) offer higher accuracy but at significantly greater computational cost, typically limiting their use to systems with fewer than a few hundred atoms [6, 7]. In recent years, machine-learned interatomic potentials (MLIPs) have emerged as a compelling middle ground. These models aim to retain the accuracy of first-principles methods while approaching the efficiency of classical force fields, by learning the potential energy surface directly from high-level electronic structure data [8–25].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Despite the rapid emergence of diverse MLIP architectures, which have significantly broadened the scope of atomistic simulations, the field continues to lack a standardised and rigorous framework for evaluating model performance in downstream applications. Many benchmarks focus on energy and force errors, which miss aspects like stability, transferability, and robustness. Recent works propose more holistic evaluations [11, 26–34], which we detail in the Literature Review section. However, all these studies highlight the need for consistent and reproducible evaluation protocols that go beyond basic error metrics, aiming to establish benchmarking practices that reflect real-world simulation demands. Therefore, a universally adopted, comprehensive benchmarking suite that can guide both model development and deployment remains an open challenge for the community.

To address this gap, we introduce MLIPAudit: an open, curated repository of benchmarks, reference datasets, and model evaluations for MLIP models applied (in its first version) to the analysis of small molecules, molecular liquids and biomolecules. MLIPAudit is designed to complement model-centric testing by shifting the focus to systematic validation and comparison. It provides:

- A diverse set of benchmark systems, including organic small molecules, flexible peptides, folded protein domains, molecular liquids and solvated systems.
- Pre-computed results for a range of published and pretrained MLIP models, enabling direct, fair comparisons.
- A continuously updated leaderboard, tracking performance across different tasks.
- A suite of tools for users to submit and evaluate their models within the same benchmarking pipeline.

By providing a shared reference point for assessing accuracy, robustness, and generalisation, MLIPAudit aims to facilitate transparency, reproducibility, and community-wide progress in the development and deployment of MLIPs for complex molecular systems.

2 Literature Review

MLIP Audit aims to expand the existing methods and tools for benchmarking MLIPs. To put this work in context, we summarise current efforts for MLIP benchmarking here.

Static regression metrics: The first and most fundamental level of MLIP evaluation involves the use of standard regression metrics to quantify a model's ability to reproduce the reference quantum-mechanical (QM) data it was trained on. The most common benchmarks in this category are the root-mean-square-error (RMSE) and mean-absolute-error (MAE) calculated for energies and atomic forces on a held-out validation dataset [35]. These benchmarks are routinely reported with the release of new MLIP models, and state-of-the-art models achieve high accuracy on these tests. Although benchmarks for atomic energies and forces are a necessary baseline for the interpolation accuracy of the models, they are insufficient to estimate their practical utility. This is demonstrated, for example, by Gonzales et al. [36], who found that three models with very similar force validation error show significant variation in performance on a structural relaxation task.

Assessment of physical and chemical behaviour: Recent MLIP benchmarks generally accompany model releases and assess performance on physical and chemical properties using QM or experimental data, typically tailored to specific use cases. For models trained on small organic molecules, standard tests include dihedral scans, conformer selection, vibrational frequencies, and interaction energies [32, 37, 38]. Biomolecular benchmarks cover backbone sampling, water properties, and folding dynamics [32, 38, 39], while models trained on reactivity data are evaluated on their ability to reproduce product, reactant, and transition state geometries, as well as reaction pathways via string or NEB methods [33, 40].

Comparative studies have also emerged, evaluating multiple MLIPs across diverse benchmarks. Fu et al. [27] propose a suite spanning organic molecules, peptides, and materials, and find that models with low force errors may still perform poorly on simulation-based metrics like energy conservation

and sampling. Similarly, Liu et al. [41] report discrepancies in atom dynamics and rare events, even for models with strong regression accuracy. These findings reflect a growing consensus that static error metrics alone are insufficient for evaluating MLIPs, and that dynamic and simulation-based benchmarks are increasingly essential.

Standardised benchmarks: While a great variety of benchmarks for accurate physical and chemical properties can be collected from individual model releases and MLIP evaluation studies, a need remains for standardised benchmarks that can be used to compare models on a level playing field and get a holistic view of their utility regarding practical tasks.

This gap is addressed by leaderboards and standardised frameworks. MLIP Arena [26] is a leaderboard based on a benchmark platform focused on physical awareness, stability, reactivity, and predictive power. The framework comprises a small but well-selected suite of benchmarks that address known problems like data leakage, transferability, and overreliance on specific errors. Matbench Discovery [42] features a leaderboard and evaluation framework that is easily extendable to additional models and focused exclusively on materials science. MOFSimBench [43] is a standardised benchmark specialised on metal-organic frameworks that highlights simulation metrics and bulk properties. MLIPX [44] provides a framework with a user-centric perspective, providing a set of reusable recipes that allow users to compose benchmarks for their specific tasks.

These standardised frameworks are valuable tools to evaluate and compare MLIP models. However, they are limited to a specific domain of application, employ a small number of benchmarks or require development by the MLIP user.

3 MLIPAudit Benchmarks

To enable a rigorous and meaningful evaluation of MLIP models, MLIPAudit includes a curated and modular suite of benchmarks that span a range of molecular systems and complexity levels (Figure 1). These benchmarks are designed to capture both general-purpose and domain-specific challenges faced by MLIPs in industrial applications. Benchmark subsets each emphasise different aspects of model performance, such as elemental molecular dynamics stability, non-covalent interactions, conformational ranking of small organic compounds, or sampling of rotamers in biomolecules. A description of the rationale for each benchmark on the different categories is given in Appendix A, including: (i) general systems designed for molecular dynamics stability and scaling, (ii) small molecules relevant to materials chemistry, (iii) molecular liquids, and (iv) biomolecules.

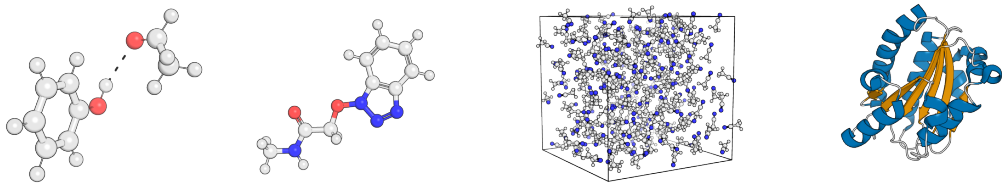

Figure 1: Representative molecular systems spanning increasing levels of structural and environmental complexity, from isolated dimers and drug-like molecules, to condensed-phase molecular liquids and folded biomolecules.

We have evaluated the performance of the three graph-based MLIPs provided in the open-source mlip library [25]: MACE [9], NequIP [11], and ViSNet [39]. All three models were trained on a subset of the SPICE2 dataset [45], which includes 1,737,896 molecular structures across 15 elements (B, Br, C, Cl, F, H, I, K, Li, N, Na, O, P, S, Si). From now on, MACE-SPICE2, NequIP-SPICE2 and Visnet-SPICE2. Training protocols and dataset curation details are available in [25]. Additionally, we trained a new version of each of these models (MACE-t1x, NequIP-t1x, Visnet-t1x) using 10%

(randomly sampled) of the original t1x dataset [46], containing a total of one million structures and four elements (H, C, N, O).

To ensure fair and consistent comparison across models, we define a composite score $S_m \in [0, 1]$ that averages soft-thresholded, normalised benchmark metric scores, rewarding models that approach DFT-level accuracy. Only benchmarks compatible with a model's element set are included, ensuring broad applicability without penalising for unsupported systems. For full details, see Appendix B.

For each benchmark, a set of test cases has been curated (Appendix C, Table 4). As public datasets increase, it becomes increasingly challenging to ensure zero overlap between the training data and the relevant chemistry that one needs to include to ensure the relevance and reliability of the benchmarks. In Appendix C-Table 5, we disclose the overlap between the MLIPAudit test cases per benchmark and the training set for the presented models. In most cases, the overlap is either zero or under 10 $\%$. But, for the conformer selection benchmark, for which two molecules (adenosine and efivarez) from the Wiggle150 [47] dataset were present in the model's training set. In the following, we will discuss the different scores and how the overlap may impact ranking.

3.1 Overall ranking

Table 1 highlights the generalisation capabilities of the top-performing models. Visnet-SPICE2 leads the leaderboard with the highest average score (0.676, followed closely by NequIP-SPICE2 and MACE-SPICE2. All three models were trained on a diverse dataset and evaluated across all 14 benchmarks. These models consistently perform well across domains, underscoring the benefits of comprehensive training and robust architectures. However, it is worth noting that model performance is reflective of training strategy, not solely the model architecture, and it shouldn't be considered an assessment of the model architecture.

Table 1: Overall MLIPAudit scores

Rank	Model Name	Average Score	Benchmarks
1	Visnet-SPICE2	0.676	14/14
2	MACE-SPICE2	0.633	14/14
3	NequIP-SPICE2	0.620	14/14
4	MACE-t1x	0.271	10/14
5	Visnet-t1x	0.270	10/14
6	NequIP-t1x	0.268	10/14

Lower-ranked models, including NequIP-t1x, MACE-t1x, and Visnet-t1x variants, show notably lower scores and narrower benchmark coverage. However, this performance disparity is expected: these models were explicitly trained for reactivity-focused tasks using the t1x dataset [46], which lacks the diversity required to generalise to broader molecular systems. As such, their lower total scores (e.g., 0.268 for NequIP-t1x, 0.270 for Visnet-t1x) do not necessarily indicate inferior model design but rather reflect the trade-off between task-specific optimisation and overall versatility.

3.2 Categorical ranking

In Appendix C-Table 6, we summarise our category-based ranking analysis, which further reveals the specialisation and limitations of each MLIP model across different chemical domains. While Visnet-SPICE2 continues to lead overall, its performance across specific categories reinforces its strength in broad generalisation. It ranks first in both Small Molecules and Biomolecules, with high average scores of 0.578 and 0.727, respectively. Additionally, Visnet-SPICE2 shares the top spot in Molecular Liquids (with an ideal score of 1.0) alongside MACE-SPICE2, further highlighting its robust adaptability. NequIP-SPICE2 performs similarly to Visnet-SPICE2 in small molecule benchmarks. It achieves robust scores in both biomolecular (0.584) and molecular liquid (0.834) benchmarks, suggesting reliable generalisation across chemically diverse systems, without reaching Visnet-SPICE2 performance.

MACE-SPICE2 displays a similar pattern, achieving high performance in molecular liquid benchmarks (average score of 1.0) but showing reduced accuracy in the biomolecular category (0.530). This may reflect limits in capturing the structural and conformational complexity of biomolecules.

The performance of the t1x-trained models (e.g., NequIP-t1x, MACE-t1x, Visnet-t1x) reflects their intended specialisation. These models were trained primarily for reactivity tasks and, as such, show reasonable results on small molecule tasks but limited performance in molecular liquids and biomolecular categories. This outcome aligns with expectations, as the t1x dataset did not include training data representative of condensed-phase systems or protein environments. In Appendix C-Table 8, we have included two Visnet (Visnet-SPICE2-t1x, Visnet-SPICE2-t1x-L) versions trained with SPICE2 and t1x from the OMOL dataset [40] and one MACE version (MACE-SPICE2-t1x) in the Small-molecule category only. These models outperform their other variants with Visnet-SPICE2-t1x-L leading the category.

3.3 Single benchmark highlighted results

3.3.1 Reactivity benchmarks

The generalist models (Visnet-SPICE2, NequIP-SPICE2, MACE-SPICE2) perform notably worse in the reactivity task Table 2. It's worth noting that all models, including the generalists, completed all test cases (100/100 for the nudge elastic band (NEB) benchmark, $\sim$12000/12000 for the transition-state-theory (TST) benchmark, indicating that performance differences stem from modelling accuracy rather than lack of elements in the training set. These results suggest that, in the context of reactivity benchmarks, domain-specific training still offers a measurable edge, especially when accurate prediction of reaction energies or barriers is the primary objective. NeuquIP-t1x leads the NEB benchmark while the new version of Visnet-SPICE2-t1x-L leads the TST benchmark, achieving DFT accuracy for the prediction of activation energies (Figure 2). Suggesting that the different DFT theory levels between the original t1x [46] and the OMOL [40] version might play a role too. However, the relatively modest top scores (e.g., 0.623 for NEB) also indicate room for further improvement, even among specialised models.

Table 2: Reactivity Benchmarks Ranking

Rank	Benchmark	Model Name	Score	Test Cases
1	Small Molecule Reactivity Neb	NequIP-t1x	0.623	100/100
2	Small Molecule Reactivity Neb	Visnet-SPICE2-t1x-L	0.565	100/100
3	Small Molecule Reactivity Neb	MACE-SPICE2-t1x	0.462	100/100
4	Small Molecule Reactivity Neb	MACE-t1x	0.460	100/100
5	Small Molecule Reactivity Neb	Visnet-SPICE2-t1x	0.450	100/100
6	Small Molecule Reactivity Neb	Visnet-t1x	0.410	100/100
7	Small Molecule Reactivity Neb	NequIP-SPICE2	0.140	100/100
8	Small Molecule Reactivity Neb	Visnet-SPICE2	0.100	100/100
9	Small Molecule Reactivity Neb	MACE-SPICE2	0.090	100/100
1	Small Molecule Reactivity Tst	Visnet-SPICE2-t1x-L	0.737	11961/11961
2	Small Molecule Reactivity Tst	MACE-SPICE2-t1x	0.574	11961/11961
3	Small Molecule Reactivity Tst	Visnet-SPICE2-t1x	0.416	11961/11961
4	Small Molecule Reactivity Tst	NequIP-t1x	0.402	11961/11961
5	Small Molecule Reactivity Tst	MACE-t1x	0.381	11961/11961
6	Small Molecule Reactivity Tst	Visnet-t1x	0.361	11961/11961
7	Small Molecule Reactivity Tst	MACE-SPICE2	0.248	11961/11961
8	Small Molecule Reactivity Tst	Visnet-SPICE2	0.246	11961/11961
9	Small Molecule Reactivity Tst	NequIP-SPICE2	0.237	11961/11961

As shown in Figure 2, all t1x-trained models outperform general-purpose MLIPs, while generalist models (e.g., Visnet-SPICE2, MACE-SPICE2) show much larger errors, especially for activation energies. These results reinforce the value of domain-specific training, though even top models leave room for improvement.

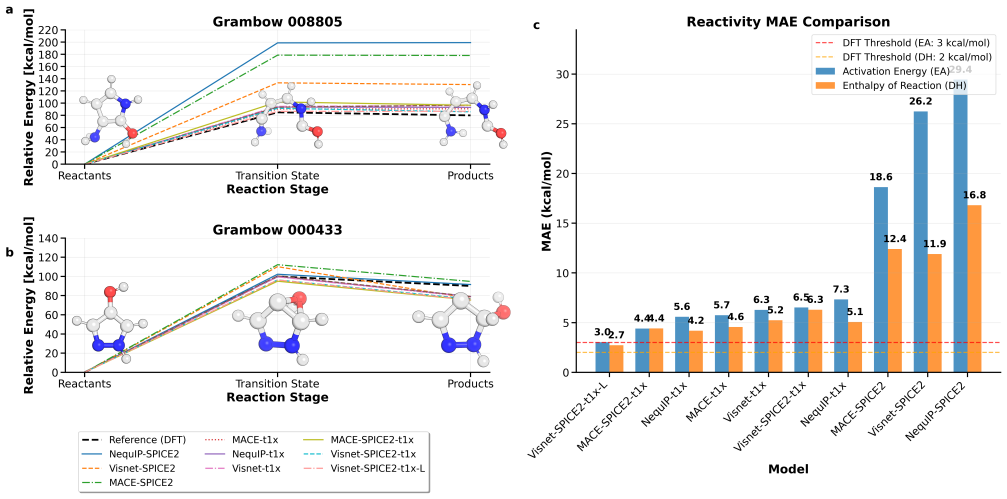

Figure 2: Reactivity benchmark performance. (a–b) Reaction energy profiles for two Grambow reactions (IDs 008805 and 000433) [48] MLIP predictions to DFT references. (c) MAEs for activation energies (EA) and reaction enthalpies across the benchmark.

3.3.2 Molecular liquids benchmark: water radial distribution function

Having a closer look at the single benchmarks, the water radial distribution function (RDF) benchmark provides a compelling illustration of the strengths of MLIPs over traditional force fields. As shown in Appendix C, Figure 6, all three MLIP models, MACE-SPICE2, Visnet-SPICE2, and NequIP-SPICE2, reproduce the experimental RDF profile with high fidelity across the full radial range, accurately reproducing both the first solvation shell peak and subsequent oscillations. In contrast, TIP3P and TIP4P [49], two of the most widely used classical water models, show notable deviations, particularly in the overstructured and exaggerated height of the first peak, a known artefact in rigid water models [50].

This alignment between MLIP predictions and experimental data strongly supports the notion that learned potentials, trained on accurate quantum data, can capture the subtle balance of hydrogen bonding and thermal fluctuations that define liquid water structure, without the need for hand-tuned parameterisation. This not only reflects the higher representational capacity of MLIPs but also demonstrates their ability to generalise to bulk-phase properties, a capability that classical force fields struggle to match without introducing complex polarizable terms or many-body corrections.

3.3.3 Small molecules benchmarks: dihedral scans

The dihedral scan benchmark highlights another area where MLIP models show outstanding agreement with quantum reference data. As shown in Figure 3, the energy profiles predicted by all MLIP models align nearly perfectly with DFT-calculated torsional energy curves across a representative scan. This agreement is not only qualitative—preserving the positions and heights of barriers, but also quantitatively precise, with RMSE values all well below the 1.0 kcal/mol DFT-level convergence threshold. This strong performance is further reflected in the ranking table (Appendix C, Table 8), where Visnet-SPICE2 and Visnet-SPICE2-t1x-L lead the benchmark scoring $\sim$1.0, followed closely by NequIP-SPICE2 and MACE-SPICE2, MACE-SPICE2-t1x. Notably, all models completed the full set of 500 fragments, demonstrating not only accuracy but robustness and generalisability across a diverse chemical space.

The error bars shown on the right panel of Figure 3 underscore how consistent the models are, with MAE values under 0.12 kcal/mol for all methods—well within chemical accuracy. MLIPs outperform classical parameters like GAFF2 [51]. These results validate the capability of current

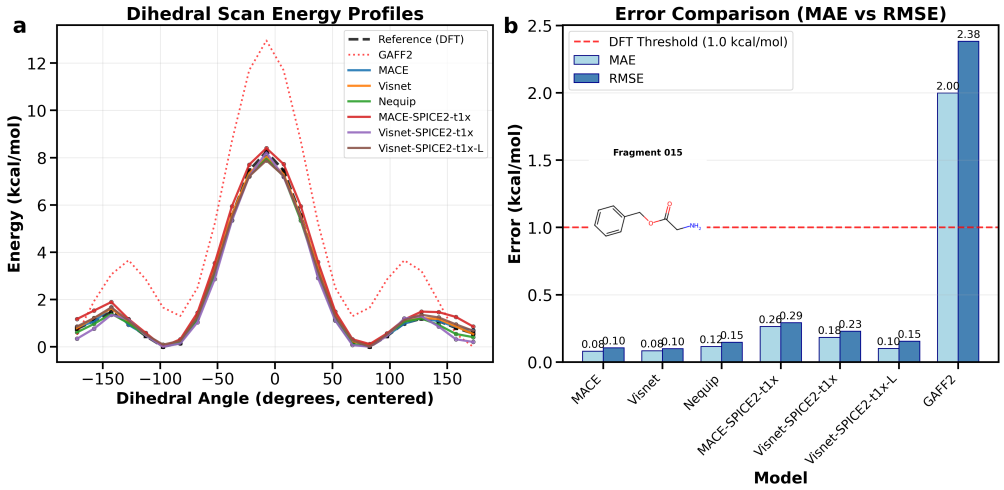

Figure 3: Dihedral scan benchmark. (a) Dihedral energy profiles for fragment 015 compared to DFT reference values. (b) MAE and RMSE for each model. DFT-level error threshold (red dashed line).

MLIPs to accurately model intramolecular potential energy surfaces, a critical requirement for reliable conformational sampling, molecular docking, or pharmacophore prediction.

Taken together, this benchmark provides a clear example of how MLIPs can match DFT accuracy at a fraction of the computational cost, making them practical for high-throughput screening or molecular simulations involving flexible, drug-like molecules.

3.3.4 Small molecules benchmarks: conformer ranking

Figure 4 presents model performance on the conformer benchmark, showing MAE values by molecule for three general-purpose MLIPs: NequIP-SPICE2, Visnet-SPICE2, and MACE-SPICE2. All models were trained on datasets that included adenosine (ADO) and efavirenz (EFA), while benzylpenicillin (BPN) was excluded from training and thus acts as a stronger generalisation test.

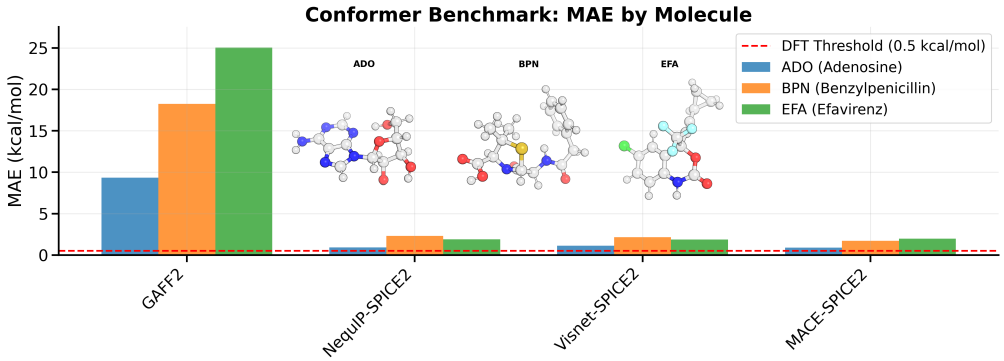

Figure 4: Conformer selection benchmark across three pharmaceutically relevant molecules: adenosine (ADO), benzylpenicillin (BPN), and efavirenz (EFA). MAE is computed with respect to DFT reference conformer energies. DFT threshold (red dashed line at 0.5 kcal/mol). Insets depict representative 3D conformers for each molecule.

Despite having seen ADO and EFA during training, none of the models reach the DFT-level MAE threshold of 0.5 kcal/mol, pointing to persistent difficulty in accurately ranking conformers. ADO is best predicted, while EFA shows higher errors due to its flexibility. BPN, which was unseen during training, is the most challenging, though MACE-SPICE2 shows slightly better generalisation. All models outperform GAFF2 [51], especially on EFA. Still, as seen in Appendix C, Figure 7, predicted

vs. DFT energy plots show strong agreement and near-perfect Spearman correlations across all molecules.

This consistency suggests that while the models may struggle to reproduce exact conformer energy magnitudes (as seen in the MAE analysis), they are highly effective at preserving the correct energetic ordering. In practical applications like conformer selection or ranking, such ordinal accuracy can be more important than precise energetic reproduction, particularly when used in combination with scoring functions or downstream screening.

Interestingly, the performance gap between in-training-set molecules (ADO, EFA) and the out-of-distribution case (BPN) is far less pronounced here than in absolute MAE terms—highlighting that model generalisation, at least in terms of correlation, is relatively robust. These findings reinforce the importance of using multiple complementary metrics (e.g., MAE and rank correlation) when evaluating MLIP performance for conformational energetics.

3.3.5 Biomolecules benchmarks

The biomolecules benchmark (Appendix C, Table 6) provides a fitting conclusion to our comprehensive assessment, highlighting the capacity of MLIP models to operate effectively in complex, biologically relevant regimes. All top models successfully completed the protein folding stability benchmark (6/6 test cases, see Appendix C), all models achieve similar scores $\sim$0.525, but there is room for improvement. This level of agreement underscores the growing maturity of MLIPs for macromolecular tasks. The Protein Sampling benchmark across different MLIP models shows that models trained on the SPICE2 dataset (e.g., Visnet-SPICE2, NequIP-SPICE2, MACE-SPICE2) significantly outperform their t1x-trained counterparts, with Visnet-SPICE2 achieving the highest score (0.928) and full coverage (12/12 systems). Taken together, the results from this and all previous benchmarks reinforce a central conclusion: while task-specific training offers advantages in specialised domains, the leading generalist models demonstrate strong, transferable performance across molecular scales and properties, setting the stage for robust deployment in real-world chemistry and biology applications.

3.4 Conclusions and future outlook

The MLIPAudit suite provides a comprehensive and diverse evaluation framework for MLIPs, spanning small-molecule geometrical and conformational energetics, reactivity, molecular liquids, and biomolecular stability and sampling. Our results show that while specialised models trained on the t1x dataset excel in targeted tasks such as reaction barrier prediction, general-purpose architectures like Visnet-SPICE2, NequIP-SPICE2, and MACE-SPICE2 exhibit strong and transferable accuracy across a wide range of benchmarks, often surpassing classical force fields and closely matching DFT reference data in others. Notably, the Visnet model trained on SPICE2 and t1x from the OpenMolecules (OMOL) dataset leads the small-molecule benchmarks, highlighting the promise of hybrid training strategies and possibly reflecting the importance of the underlying level of theory used in data generation.

Despite this progress, performance gaps persist, especially in condensed-phase systems and energetically subtle regimes, indicating that further improvements are needed. Looking ahead, we plan to expand the MLIPAudit suite with new test cases targeting larger, more complex systems, as well as emerging tasks such as vibrational frequencies and binding free energies. We also aim to include newly released open-source models to keep the benchmark current and representative. By continually broadening the scope and complexity of MLIPAudit, we hope to accelerate the development of MLIPs that are not only accurate but also general, scalable, and ready for real-world deployment across the chemical sciences.

References

[1] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, 2012.

[2] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[3] Gerbrand Ceder and Kristin Persson. The stuff of dreams. *Scientific American*, 309(3):36–41, 2013.

[4] William D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[5] Alexander D MacKerell Jr, Donald Bashford, Michael Bellott, Roland L Dunbrack Jr, John D Evanseck, Michael J Field, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[6] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133, 1965.

[7] Robert G Parr and Weitao Yang. *Density-functional theory of atoms and molecules*. Oxford University Press, 1989.

[8] Yury Lysogorskiy, Chris Van Den Oord, Alexey Bochkarev, Shyue Ping Menon, Matteo Rinaldi, Tobias Hammerschmidt, Michael Mrovec, Alexander Thompson, Gábor Csányi, Christoph Ortner, et al. Performant implementation of the atomic cluster expansion (pace) and application to copper and silicon. *npj Computational Materials*, 7:97, 2021.

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

[10] Dávid Péter Kovács, Ilyes Batatia, Eszter Sara Arany, and Gabor Csanyi. Evaluation of the mace force field architecture: From medicinal chemistry to materials science. *The Journal of Chemical Physics*, 159(4):044118, 2023.

[11] Sebastian Batzner, Alexander Musaelian, Linfeng Sun, Michael Geiger, Jonathan P Mailoa, Marc Kornbluth, Nicola Molinari, Tyle Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13:2453, 2022.

[12] Vitalii Zaverkin, Daniel Holzmüller, Luca Bonfirraro, and Johannes Kästner. Transfer learning for chemically accurate interatomic neural network potentials. *Physical Chemistry Chemical Physics*, 25:5383, 2023.

[13] Mehdi Haghighatlari, Jia Li, Xiangyu Guan, Oliver Zhang, Abhishek Das, Christoph J Stein, Fatemeh Heidar-Zadeh, Meng Liu, Martin Head-Gordon, Lucas Bertels, et al. Newtonnet: A newtonian message passing network for deep learning of interatomic potentials and forces. *Digital Discovery*, 1:333, 2022.

[14] Alexander V Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3):1153–1173, 2016.

[15] David Anstine, Roman Zubatyuk, and Olexandr Isayev. Aimnet2: A neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chemical Science*, 2025.

[16] A Kabylda, V Vassilev-Galindo, S Chmiela, I Poltavsky, and Alexandre Tkatchenko. Efficient interatomic descriptors for accurate machine learning force fields of extended molecules. *Nature Communications*, 14:3562, 2023.

[17] Jörg Behler. Four generations of high-dimensional neural network potentials. *Chemical Reviews*, 121(16):10037–10072, 2021.

[18] Federico Musil, Andrea Grisafi, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. Physics-inspired structural representations for molecules and materials. *Chemical Reviews*, 121(16):9759–9815, 2021.

[19] Volker L Deringer, Albert P Bartók, Noam Bernstein, Daniel M Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian process regression for materials and molecules. *Chemical Reviews*, 121(16):10073–10141, 2021.

[20] Bing Huang and O Anatole Von Lilienfeld. Ab initio machine learning in chemical compound space. *Chemical Reviews*, 121(16):10001–10036, 2021.

[21] Murray S Daw and MI Baskes. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Physical Review B*, 29(12):6443, 1984.

[22] Volker L Deringer, Noam Bernstein, Gábor Csányi, C Ben Mahmoud, Michele Ceriotti, Mark Wilson, David A Drabold, and Steven R Elliott. Origins of structural and electronic transitions in disordered silicon. *Nature*, 589:59–64, 2021.

[23] William J Baldwin, Xiaoxuan Liang, Johan Klarbring, Marija Dubajic, Diego Dell'Angelo, Charles Sutton, Chiara Caddeo, Samuel D Stranks, Alessandro Mattoni, Aron Walsh, et al. Dynamic local structure in caesium lead iodide: Spatial correlation and transient domains. *Small*, 20(2303565), 2023.

[24] Christopher W Rosenbrock, Konstantin Gubaev, Alexander V Shapeev, László B Pártay, Noam Bernstein, Gábor Csányi, and Gus L W Hart. Machine-learned interatomic potentials for alloys and alloy phase diagrams. *npj Computational Materials*, 7:24, 2021.

[25] Christoph Brunken, Olivier Peltre, Heloise Chomet, Lucien Walewski, Manus McAuliffe, Valentin Heyraud, Solal Attias, Martin Maarand, Yessine Khanfir, Edan Toledo, Fabio Falcioni, Marie Bluntzer, Silvia Acosta-Gutiérrez, and Jules Tilly. Machine learning interatomic potentials: library for efficient training, model development and simulation of molecular systems. *arXiv preprint*, 2025.

[26] Yuan Chiang, Tobias Kreiman, Elizabeth Weaver, Ishan Amin, Matthew Kuner, Christine Zhang, Aaron Kaplan, Daryl Chrzan, Samuel Blau, Aditi Krishnapriyan, and Mark Asta. MLIP Arena: Fair and transparent benchmark of machine-learned interatomic potentials. *AI4Mat ICLR*, 2025.

[27] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations. *Transactions on Machine Learning Reasearch*, 4, 2023.

[28] Tristan Maxson, Ademola Soyemi, Benjamin W. J. Chen, and Tibor Szilvási. Enhancing the quality and reliability of machine learning interatomic potentials through better reporting practices. *The Journal of Physical Chemistry C*, 2024.

[29] Christoph Ortner and Yangshuai Wang. A framework for a generalisation analysis of machine-learned interatomic potentials. *arXiv preprint*, 2022.

[30] Michael J. Waters and James M. Rondinelli. Benchmarking structural evolution methods for training of machine learned interatomic potentials. *arXiv preprint*, 2022.

[31] Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jörg Behler, Gábor Csányi, Alexander V. Shapeev, Aidan P. Thompson, Mitchell A. Wood, and Shyue Ping Ong. A performance and cost assessment of machine learning interatomic potentials. *arXiv preprint*, 2019.

[32] Dávid Péter Kovács, J. Harry Moore, Nicholas J. Browning, Ilyes Batatia, Joshua T. Horton, Yixuan Pu, Venkat Kapil, William C. Witt, Ioan-Bogdan Magdău, Daniel J. Cole, and Gábor Csányi. MACE-OFF: Transferable Short Range Machine Learning Force Fields for Organic Molecules. *Journal of the American Chemical Society*, 2025.

[33] Dylan M. Anstine, Qiyuan Zhao, Roman Zubatiuk, Shuhao Zhang, Veerupaksh Singla, Filipp Nikitin, Brett M. Savoie, and Olexandr Isayev. AIMNet2-rxn: A Machine Learned Potential for Generalized Reaction Modeling on a Millions-of-Pathways Scale. *ChemRxiv preprint*, 2025.

[34] Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. Uma: A family of universal models for atoms. *arXiv preprint*, 2025.

[35] Joe D. Morrow, John L. A. Gardner, and Volker L. Deringer. How to validate machine-learned interatomic potentials. *The Journal of Chemical Physics*, 158:121501, 2023.

[36] Carmelo Gonzales, Eric Fuemmeler, Ellad Tadmor, Stefano Martiniani, and Santiago Miret. Benchmarking of Universal Machine Learning Interatomic Potentials for Structural Relaxation. *AI4Mat NeurIPS*, 2024.

[37] Anders S. Christensen, Sai Krishna Sirumalla, Zhuoran Qiao, Michael B. O'Connor, Daniel G. A. Smith, Feizhi Ding, Peter J. Bygrave, Animashree Anandkumar, Matthew Welborn, Frederick R. Manby, and Thomas F. Miller. OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *The Journal of Chemical Physics*, 155:204103, 2021.

[38] John L. Weber, Rishabh D. Guha, Garvit Agarwal, Yujing Wei, Aidan A. Fike, Xiaowei Xie, James Stevenson, Karl Leswing, Mathew D. Halls, Robert Abel, and Leif D. Jacobson. Efficient Long-Range Machine Learning Force Fields for Liquid and Materials Properties. *arXiv preprint*, 2025.

[39] Yusong Wang, Tong Wang, Shaoning Li, Xinheng He, Mingyu Li, Zun Wang, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15(313), 2024.

[40] Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models. *arXiv preprint*, 2025.

[41] Yunsheng Liu, Xingfeng He, and Yifei Mo. Discrepancies and error evaluation metrics for machine learning interatomic potentials. *npj Computational Materials*, 9(174), 2023.

[42] Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. Matbench Discovery – A framework to evaluate machine learning crystal stability predictions. *arXiv preprint*, 2024.

[43] Hendrik Kraß, Ju Huang, and Seyed Mohamad Moosavi. MOFSimBench: Evaluating Universal Machine Learning Interatomic Potentials In Metal–Organic Framework Molecular Modeling. *arXiv preprint*, 2025.

[44] Fabian Zills, Sheena Agarwal, Tiago Goncalves, Srishti Gupta, Edvin Fako, Shuang Han, Imke Mueller, Christian Holm, and Sandip De. MLIPX: Machine Learned Interatomic Potential eXploration. *ChemRxiv preprint*, 2025.

[45] Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(11), 2023.

[46] Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9 (779), 2022.

[47] Rebecca Brew, Ian Nelson, Meruyert Binayeva, Amlan Nayak, Wyatt Simmons, Joseph Gair, and Corin Wagen. Wiggle150: Benchmarking density functionals and neural network potentials on highly strained conformers. *ChemRxiv preprint*, 2025.

[48] L. Pattanaik Grambow and W. H. Green. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data*, 7(137), 2020.

[49] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics*, 79: 926–935, 1983.

[50] Gaia Camisasca, Harshad Pathak, Kjartan Thor Wikfeldt, and Lars G. M. Pettersson. Radial distribution functions of water: Models vs experiments. *The Journal of Chemical Physics*, 151: 044502, 2019.

[51] Xibing He, Viet H. Man, Wei Yang, Tai-Sung Lee, and Junmei Wang. A fast and high-quality charge model for the next generation general amber force field. *The Journal of Chemical Physics*, 153:114502, 2020.

[52] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528 – 1532, 2015. doi: 10.1016/j.bpj.2015.08.015.

[53] Skolnick J. Zhang Y. Scoring function for automated assessment of protein structure template quality. *Proteins.*, 57(4):702–710, 2004.

[54] Sander C. Kabsch W. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded networks in three-dimensional structures. *Biopolymers*, 22(12):2577–637, 1983.

[55] S.C. Lovell, J.M. Word, J.S. Richardson, and D.C. Richardson. The penultimate rotamer library. *Proteins*, 40:389–408, 2000.

[56] David M. Mount Songrit Maneewongvatana. Analysis of approximate nearest neighbor searching with clustered point sets. *ArXiv*, 1999. doi: https://doi.org/10.48550/arXiv.cs/9901013.

[57] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[58] Lorenzo D'Amore, David F. Hahn, David L. Dotson, Joshua T. Horton, Jamshed Anwar, Ian Craig, Thomas Fox, Alberto Gobbi, Sirish Kaushik Lakkaraju, Xavier Lucas, Katharina Meier, David L. Mobley, Arjun Narayanan, Christina E. M. Schindler, William C. Swope, Pieter J. in 't Veld, Jeffrey Wagner, Bai Xue, and Gary Tresadern. Collaborative assessment of molecular geometries and energies from the open force field. *Journal of Chemical Information and Modeling*, 62(23):6094—-6104, 2022.

[59] Brajesh K. Rai, Vishnu Sresht, Qingyi Yang, Ray Unwalla, Meihua Tu, Alan M. Mathiowetz, and Gregory A. Bakken. Torsionnet: A deep neural network to rapidly predict small-molecule torsional energy profiles with the accuracy of quantum mechanics. *Journal of Chemical Information and Modeling*, 62(4):785–800, 2022.

[60] Oya Wahl and Thomas Sander. Tautobase: An open tautomer database. *Journal of Chemical Information and Modeling*, 60(3):1085–1089, 2020.

[61] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

[62] Lawrie B. Skinner; Congcong Huang; Daniel Schlesinger; Lars G. M. Pettersson; Anders Nilsson; Chris J. Benmore. Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide q-range. *The Journal of Chemical Physics*, 138:074506, 2013.

[63] Yoshitada Murata Keiko Nishikawa. Liquid structure of carbon tetrachloride and long-range correlation. *Bulletin of the Chemical Society of Japan*, 52:293–298, 1979.

[64] Evert Jan Meijer Jan-Willem Handgraaf, Titus S van Erp. Ab initio molecular dynamics study of liquid methanol. *Chemical Physics Letters*, 367:617–624, 2003.

[65] László Pusztai Szilvia Pothoczki. Intermolecular orientations in liquid acetonitrile: New insights based on diffraction measurements and all-atom simulations. *Journal of Molecular Liquids*, 225:160–166, 2017.

A Benchmarks overview

Each benchmark in MLIP-Audit includes a brief introduction that outlines its purpose, helping users understand the relevance of the task and how it reflects molecular challenges. A link to the documentation is provided for users who want a deeper explanation of the benchmark's design, scientific context, datasets and implementation details. A description of each benchmark's dataset can be found in Appendix C-Table 4. This is followed by key performance metrics for the best-performing model, along with a summary of results across all analysed MLIP models. Depending on the nature of the benchmark, additional visualisations may be included, such as radial distribution functions for molecular liquids or torsion energy profiles for small molecules, which users can explore interactively or download for further analysis (Figure 5).

In the following subsections, we describe the composition, rationale, and evaluation criteria for each benchmark category: (i) general systems designed for molecular dynamics stability and scaling, (ii) small molecules relevant to pharmaceutical and materials chemistry, and (iii) biomolecules, which pose unique challenges due to their size, flexibility, and hierarchical structure.

A.1 General benchmarks

The general benchmarks implemented in MLIP Audit are system-agnostic and focus on fundamental molecular dynamics (MD) stability and performance metrics that are applicable across molecular systems. Two benchmarks are included in this category:

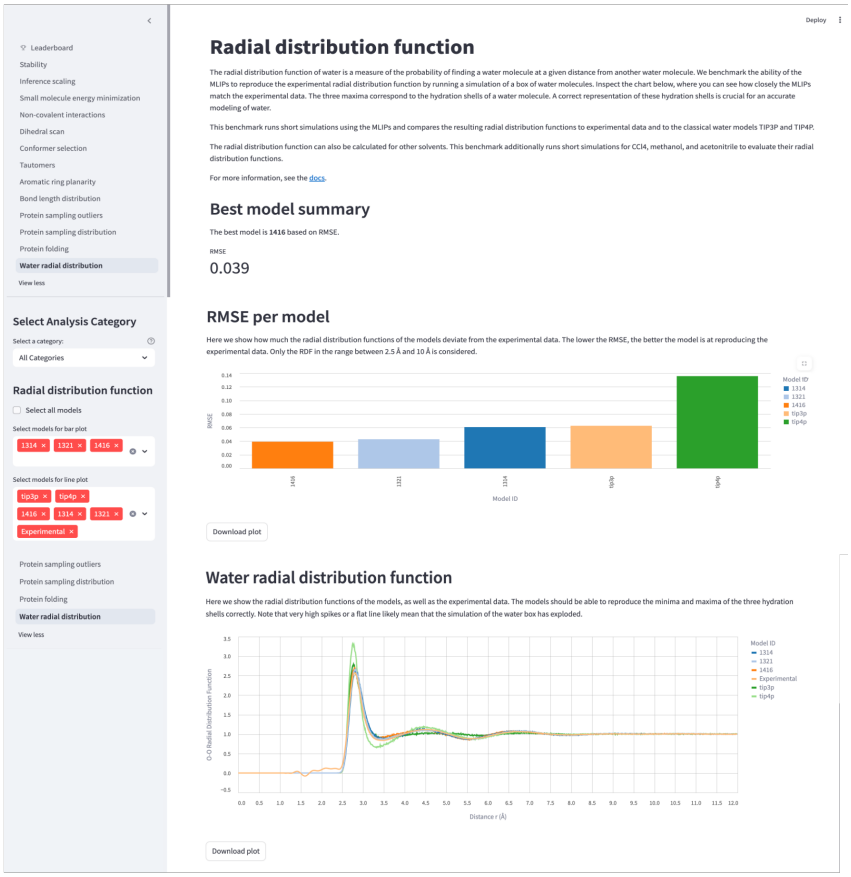

Figure 5: MLIPAudit interface

- **Stability**: assesses the dynamical stability of an MLIP during an MD simulation for a diverse set of large biomolecular systems. For each system, the benchmark performs an MD simulation using the MLIP model in the NVT ensemble at 300 K for 100,000 steps (100 ps), leveraging the jax-md engine, as integrated via the mlip library[25]. The test monitors the system for signs of instability by detecting abrupt temperature spikes ("explosions") and hydrogen atom drift. These indicators help determine whether the MLIP maintains stable and physically consistent dynamics over extended simulation times.
- **Inference Scaling**: evaluates how the computational cost of an MLIP scales with the system size. By running single, long MD episodes on a series of molecular systems of increasing size, we systematically assess the relationship between molecular complexity and inference performance. This benchmark is not used for scoring, but it aims at helping the user to pick the best model in terms of time-to-solution for the application task.

A.2 Small Molecules

MLIPAudit small-molecule benchmarks focus on the ability of MLIPs to reproduce the properties and dynamics of small organic molecules, including their conformational sampling and interactions with other molecules. In order of task complexity:

- **Bond Length**: evaluates the ability of MLIPs to accurately model the equilibrium bond lengths of small organic molecules during MD simulations. This is an important test to understand whether the MLIP respects basic chemistry throughout simulations. Accurate prediction of bond length is crucial for capturing the structural and electronic properties of any chemically relevant compounds. For each molecule in the dataset, the benchmark

performs an MD simulation with the same configuration described in the stability benchmark. Throughout the trajectory, the positions of the bond atoms are tracked, and their deviation from a reference bond length of the QM optimised starting structure is calculated. The average deviation over the trajectory provides a direct measure of the MLIP's ability to maintain bond lengths under thermal fluctuations, enabling quantitative comparison to reference data or other models.

- **Ring Planarity**: evaluates the ability of MLIPs to preserve the planarity of aromatic and conjugated rings in small organic molecules during molecular dynamics simulations. Aromatic rings (e.g., benzene) are inherently planar due to delocalised π electrons. Ring planarity enforcement is crucial in molecular dynamics simulations because it preserves the correct geometry, electronic structure, and interactions of aromatic and conjugated systems. Without proper planarity (e.g., via improper torsions), simulations can produce chemically unrealistic distortions that compromise accuracy in energy, flexibility, and binding predictions. This is especially important in molecules like benzene, tyrosine side chains, nucleobases, and drug scaffolds, where planarity governs stacking, hydrogen bonding, and overall stability. For each molecule in the dataset, the benchmark performs an MD simulation with the same configuration described in the stability benchmark. Throughout the trajectory, the positions of the ring atoms are tracked, and their deviation from a perfect plane is quantified using the root mean square deviation (RMSD) from planarity. The ideal plane of the ring is computed using a principal component analysis of the ring's atoms. The average deviation over the trajectory provides a direct measure of the MLIP's ability to maintain ring planarity under thermal fluctuations, enabling quantitative comparison to reference data or other models.

- **Dihedral Scan**: evaluates the MLIP's ability to reproduce torsional energy profiles of rotatable bonds in small molecules, aiming to approach the quantum-mechanical QM reference quality. Dihedral scans are essential for mapping how a molecule's energy changes as bonds rotate, revealing preferred conformations and energy barriers. Beyond force field development, they are also used in studying reaction mechanisms, analysing conformational dynamics in drug discovery, validating quantum chemistry methods, and guiding the design of flexible or constrained molecules. For each molecule, the benchmark leverages the mlip library for model inference, comparing the predicted energies along a dihedral scan to QM reference energy profiles. The reference profile is shifted so that its global minimum is zero, and the MLIP profile is aligned to the same conformer. Performance is quantified using the following metrics: MAE and RMSE. The Pearson correlation coefficient between the MLIP-predicted and reference datapoints and the mean barrier height error.

- **Non-covalent Interactions**: tests if the MLIP can reproduce interaction energies of molecular complexes driven by non-covalent interactions. Non-covalent interactions are of the highest importance for the structure and function of every biological molecule. This benchmark assesses a broad range of interaction types: London dispersion, hydrogen bonds, ionic hydrogen bonds, repulsive contacts and sigma hole interactions. Assessing the accuracy of non-covalent interactions is crucial for evaluating how well computational models capture key forces like hydrogen bonding, π-π stacking, and van der Waals interactions that govern molecular recognition, binding, and assembly. This is essential not only for force field development, but also for validating quantum methods, guiding molecular design, modelling biomolecular interfaces, and studying condensed-phase behaviour such as solvation and aggregation. The benchmark runs energy inference on all structures of the distance scans of bi-molecular complexes in the dataset. The key metric is the RMSE of the interaction energy, which is the minimum of the energy well in the distance scan, relative to the energy of the dissociated complex, compared to the reference data. For repulsive contacts, the maximum of the energy profile is used instead. Some of the molecular complexes in the benchmark dataset contain exotic elements (see dataset section). In case the MLIP has never seen an element of a molecular complex, this complex will be skipped in the benchmark.

- **Geometrical Minimisation**: assesses the MLIP's capability to preserve the ground-state geometry of organic small molecules during energy minimisation, ensuring that initial X-ray or DFT-optimised structures remain accurate and physically consistent. Each system is minimised over 1,000 steps using the FIRE (Fast Inertial Relaxation Engine) algorithm (default parameters). After minimisation, structural fidelity is assessed by computing the RMSD of all heavy atoms relative to the initial geometry, using the RMSD implementation provided by mdtraj [52].
- **Conformer Selection**: evaluates the MLIP's ability to identify the most stable conformers within an ensemble of flexible organic molecules and accurately predict their relative energy differences. It focuses on capturing subtle intramolecular interactions and strain effects that influence conformational energies. These metrics assess both numerical accuracy and the MLIP's ability to preserve relative conformer energetics, which is crucial for downstream applications such as conformational sampling and compound ranking.
- **Tautomers**: assesses the ability of MLIP to accurately predict the relative energies and stabilities of tautomeric forms of small molecules in vacuum. Tautomers are structural isomers that interconvert via proton transfer and/or double bond rearrangement, and accurately estimating the energy gap between them is an important measure of chemical accuracy in the MLIP framework. Tautomer ranking assesses a model's ability to predict the relative stability of different tautomeric forms of a molecule, which is critical for accurately modelling protonation states, reactivity, and binding affinities. It is especially important in drug discovery, quantum method benchmarking, and cheminformatics, where tautomers can dramatically affect molecular properties and biological activity. For each molecule, the benchmark compares MLIP-predicted energies against QM reference data. Performance is quantified by comparing the absolute deviation of the energy difference between the tautomeric forms from the DFT data.
- **Reactivity**: assesses the MLIP's capability to model chemical reactivity. The reactivity-tst benchmark tests the ability to predict the energy of transition states relative to the reaction's reactants and products and thereby the activation energy and enthalpy of a reaction. This benchmark calculates the energy of reactants, products and transition states of a large dataset of reactions. From the difference between these states, the activation energy and enthalpy of formation can be calculated. The performance is quantified using the MAE and RMSE in activation energy and enthalpy of formation. The reactivity-neb benchmark evaluates the capability to converge a set of nudged elastic band calculations with a known transition state. The performance is quantified by the percentage of converged calculations.

A.3 Molecular Liquids

The MLIP Audit molecular liquids benchmark focuses on assessing long-range interactions by computing the radial distribution function for different molecular liquids.

- **Radial Distribution Function**: assesses the ability of MLIP to accurately reproduce the radial distribution function (RDF) of liquids. The RDF characterises the local and intermediate-range structure of a liquid by describing how particle density varies as a function of distance from a reference particle. Accurate modelling of the RDF is essential for capturing both short-range ordering and long-range interactions, which are critical for understanding the microscopic structure and emergent properties of liquid systems. The benchmark performs an MD simulation using the MLIP model in the NVT ensemble at 300 K for 500,000 steps, leveraging the jax-md engine from the mlip library. The starting configuration is already equilibrated. For every specific atom pair (e.g., oxygen-oxygen in water), the radial distribution function (RDF or g(r)) is calculated from the simulation, as:

$$g(r) = \frac{1}{4\Pi r^2 \rho N} \langle \sum_{i=1}^{N} \sum_{j \neq i}^{N} \delta(r - r_{ij}) \rangle \tag{1}$$

where: r is the distance from a reference particle,ρ is the average number density, N is the number of particles, r_{ij} is the distance between particles and δ is the Dirac delta function. Angle brackets denote an ensemble average. For each test case, the benchmark computes $r_{\text{peak}} = \arg\max_r g(r)$ and compares it with the experimental value for the first solvation shell.

A.4 Biomolecules

MLIP Audit biomolecule benchmarks focus on assessing the properties and dynamics of proteins, including their folding behaviour, structural stability, and conformational sampling.

- **Protein folding**: evaluates the ability of an MLIP to preserve the structural integrity of experimentally determined protein conformations during MD simulations. It assesses the retention of secondary structure elements and overall compactness across a set of known protein structures. This module analyses the folding trajectories of proteins in MLIP simulations. For each molecule in the dataset, the benchmark performs an MD simulation with the same configuration described in the stability benchmark. We track how Root Mean Square Deviation (RMSD), TM Score [53], Dictionary of Secondary Structure in Proteins (DSSP) [54] and Radius of Gyration change over time.
- **Sampling Outlier Detection**: Assesses the structural quality of sampled conformations by computing backbone Ramachandran angles (ϕ/ψ) and side-chain rotamer angles (χ), and identifying outliers through comparison with reference rotamer libraries [55]. For each molecule in the dataset, the benchmark performs an MD simulation with the same configuration described in the stability benchmark. The outlier detection identifies residues whose dihedral angles fall outside expected ranges, relying on the fast KDtree [56] scipy [57] implementation. The analysis provides a global percentage of outliers for backbone and rotamers per structure, as well as a more detailed analysis per residue type.

B Benchmarks scoring

To enable consistent and fair comparison across models, we define a composite score that aggregates performance over all compatible benchmarks. Each benchmark $b \in \mathcal{B}$ may report one or more metrics $x_{m,b}^{(i)}$, where $i = 1, \ldots, N_b$ indexes the N_b metrics evaluated for the model m. For each metric, we compute a normalised score using a soft thresholding function based on a DFT-derived reference tolerance $t_b^{(i)}$ (see 3):

$$s_{m,b}^{(i)} = \begin{cases} 1, & \text{if } x_{m,b}^{(i)} \leq t_b^{(i)} \\ \exp\left(-\alpha \cdot \frac{x_{m,b}^{(i)} - t_b^{(i)}}{t_b^{(i)}}\right), & \text{otherwise} \end{cases}$$

where α is a tunable parameter controlling the steepness of the penalty (e.g., $\alpha = 3$). The per-benchmark score is then computed as the average over all its metric scores:

$$s_{m,b} = \frac{1}{N_b} \sum_{i=1}^{N_b} s_{m,b}^{(i)}$$

Let $\mathcal{B}_m \subseteq \mathcal{B}$ denote the subset of benchmarks for which the model m has valid data (i.e., benchmarks compatible with its element set). The final model score is the mean over all benchmarks on which the model could be evaluated:

$$S_m = \frac{1}{|\mathcal{B}_m|} \sum_{b \in \mathcal{B}_m} s_{m,b}$$

This scoring framework ensures that models are rewarded for meeting or exceeding DFT-level accuracy, but are not penalised for benchmarks they cannot run due to missing chemical elements. Benchmarks with multiple metrics contribute proportionally, and the result is a single interpretable score $S_m \in [0, 1]$ that balances physical fidelity, chemical coverage, and overall model robustness.

Table 3: Acceptable error ranges for Classical Force Fields (FF), DFT, and MLIPs across benchmarks.

Benchmark	**Metric**	**Classical FF**	**DFT**
Small molecule minimisation	RMSD (Å)	0.2–0.5	$\leq$0.01–0.075
Non-covalent interactions	Absolute deviation from reference interaction energy (kcal/mol)	1.0–2.0	$\leq$0.2–1.0
	RMSE per interaction group (kcal/mol)	1.0–2.0	$\leq$1.0
Dihedral scan	Mean barrier error (kcal/mol)	$\leq$ 2.0	$\leq$0.5–1.0
Conformer selection	MAE (kcal/mol)	2.0–5.0	$\leq$0.5
	RMSE (kcal/mol)	3.0–6.0	$\leq$1.5
Tautomers	Absolute deviation (ΔG)	2.0–5.0	$\leq$0.05
Ring planarity	Deviation from plane (Å)	0.05–0.20	$\leq$0.01–0.05
Bond length distribution	Avg. fluctuation (Å)	0.03–0.08	$\leq$0.005–0.05
Reactivity-TST	Activation Energy (kcal/mol)	–	$\leq$2.0–3.0
	Enthalpy (kcal/mol)	-	$\leq$ 2.0
Reactivity-NEB	Final force convergence (eV/Å)	-	$\leq$0.05
Radial Distribution Function	RMSE (Å)	0.10–0.30	$\leq$0.1
Protein Sampling outliers	Ramachandran ratio	0.05–0.15	$\leq$0.005–0.1
	Rotamers ratio	0.10–0.25	$\leq$0.02–0.03
Protein Folding Stability	min(RMSD) (Å)	$\leq$3.0	$\leq$ 2.0
	max(TM-Score)	0.25–0.80	$\leq$0.5

C Supporting Figures and Tables

Table 4: Datasets used for the different benchmarks in MLIPAudit.

Benchmark	Dataset name	Link/Citation	Content description
Stability test	In-house dataset	released with MLIPAudit	PDBids: 1UAO, 1AB7, 1P79, 1BIP, 1A5E, 1A7M, 2BQV, 1J7H, 5KGZ, 1VSQ, 1JRS.
Inference scaling	In-house dataset	released with MLIPAudit	PDBids: 1AY3, 1UAO, 1AB7, 1P79, 1BIP, 1A5E, 1A7M, 2BQV, 1J7H, 5KGZ, 1VSQ, 1JRS.
Small molecule minimisation	OpenFF	[58]	100 molecules for the neutral dataset and 10 for the charged dataset. The subsets are constructed so that the chemical diversity, as represented by Morgan fingerprints, is maximised.
Non-covalent interactions	NCI-ATLAS subsets: D442x10, HB375x10, HB300SPXx10, IHB100x10, R739x5, SH250x10	http://www.nciatlas.org/	QM optimised geometries of distance scans of bi-molecular complexes, where the two molecules interact via non-covalent interactions with associated energies.
Dihedral scan	In-house recomputed TorsionNet 500 dataset at ωB97M-D3(BJ) DFT-level.	[59]	500 structures of drug-like molecules and their energy profiles around selected rotatable bonds at wB97M-D3(BJ) DFT-level.
Conformer selection	Wiggle 150	[47]	50 conformers each of three molecules: Adenosine, Benzylpenicillin, and Efavirenz.
Tautomers	In-house recomputed Tautobase dataset at ωB97M-D3(BJ) DFT-level.	[60]	2,792 tautomer pairs sourced from the Tautobase dataset. After generation of the structures and minimisation at xtb level, the QM energies were computed in-house using ωB97M-D3(BJ)/def2-TZVPPD level of theory.
Ring planarity	QM9 subset	[61]	One representative molecule each, containing substructures for benzene, furan, imidazole, purine, pyridine and pyrrole.
Bond length	QM9 subset	[61]	One representative molecule each, containing the bond types C-C, C=C, C#C, C-N, C-O, C=O and C-F.
Reactivity	Grambow dataset	[48]	Reactants, products and transition states of 11960 reactions.
Radial Distribution Function	different sources	[62–65]	Water, CCl4, Acetonitrile, Methanol.
Protein Folding	In-house dataset	released with MLIPAudit	PDBids: 1CQ0, 1UAO, 2JOF, 1BA6, 1E0L.
Protein Sampling	In-house dataset	released with MLIPAudit	ala-leu-glu-lys, gln-arg-asp-ala, glu-gly-ser-arg, gly-thr-trp-gly, gly-tyr-ala-val, met-ser-asn-gly, met-val-his-asn, pro-met-ile-gln, pro-met-phe-ala, ser-ala-cys-pro, trp-phe-gly-ala, val-glu-lys-ala.

Table 5: MLIPAudit test-cases overlap with models training dataset

Benchmark Category	Benchmark	Overlap [%]
Small-Molecule	Minimisation	0
Small-Molecule	Bond Length distribution	0
Small-Molecule	Ring Planarity	0
Small-Molecule	Conformer selection	66.7
Small-Molecule	Dihedral scan	1.4
Small-Molecule	Tautomers	8.4
Small-Molecule	Non-covalent interactions	–
Small-Molecule	Reactivity	–
Molecular liquids	RDF	0
Biomolecules	Folding stability	0
Biomolecules	Sampling	0

Table 6: Category-based rankings (aggregated scores by benchmark category)

Rank	Category	Model Name	Score	Benchmarks
1	General	MACE-SPICE2	0.900	1/1
2	General	NequIP-SPICE2	0.900	1/1
3	General	Visnet-SPICE2	0.810	1/1
5	General	Visnet-t1x	0.160	1/1
6	General	MACE-t1x	0.040	1/1
7	General	NequIP-t1x	0.040	1/1
1	Small-molecules	Visnet-SPICE2	0.578	9/9
2	Small-molecules	NequIP-SPICE2	0.550	9/9
3	Small-molecules	MACE-SPICE2	0.545	9/9
4	Small-molecules	NequIP-t1x	0.379	6/9
5	Small-molecules	MACE-t1x	0.351	6/9
6	Small-molecules	Visnet-t1x	0.306	6/9
1	Molecular-liquids	MACE-SPICE2	1.000	2/2
2	Molecular-liquids	Visnet-SPICE2	1.000	2/2
3	Molecular-liquids	NequIP-SPICE2	0.834	2/2
4	Molecular-liquids	Visnet-t1x	0.000	1/2
5	Molecular-liquids	MACE-t1x	0.000	1/2
6	Molecular-liquids	NequIP-t1x	0.000	1/2
1	Biomolecules	Visnet-SPICE2	0.727	2/2
2	Biomolecules	NequIP-SPICE2	0.584	2/2
3	Biomolecules	MACE-SPICE2	0.530	2/2
4	Biomolecules	Visnet-t1x	0.353	2/2
5	Biomolecules	MACE-t1x	0.284	2/2
6	Biomolecules	NequIP-t1x	0.183	2/2

Table 7: Single benchmarks rankings

Rank	Benchmark	Model Name	Score	Test Cases
1	General Stability	NequIP-SPICE2	0.900	10/9
2	General Stability	MACE-SPICE2	0.900	10/9
3	General Stability	Visnet-SPICE2	0.810	9/9
4	General Stability	MACE-SPICE2-t1x	0.300	10/9
5	General Stability	Visnet-t1x	0.160	4/9
6	General Stability	MACE-t1x	0.040	4/9
7	General Stability	NequIP-t1x	0.040	4/9
1	Molecular Liquids Solvents Rdf	MACE-SPICE2	1.000	3/3
2	Molecular Liquids Solvents Rdf	Visnet-SPICE2	1.000	3/3
3	Molecular Liquids Solvents Rdf	NequIP-SPICE2	0.669	3/3
1	Molecular Liquids Water Rdf	MACE-SPICE2	1.000	1/1
2	Molecular Liquids Water Rdf	NequIP-SPICE2	1.000	1/1
3	Molecular Liquids Water Rdf	Visnet-SPICE2	1.000	1/1
1	Protein Folding	MACE-t1x	0.525	3/5
2	Protein Folding	Visnet-t1x	0.525	3/5
3	Protein Folding	Visnet-SPICE2	0.525	5/5
4	Protein Folding	MACE-SPICE2	0.525	5/5
5	Protein Folding	NequIP-SPICE2	0.525	5/5
1	Protein Sampling	Visnet-SPICE2	0.928	12/12
2	Protein Sampling	NequIP-SPICE2	0.643	9/12
3	Protein Sampling	MACE-SPICE2	0.535	12/12
4	Protein Sampling	NequIP-t1x	0.366	9/12
5	Protein Sampling	Visnet-t1x	0.181	7/12
6	Protein Sampling	MACE-t1x	0.043	7/12

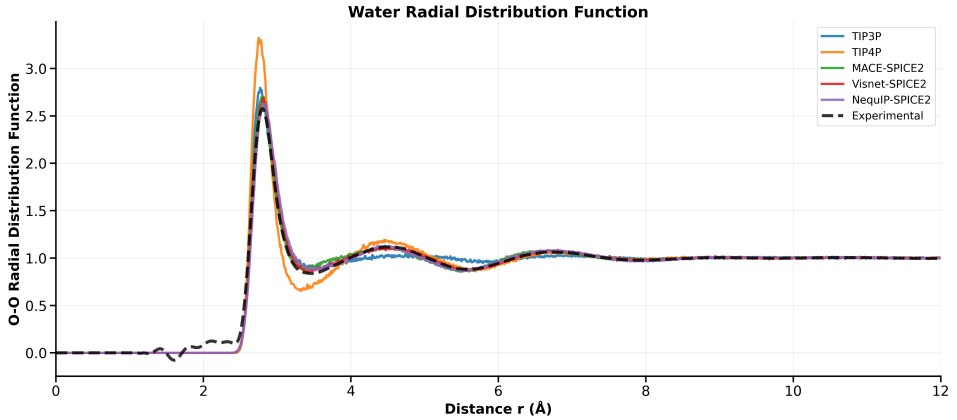

Figure 6: Water radial distribution function for MACE-SPICE2, NequIP-SPICE2 and Visnet-SPICE2, compared with the experimental observable and two water classical forcefields TIP3P and TIP4P [49]

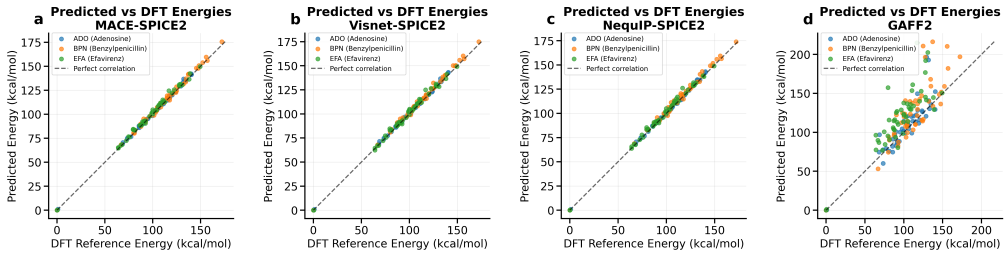

Figure 7: Predicted vs. DFT conformer energies for adenosine (ADO, blue), benzylpenicillin (BPN, orange), and efavirenz (EFA, green).

Table 8: Small-Molecules single benchmarks rankings

Rank	Benchmark	Model Name	Score	Test Cases
1	Small Molecule Bond Lenght	MACE-SPICE2	0.000	8/8
2	Small Molecule Bond Lenght	MACE-t1x	0.000	8/8
3	Small Molecule Bond Lenght	NequIP-t1x	0.000	8/8
4	Small Molecule Bond Lenght	NequIP-SPICE2	0.000	8/8
5	Small Molecule Bond Lenght	Visnet-SPICE2	0.000	8/8
6	Small Molecule Bond Lenght	Visnet-SPICE2-t1x-L	0.000	8/8
7	Small Molecule Bond Lenght	Visnet-SPICE2-t1x	0.000	8/8
8	Small Molecule Bond Lenght	MACE-SPICE2-t1x	0.000	8/8
9	Small Molecule Bond Lenght	Visnet-t1x	0.000	8/8
1	Small Molecule Conformer Selection	Visnet-SPICE2-t1x-L	0.748	3/3
2	Small Molecule Conformer Selection	MACE-SPICE2-t1x	0.672	3/3
3	Small Molecule Conformer Selection	MACE-SPICE2	0.471	3/3
4	Small Molecule Conformer Selection	Visnet-SPICE2	0.416	3/3
5	Small Molecule Conformer Selection	NequIP-SPICE2	0.390	3/3
6	Small Molecule Conformer Selection	Visnet-SPICE2-t1x	0.025	3/3
1	Small Molecule Dihedral Scan	Visnet-SPICE2	0.998	500/500
2	Small Molecule Dihedral Scan	Visnet-SPICE2-t1x-L	0.998	500/500
3	Small Molecule Dihedral Scan	NequIP-SPICE2	0.994	500/500
4	Small Molecule Dihedral Scan	MACE-SPICE2	0.987	500/500
5	Small Molecule Dihedral Scan	MACE-SPICE2-t1x	0.945	500/500
6	Small Molecule Dihedral Scan	Visnet-SPICE2-t1x	0.796	500/500
1	Small Molecule Noncovalent Interaction	MACE-SPICE2	0.525	1807/1807
2	Small Molecule Noncovalent Interaction	Visnet-SPICE2-t1x-L	0.519	1807/1807
3	Small Molecule Noncovalent Interaction	NequIP-SPICE2	0.515	1807/1807
4	Small Molecule Noncovalent Interaction	Visnet-SPICE2	0.514	1807/1807
5	Small Molecule Noncovalent Interaction	MACE-SPICE2-t1x	0.459	1807/1807
6	Small Molecule Noncovalent Interaction	Visnet-SPICE2-t1x	0.380	1807/1807
7	Small Molecule Noncovalent Interaction	NequIP-t1x	0.309	689/1807
8	Small Molecule Noncovalent Interaction	MACE-t1x	0.303	689/1807
9	Small Molecule Noncovalent Interaction	Visnet-t1x	0.152	689/1807
1	Small Molecule Ring Planarity	MACE-SPICE2	1.000	6/6
2	Small Molecule Ring Planarity	Visnet-SPICE2-t1x	1.000	6/6
3	Small Molecule Ring Planarity	MACE-SPICE2-t1x	1.000	6/6
4	Small Molecule Ring Planarity	Visnet-SPICE2	1.000	6/6
5	Small Molecule Ring Planarity	NequIP-SPICE2	1.000	6/6
6	Small Molecule Ring Planarity	Visnet-SPICE2-t1x-L	1.000	6/6
7	Small Molecule Ring Planarity	MACE-t1x	0.961	6/6
8	Small Molecule Ring Planarity	NequIP-t1x	0.938	6/6
9	Small Molecule Ring Planarity	Visnet-t1x	0.912	6/6
1	Small Molecule Rmsd	Visnet-SPICE2	1.000	220/220
2	Small Molecule Rmsd	Visnet-SPICE2-t1x-L	0.996	220/220
3	Small Molecule Rmsd	MACE-SPICE2	0.942	220/220
4	Small Molecule Rmsd	NequIP-SPICE2	0.760	220/220
5	Small Molecule Rmsd	MACE-SPICE2-t1x	0.464	220/220
6	Small Molecule Rmsd	Visnet-SPICE2-t1x	0.015	220/220
7	Small Molecule Rmsd	MACE-t1x	0.000	220/220
8	Small Molecule Rmsd	Visnet-t1x	0.000	220/220
9	Small Molecule Rmsd	NequIP-t1x	0.000	220/220
1	Small Molecule Tautomers	Visnet-SPICE2	0.927	1400/1400
2	Small Molecule Tautomers	NequIP-SPICE2	0.914	1400/1400
3	Small Molecule Tautomers	Visnet-SPICE2-t1x-L	0.911	1400/1400
4	Small Molecule Tautomers	MACE-SPICE2	0.644	1400/1400
5	Small Molecule Tautomers	MACE-SPICE2-t1x	0.537	1400/1400
6	Small Molecule Tautomers	Visnet-SPICE2-t1x	0.357	1400/1400