Universal Machine Learning Interatomic Potentials Enable Accurate Metal–Organic Framework Molecular Modeling

Hendrik Kraß[1,2] **Ju Huang**[1] **Seyed Mohamad Moosavi**[1,3]
[1]Chemical Engineering & Applied Chemistry, University of Toronto, Toronto, Canada
[2]Department of Computer Science, University of Tübingen, Tübingen, Germany
[3]Vector Institute for Artificial Intelligence, Toronto, Canada
mohamad.moosavi@utoronto.ca

Abstract

Universal machine learning interatomic potentials (uMLIPs) have emerged as powerful tools for accelerating atomistic simulations, offering scalable and efficient modeling with accuracy close to quantum calculations. However, their reliability and effectiveness in practical, real-world applications remain an open question. Metal-organic frameworks (MOFs) and related nanoporous materials are highly porous crystals with critical relevance in carbon capture, energy storage, and catalysis applications. Modeling nanoporous materials presents distinct challenges for uMLIPs due to their diverse chemistry, structural complexity, including porosity and coordination bonds, and the absence from existing training datasets. Here, we introduce MOFSimBench, a benchmark to evaluate uMLIPs on key materials modeling tasks for nanoporous materials. Evaluating 20 models from various architectures on a chemically and structurally diverse materials set, we find that top-performing uMLIPs consistently outperform classical force fields and fine-tuned machine learning potentials across structural optimizations, molecular dynamics simulations, and bulk modulus and heat capacity predictions. Our modular and extensible benchmarking framework is available at `https://github.com/AI4ChemS/mofsim-bench`.

1 Introduction

Molecular modeling is a powerful tool for understanding the structure and interactions of molecules and materials, and it plays a crucial role in predicting material properties to accelerate discovery [Frenkel and Smit, 2023, Tuckerman, 2023, Noé et al., 2020]. However, computing interatomic interactions in a molecular simulation is bound to an accuracy-efficiency trade-off: *Ab initio* quantum methods offer high accuracy, but are computationally expensive and do not scale well to large systems; in contrast, empirical potentials are computationally efficient, but often lack sufficient accuracy for practical applicability [Musil et al., 2021, Huang et al., 2023, Xie et al., 2023]. Machine learning interatomic potentials (MLIPs) have emerged as a promising tool to bridge this gap between efficiency and accuracy [Behler and Parrinello, 2007, Deringer and Csányi, 2017, Batatia et al., 2023, Lysogorskiy et al., 2021, Xie et al., 2023, Schütt et al., 2017, Batzner et al., 2022, Liao et al., 2024, Fu et al., 2025]. To enable broad applicability across a wide range of elements, universal MLIPs (uMLIPs) have been introduced to benefit from the scale of data and models, enabling the general deployment of MLIPs in many domains [Chen and Ong, 2022, Deng et al., 2023, Batatia et al., 2024, Barroso-Luque et al., 2024, Yang et al., 2024, Neumann et al., 2024, Rhodes et al., 2025, Bochkarev et al., 2024].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat NeurIPS 2025.

While the promise of "universal" interatomic potentials is compelling, their real-world deployment often encounters additional challenges, such as out-of-distribution (OOD) materials and tasks, which require rigorous testing of the models. Existing open benchmarks and evaluation frameworks, such as Matbench Discovery [Riebesell et al., 2025], often evaluate metrics closely aligned with the training objectives (e.g., energy, force, and stress fitting), which may not adequately reflect model performance on domain-relevant tasks or OOD scenarios [Focassio et al., 2025]. Recent efforts have begun incorporating downstream assessments, such as molecular dynamics (MD) stability and prediction of materials properties [Póta et al., 2024, Loew et al., 2024, Fu et al., 2023, Chiang et al., 2025], which may offer more reliable indicators of practical modeling performance.

This study focuses on applications of uMLIPs for metal-organic frameworks (MOFs), a class of highly porous materials with applications in carbon capture [Lin et al., 2021, Ye et al., 2025, Boyd et al., 2019, Chen et al., 2025] energy storage [Chen et al., 2024, Shin et al., 2023, Gittins et al., 2024], and catalysis [Mourino et al., 2025, Yu et al., 2025, Fumanal et al., 2020], as well as related nanoporous materials. MOFs are formed by self-assembly of metal nodes and organic linkers, resulting in modular porous frameworks with infinite design space [Moosavi et al., 2020]. Their low symmetry and large unit cells make *ab initio* quantum calculations prohibitively expensive for large-scale or long-timescale dynamic simulations. As a result, large-scale molecular modeling of MOFs has traditionally relied on classical force fields [Evans et al., 2016, Islamov et al., 2023, Moosavi et al., 2018]. While specialized force fields have been developed for a few prototypical frameworks, these force fields still fail to cover the diverse chemistries of MOFs as they retain a rigid functional form and rely on fixed parameters for specific coordination geometries [Boyd et al., 2017].

uMLIPs provide an opportunity to address these shortcomings by offering quantum-level accuracy at a computational cost closer to classical force fields. To assess their applicability in MOF molecular modeling, we investigate their performance on a variety of MOFs in structural optimization, MD, and derived properties, including elastic moduli and heat capacities. In addition, we compare them to fine-tuned uMLIPs specifically on MOF data [Elena et al., 2025] and the Universal Force Field (UFF) parametrized for MOFs (UFF4MOF) [Rappe et al., 1992, Coupry et al., 2016].

2 Benchmarking results

Our benchmark includes 9 state-of-the-art uMLIPs, spanning a range of architectural designs and training datasets. The selected architectures cover equivariant graph neural networks (GNNs) (MACE [Batatia et al., 2024, 2023], MatterSim [Yang et al., 2024], SevenNet [Park et al., 2024]), graph transformers (GTs) (EquiformerV2 (eqV2) [Liao et al., 2024, Barroso-Luque et al., 2024], eSEN [Fu et al., 2025]), graph network-based simulators (GNS) (orb-v2 [Neumann et al., 2024], orb-v3 [Rhodes et al., 2025]), and the graph basis function-based GRACE [Bochkarev et al., 2024]. All models, except MatterSim [Yang et al., 2024], which was trained on a proprietary dataset, were trained on large-scale open datasets, including MPtrj [Deng et al., 2023], Alexandria [Schmidt et al., 2023], and/or OMat24 [Barroso-Luque et al., 2024]. While MPtrj and Alexandria primarily contain equilibrium conformations, OMat24 includes extensive out-of-equilibrium data, enabling broader coverage of the potential energy surface (PES). We furthermore investigate the performance of MACE-MP-MOF0 [Elena et al., 2025], a uMLIP fine-tuned on MOF data, to study the effects of domain-specific training, as well as UFF4MOF [Coupry et al., 2016].

Each model is evaluated on a consistent set of tasks critical for characterizing nanoporous materials: structural optimization, molecular dynamics stability, and bulk property prediction, namely bulk modulus and the specific heat capacity. Tasks are performed on a set of 100 nanoporous structures curated from QMOF [Rosen et al., 2021], MOSAEC-DB [Gibaldi et al., 2025], IZA [Baerlocher et al.], and CURATED-COF [Ongari et al., 2019]. Performance is assessed by comparing uMLIP predictions against DFT references. Structures and DFT references for the heat capacity analysis were obtained from Moosavi et al. [2022].

2.1 Structural optimization

Structural optimization is a foundational task in atomistic modeling used to identify minimum-energy configurations, assess local stability, and reduce artifacts introduced by experimentally determined or hypothetical initial structures. We investigate the number of steps required to reach a force convergence criterion of 10^{-3} eV/Å or at most $5\,000$ optimization steps, as well as the

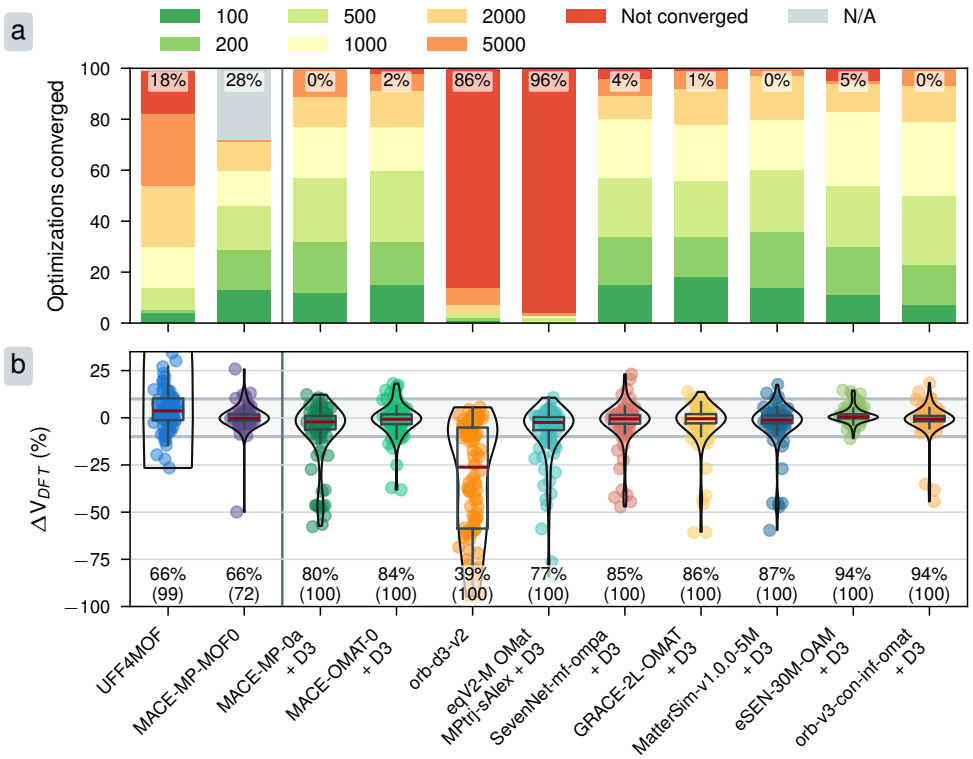

Figure 1: a) Structural minimization convergence grouped by number of optimization steps required to reach a force convergence criterion of 10^{-3} eV/Å. Numbers at the top indicate the percentage of N/A results and unconverged structural optimizations. b) Relative difference between uMLIP-optimized and DFT-optimized cell volumes. Numbers below violin plots denote the percentage of successfully computed structures with volume deviations of less than 10% to the DFT reference and the number of successfully computed structures.

relative volume difference to DFT-optimized structures in Figure 2. UFF4MOF struggles with slow convergence and inaccurate volumes. The fine-tuned MACE-MP-MOF0 is unable to compute 28 % of structures due to unsupported atom types, but captures volumes on the remaining structures accurately. Two non-conservative models, orb-d3-v2 and eqV2-M-OMsA, show large fractions of unconverged optimizations due to direct predictions of forces as an output head. This results in small oscillations in the forces that prevent stable convergence. In comparison, all other models compute forces as energy derivatives, resulting in fast and stable convergence. eSEN-OAM and orb-v3-con-inf show the best convergence and volumes closest to DFT. While most models exhibit outliers, no structure fails for all models, and no element is predominantly present in these structures (see Figure S3).

2.2 Molecular dynamics stability

Evaluating machine learning potentials solely on static samples at 0 K for energy and force prediction is insufficient to establish their usability in molecular dynamics simulations (MD) [Fu et al., 2023]. These simulations offer insight into the thermodynamics-dependent behavior of materials, and it is crucial that uMLIPs can perform them robustly. Each uMLIP simulates the structures over 50 ps in the *NpT* ensemble at 300 K and 1 bar. Comparing initial and final simulation volume, we find that simulation stability is closely related to optimization stability (see Figure S4), with a few structures showing significant volume deviations, but a majority completing the simulation with deviations of less than ± 10 %. However, overall stability is slightly higher. In simulations, the energy gradient is not as strictly followed as in structural optimizations, allowing potentials to recover to stable configurations more easily.

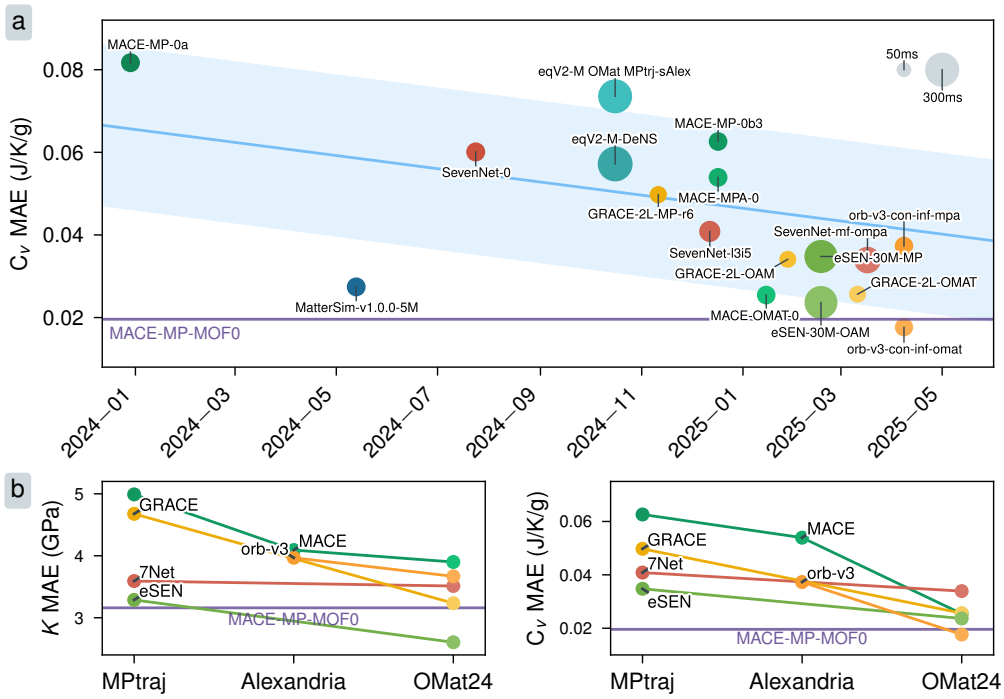

Figure 2: a) Evolution of heat capacity C_V mean absolute error between January 2024 and June 2025, showing a substantial decline in errors. The size of the points indicates the inference time of a single structural optimization step on MOF-5 (424 atoms) with the FIRE optimizer. b) Bulk modulus K and heat capacity C_V prediction mean absolute errors for different architectures with different largest datasets used (MPtrj < Alexandria < OMat24). All model architectures exhibit decreased errors with larger, more diverse training data.

2.3 Model performance evolution

The development of new MLIP architectures and datasets has led to rapid iterations of models and substantial performance improvements over a short period of time. To investigate this progress, we shift the focus to bulk property predictions, which complement the prior structural and dynamic analysis. Figure 2a displays the decrease in errors for heat capacity C_V predictions over 18 months. A strong linear trend shows errors decreasing over 75 % over this period, with a significant drop occurring in January 2025, two months after the release of the OMat24 dataset [Barroso-Luque et al., 2024]. The adoption of this large-scale dataset featuring non-equilibrium structures showcases the importance of diverse training data to mitigate issues such as PES softening [Deng et al., 2025] and its impact on improved accuracy. We further show this dataset-driven performance improvement in Figure 2b by comparing bulk modulus K and heat capacity C_V with different largest training datasets on the same architectures. For both properties, every architecture benefits from OMat24 training data, decreasing the error to the DFT references, and in two cases surpassing the fine-tuned MACE-MP-MOF0. This marks a significant milestone and further emphasizes the importance of carefully designed training data.

3 Discussion

We introduced MOFSimBench, an open and modular benchmarking framework for uMLIPs in nanoporous materials modeling. Our benchmark shows that recent uMLIPs show strong performance across the investigated tasks and surpass fine-tuned baselines as well as UFF/UFF4MOF, indicating their readiness for deployment. Non-conservative architectures trail conservative models in performance and show insufficient stability in structural optimization and MD simulations, highlighting the importance of exact gradient-based forces. Other architectural design decisions impact model performance and data efficiency. However, the size and diversity of the training dataset have the

largest impact on model performance, as out-of-equilibrium samples increase the stability of uMLIPs in advanced tasks. While uMLIPs show promising advances, limitations in the underlying DFT references must be acknowledged, and developing models beyond the PBE level of theory will be an important future step. Overall, the rapid advances in uMLIPs make them a crucial and accurate tool in MOF molecular modeling.

References

Ch. Baerlocher, Darren Brouwer, Bernd Marler, and L.B. McCusker. Database of zeolite structures.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, and Zachary W. Ulissi. Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models, October 2024. arXiv:2410.12771 [cond-mat, physics:physics].

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, 2023.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, March 2024. arXiv:2401.00096 [cond-mat, physics:physics].

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

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14), April 2007.

Anton Bochkarev, Yury Lysogorskiy, and Ralf Drautz. Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing. *Phys. Rev. X*, 14:021036, Jun 2024.

Peter G Boyd, Seyed Mohamad Moosavi, Matthew Witman, and Berend Smit. Force-Field Prediction of Materials Properties in Metal-Organic Frameworks. *The Journal of Physical Chemistry Letters*, 8(2):357–363, January 2017.

Peter G Boyd, Arunraj Chidambaram, Enrique García-Díez, Christopher P Ireland, Thomas D Daff, Richard Bounds, Andrzej Gładysiak, Pascal Schouwink, Seyed Mohamad Moosavi, M Mercedes Maroto-Valer, et al. Data-driven design of metal–organic frameworks for wet flue gas co2 capture. *Nature*, 576(7786):253–256, 2019.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, November 2022.

Tingting Chen, Hengyue Xu, Shaopeng Li, Jiaqi Zhang, Zhicheng Tan, Long Chen, Yiwang Chen, Zhongjie Huang, and Huan Pang. Tailoring the electrochemical responses of mof-74 via dual-defect engineering for superior energy storage. *Advanced Materials*, 36(31):2402234, 2024.

Xu Chen, Dhruv Menon, Xiaoliang Wang, Meng He, Mohammad Reza Alizadeh Kiapi, Mehrdad Asgari, Yuexi Lyu, Xianhui Tang, Luke L Keenan, William Shepard, et al. Flexibility-frustrated porosity for enhanced selective co2 adsorption in an ultramicroporous metal-organic framework. *Chem*, 2025.

Yuan Chiang, Tobias Kreiman, Elizabeth Weaver, Matthew Kuner, Christine Zhang, Aaron Kaplan, Daryl Chrzan, Samuel M Blau, Aditi S. Krishnapriyan, and Mark Asta. MLIP arena: Advancing fairness and transparency in machine learning interatomic potentials through an open and accessible benchmark platform. In *AI for Accelerated Materials Design - ICLR 2025*, 2025.

Damien E. Coupry, Matthew A. Addicoat, and Thomas Heine. Extension of the universal force field for metal–organic frameworks. *Journal of Chemical Theory and Computation*, 12(10):5215–5225, September 2016.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, September 2023.

Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A. Persson, and Gerbrand Ceder. Systematic softening in universal machine learning interatomic potentials. *npj Computational Materials*, 11(1), January 2025.

Volker L. Deringer and Gábor Csányi. Machine learning based interatomic potential for amorphous carbon. *Physical Review B*, 95(9), March 2017.

Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, January 2019.

Alin Marin Elena, Prathami Divakar Kamath, Théo Jaffrelot Inizan, Andrew S. Rosen, Federica Zanca, and Kristin A. Persson. Machine learned potential for high-throughput phonon calculations of metal—organic frameworks. *npj Computational Materials*, 11(1), May 2025.

Jack D Evans, Guillaume Fraux, Romain Gaillac, Daniela Kohen, Fabien Trousselet, Jean-Mathieu Vanson, and François-Xavier Coudert. Computational chemistry methods for nanoporous materials. *Chemistry of Materials*, 29(1):199–212, 2016.

Bruno Focassio, Luis Paulo M. Freitas, and Gabriel R. Schleder. Performance Assessment of Universal Machine Learning Interatomic Potentials: Challenges and Directions for Materials' Surfaces. *ACS Applied Materials & Interfaces*, 17(9):13111–13121, March 2025.

Filip Formalik, Michael Fischer, Justyna Rogacka, Lucyna Firlej, and Bogdan Kuchta. Benchmarking of GGA density functionals for modeling structures of nanoporous, rigid and flexible MOFs. *The Journal of Chemical Physics*, 149(6), August 2018.

Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2023.

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 Research*, 2023. Survey Certification.

Xiang Fu, Brandon M. Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction, 2025.

Maria Fumanal, Andres Ortega-Guerrero, Kevin Maik Jablonka, Berend Smit, and Ivano Tavernelli. Charge separation and charge carrier mobility in photocatalytic metal-organic frameworks. *Advanced Functional Materials*, 30(49):2003792, 2020.

Marco Gibaldi, Anna Kapeliukha, Andrew White, Jun Luo, Robert Alex Mayo, Jake Burner, and Tom K. Woo. MOSAEC-DB: a comprehensive database of experimental metal–organic frameworks with verified chemical accuracy suitable for molecular simulations. *Chemical Science*, 16(9): 4085–4100, 2025.

Jamie W Gittins, Kangkang Ge, Chloe J Balhatchet, Pierre-Louis Taberna, Patrice Simon, and Alexander C Forse. Understanding electrolyte ion size effects on the performance of conducting metal–organic framework supercapacitors. *Journal of the American Chemical Society*, 146(18): 12473–12484, 2024.

Stefan Goedecker, Michael Teter, and Jürg Hutter. Separable dual-space gaussian pseudopotentials. *Physical Review B*, 54(3):1703, 1996.

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

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

Bing Huang, Guido Falk von Rudorff, and O Anatole von Lilienfeld. The central role of density functional theory in the ai age. *Science*, 381(6654):170–175, 2023.

Meiirbek Islamov, Hasan Babaei, Ryther Anderson, Kutay B Sezginel, Jeffrey R Long, Alan JH McGaughey, Diego A Gomez-Gualdron, and Christopher E Wilmer. High-throughput screening of hypothetical metal-organic frameworks for thermal conductivity. *npj Computational Materials*, 9 (1):11, 2023.

Xin Jin, Kevin Jablonka, Elias Moubarak, Yutao Li, and Berend Smit. MOFChecker: An algorithm for Validating and Correcting Metal-Organic Framework (MOF) Structures, February 2025.

Thomas D Kühne, Marcella Iannuzzi, Mauro Del Ben, Vladimir V Rybkin, Patrick Seewald, Frederick Stein, Teodoro Laino, Rustam Z Khaliullin, Ole Schütt, Florian Schiffmann, et al. Cp2k: An electronic structure and molecular dynamics software package-quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics*, 152(19), 2020.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations, 2024.

Jian-Bin Lin, Tai TT Nguyen, Ramanathan Vaidhyanathan, Jake Burner, Jared M Taylor, Hana Durekova, Farid Akhtar, Roger K Mah, Omid Ghaffari-Nik, Stefan Marx, et al. A scalable metal-organic framework as a durable physisorbent for carbon dioxide capture. *Science*, 374(6574): 1464–1469, 2021.

Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Universal machine learning interatomic potentials are ready for phonons, 2024.

Yury Lysogorskiy, Cas van der Oord, Anton Bochkarev, Sarath Menon, Matteo Rinaldi, Thomas Hammerschmidt, Matous Mrovec, Aidan Thompson, Gábor Csányi, Christoph Ortner, and Ralf Drautz. Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon. *npj Computational Materials*, 7(1):1–12, June 2021.

Seyed Mohamad Moosavi, Peter G Boyd, Lev Sarkisov, and Berend Smit. Improving the mechanical stability of metal–organic frameworks using chemical caryatids. *ACS Central Science*, 2018.

Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G Boyd, Yongjin Lee, Berend Smit, and Heather Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11:4068, 2020.

Seyed Mohamad Moosavi, Balázs Álmos Novotny, Daniele Ongari, Elias Moubarak, Mehrdad Asgari, ´Ozge Kadioglu, Charithea Charalambous, Andres Ortega-Guerrero, Amir H. Farmahini, Lev Sarkisov, Susana Garcia, Frank Noé, and Berend Smit. A data-science approach to predict the heat capacity of nanoporous materials. *Nature Materials*, 21(12):1419–1425, December 2022.

Beatriz Mourino, Sauradeep Majumdar, Xin Jin, Fergus McIlwaine, Joren Van Herck, Andres Ortega-Guerrero, Susana Garcia, and Berend Smit. Exploring the chemical design space of metal–organic frameworks for photocatalysis. *Chemical Science*, 2025.

Felix 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.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential, 2024.

Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual Review of Physical Chemistry*, 71(1):361–390, April 2020.

Daniele Ongari, Aliaksandr V. Yakutovich, Leopold Talirz, and Berend Smit. Building a Consistent and Reproducible Database for Adsorption Evaluation in Covalent–Organic Frameworks. *ACS Central Science*, 5(10):1663–1675, October 2019.

Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *J. Chem. Theory Comput.*, 20(11):4857–4868, 2024.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, October 1996.

Giovanni Pizzi, Andrea Cepellotti, Riccardo Sabatini, Nicola Marzari, and Boris Kozinsky. Aiida: automated interactive infrastructure and database for computational science. *Computational Materials Science*, 111:218–230, 2016.

Balázs Póta, Paramvir Ahlawat, Gábor Csányi, and Michele Simoncelli. Thermal conductivity predictions with foundation atomistic models, 2024.

A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, 114(25):10024–10035, December 1992.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale, April 2025. arXiv:2504.06231.

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. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7(6):836–847, June 2025. ISSN 2522-5839.

Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin M. Notestein, and Randall Q. Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, May 2021.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks, 2020.

Lev Sarkisov, Rocio Bueno-Perez, Mythili Sutharson, and David Fairen-Jimenez. Materials Informatics with PoreBlazer v4.0 and the CSD MOF Database. *Chemistry of Materials*, 32(23):9849–9867, December 2020.

Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro J. M. A. Carriço, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. Machine-Learning-Assisted Determination of the Global Zero-Temperature Phase Diagram of Materials. *Advanced Materials*, 35(22): 2210788, June 2023.

K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 992–1002, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Seung-Jae Shin, Jamie W Gittins, Matthias J Golomb, Alexander C Forse, and Aron Walsh. Microscopic origin of electrochemical capacitance in metal–organic frameworks. *Journal of the American Chemical Society*, 145(26):14529–14538, 2023.

So Takamoto, Chikashi Shinagawa, Daisuke Motoki, Kosuke Nakago, Wenwen Li, Iori Kurata, Taku Watanabe, Yoshihiro Yayama, Hiroki Iriguchi, Yusuke Asano, Tasuku Onodera, Takafumi Ishii, Takao Kudo, Hideki Ono, Ryohto Sawada, Ryuichiro Ishitani, Marc Ong, Taiki Yamaguchi, Toshiki Kataoka, Akihide Hayashi, and Takeshi Ibuka. Pfp: Universal neural network potential for material discovery, 2021.

A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. LAMMPS - A flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, 2022.

Atsushi Togo. First-principles phonon calculations with phonopy and phono3py. *J. Phys. Soc. Jpn.*, 92(1):012001, 2023.

Atsushi Togo, Laurent Chaput, Terumasa Tadano, and Isao Tanaka. Implementation strategies in phonopy and phono3py. *J. Phys. Condens. Matter*, 35(35):353001, 2023.

Mark E Tuckerman. *Statistical mechanics: Theory and molecular simulation*. Oxford Graduate Texts. Oxford University Press, London, England, 2 edition, August 2023.

Joost VandeVondele, Matthias Krack, Fawzi Mohamed, Michele Parrinello, Thomas Chassaing, and Jürg Hutter. Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach. *Computer Physics Communications*, 167(2):103–128, 2005.

Stephen R. Xie, Matthias Rupp, and Richard G. Hennig. Ultra-fast interpretable machine-learning potentials. *npj Computational Materials*, 9(1), September 2023.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures, May 2024. arXiv:2405.04967 [cond-mat].

Zi-Ming Ye, Yi Xie, Kent O Kirlikovali, Shengchang Xiang, Omar K Farha, and Banglin Chen. Architecting metal–organic frameworks at molecular level toward direct air capture. *Journal of the American Chemical Society*, 2025.

Zhichao Yu, Zhenjin Xu, Ruijin Zeng, Man Xu, Minglang Zou, Da Huang, Zuquan Weng, and Dianping Tang. Tailored metal–organic framework-based nanozymes for enhanced enzyme-like catalysis. *Angewandte Chemie*, 137(7):e202420200, 2025.

A Supplementary Information

A.1 Computational details

All computations use the Atomic Simulation Environment (ASE) [Hjorth Larsen et al., 2017], which enables a reproducible pipeline for all uMLIPs. All model tasks are evaluated on Nvidia V100 32GB SXM, A100 80GB SXM, or H100 94GB SXM GPUs. Speed benchmarking is performed using a single H100 94GB SXM GPU.

A.1.1 Structure Curation

Our analysis focuses on MOFs but additionally includes COFs and zeolite structures, which have also been investigated for CO_2 capture applications. We curate structures from several sources. Our curated set is divided into two subsets:

1. Main structure set: Four MOF structures ubiquitous in the literature: IRMOF-1, IRMOF-10, UiO-66, and HKUST-1. In addition, we curated a set of 96 structures comprising 50 MOFs from MOSAEC-DB [Gibaldi et al., 2025], 23 MOFs from QMOF [Rosen et al., 2021], 10 zeolites, and 7 COFs from the CURATED-COF dataset [Ongari et al., 2019]. Structures from MOSAEC and QMOF were selected using stratified sampling across the largest pore diameter, computed with Poreblazer [Sarkisov et al., 2020], to feature diverse chemistries and metal variance. Zeolites and COFs were selected based on overlap with Ref. [Moosavi et al., 2022]. In addition, six well-known structures are included in this set: AlPMOF, AlPyrMOF, CALF-20, Co-MOF-74, Mg-MOF-74, and ZIF-8.
2. Heat capacity set: 231 MOF, COF, and zeolite structures from Ref. [Moosavi et al., 2022] with available DFT heat capacity. The heat capacity values were recomputed using the available phonopy files with settings matching this study.

To ensure high quality of the structures, we performed error checks with MOFChecker 2.0 [Jin et al., 2025] and only selected structures with no irregularities. Several structures were reduced to primitive unit cells to reduce the computational complexity of the reference DFT calculations.

Figures S1 and S2 show details on the distribution of element species, number of atoms, and total molecular weights.

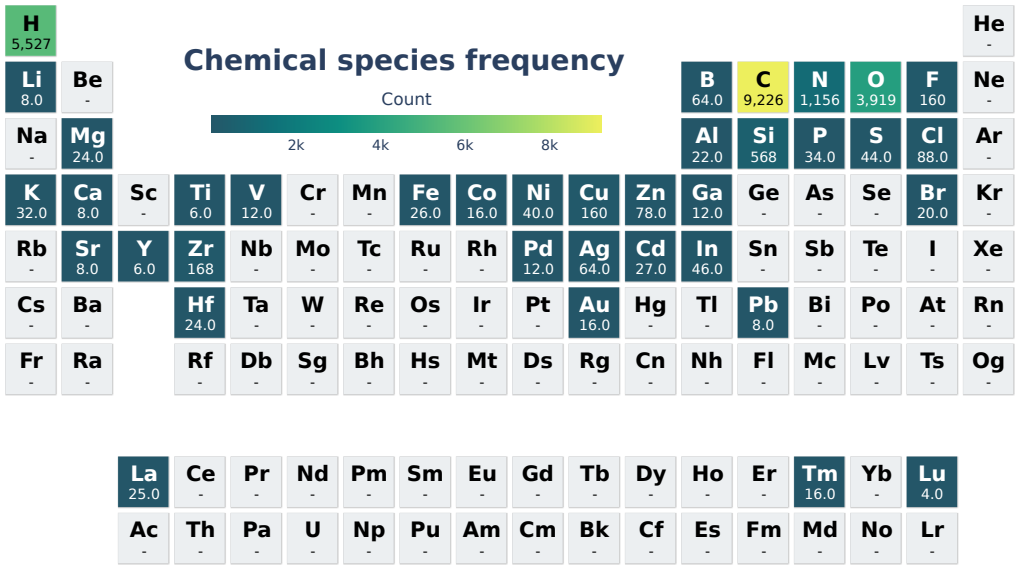

Fig. S1: Frequency of element species present in the main structure set of this study, encompassing 100 MOF, COF, and zeolite structures.

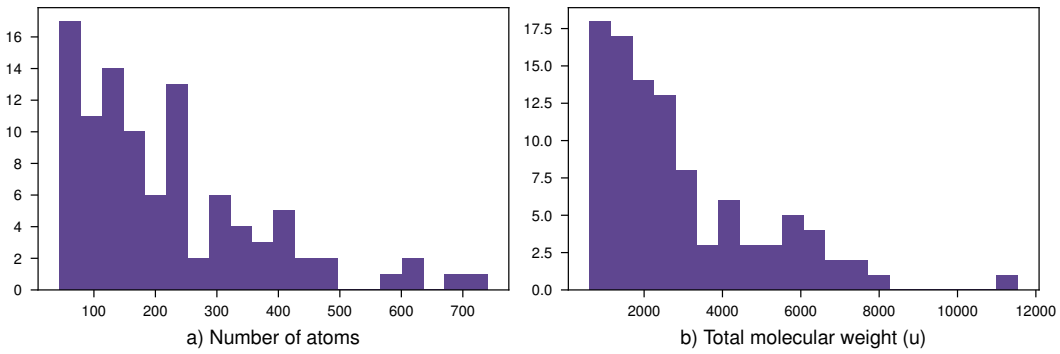

Fig. S2: a) Distribution of atom counts of structures in the main structure set. b) Distribution of total molecular weights of the structures in the main structure set.

A.1.2 Machine Learning Interatomic Potentials

We evaluate 19 uMLIPs and MACE-MP-MOF0:

1. MACE-MP-MOF0 [Elena et al., 2025]: A fine-tuned MACE-MP-0b (medium) [Batatia et al., 2024] model on a dataset of 127 MOFs and 4764 DFT calculations. In contrast to the original MACE-MP-0b model, it supports fewer elements and cannot compute all structures from our selection.
2. MACE-MP-0a* [Batatia et al., 2024]: The first universal MACE [Batatia et al., 2023] model trained on MPtrj [Deng et al., 2023].
3. MACE-MP-0b3 (medium)* [Batatia et al., 2024]: A MACE [Batatia et al., 2023] model with improved pair repulsion, correct isolated atoms, and better stability at high pressures. Trained only on MPtrj [Deng et al., 2023].
4. MACE-MPA-0 [Batatia et al., 2024]: A MACE [Batatia et al., 2023] model trained on the MPtrj [Deng et al., 2023] and sAlex [Schmidt et al., 2023, Barroso-Luque et al., 2024] dataset.
5. MACE-OMAT-0 [Batatia et al., 2024]: A MACE [Batatia et al., 2023] model trained on the OMat24 [Barroso-Luque et al., 2024] dataset.
6. MatterSim-v1 (5M) [Yang et al., 2024]: Based on the M3GNet architecture [Chen and Ong, 2022] that models three-body interactions using a graph-neural network approach. Its training uses an uncertainty-aware active-learning pipeline using model ensembles for uncertainty estimation.
7. orb-d3-v2 [Neumann et al., 2024]: A Graph Network-based Simulator (GNS) architecture [Sanchez-Gonzalez et al., 2020] trained on the MPtrj [Deng et al., 2023] and Alexandria [Schmidt et al., 2023] datasets. Notably, the training was performed on D3-corrected energies and forces, eliminating the need to compute dispersion corrections at inference time.
8. orb-mptraj-only-v2*: A GNS trained only on MPtrj [Deng et al., 2023]. In contrast to orb-d3-v2, this model does not predict D3-corrected outputs.
9. orb-v3-con-inf-omat [Rhodes et al., 2025]: The third-generation orb model with uncapped neighbor limit and conservative forces, trained on the OMat24 [Barroso-Luque et al., 2024] dataset.
10. orb-v3-con-inf-mpa [Rhodes et al., 2025]: The third-generation orb model with uncapped neighbor limit and conservative forces, trained on the MPTraj [Deng et al., 2023] and Alexandria [Schmidt et al., 2023] datasets.
11. eqV2-M OMat MPtrj-sAlex [Barroso-Luque et al., 2024]: A model based on the EquiformerV2 architecture [Liao et al., 2024], trained on the OMat24 dataset [Barroso-Luque et al., 2024], and fine-tuned on the MPtrj [Deng et al., 2023] and sAlex [Barroso-Luque et al., 2024, Schmidt et al., 2023] datasets.

12. eqV2-M-DeNS* [Barroso-Luque et al., 2024]: An EquiformerV2 [Liao et al., 2024] model that only uses the MPtrj training data.
13. eSEN-30M-OAM [Fu et al., 2025]: Based on the eSEN architecture which ensures smooth and expressive potential energy surfaces, trained on OMat24 [Barroso-Luque et al., 2024], MPtrj [Deng et al., 2023], and sAlex [Barroso-Luque et al., 2024, Schmidt et al., 2023].
14. eSEN-30M-MP* [Fu et al., 2025]: An eSEN model only trained on MPtrj [Deng et al., 2023].
15. GRACE-2L-MP (r6)* [Bochkarev et al., 2024]: The GRACE model extends the Atomic Cluster Expansion (ACE) [Drautz, 2019] to incorporate graph basis functions. This model was only trained on MPtrj[Deng et al., 2023].
16. GRACE-2L-OMAT [Bochkarev et al., 2024]: This GRACE model was only trained on OMat24 [Barroso-Luque et al., 2024].
17. GRACE-2L-OAM (r6) [Bochkarev et al., 2024]: This GRACE model was pre-fitted on OMat24 [Barroso-Luque et al., 2024] and fine-tuned on the sAlex [Barroso-Luque et al., 2024, Schmidt et al., 2023] and MPTraj[Deng et al., 2023] datasets.
18. SevenNet-0* [Park et al., 2024]: A model based on the NequIP architecture [Batzner et al., 2022] trained on MPtrj [Deng et al., 2023].
19. SevenNet-l3i5* [Park et al., 2024]: A SevenNet model with increased complexity trained on MPtrj [Deng et al., 2023].
20. SevenNet-ompa [Park et al., 2024]: A SevenNet trained on OMat24 [Barroso-Luque et al., 2024], sAlex [Barroso-Luque et al., 2024, Schmidt et al., 2023], and MPtrj [Deng et al., 2023], the selected output modal was `mpa` which resulted in better performance than `omat` except for heat capacity.

Matbench Discovery-compliant models are marked with an asterisk (*). While their performance is expected to lag behind models with broader training data, their inclusion enables direct architectural comparisons. Extended results using all model checkpoints are reported in the SI.

To the best of our knowledge, the training data sets that were used in the construction of these MLIPs do not contain MOFs structures, challenging them to make zero-shot predictions for this class of materials. The MACE-MP-MOF0 model is the only model that explicitly includes a training procedure on MOFs.

Dispersion Correction Dispersion corrections are critical for accurate modeling of MOFs [Formalik et al., 2018]. Therefore, all models either predict D3-corrected outputs (orb-d3-v2 and MACE-MP-MOF0) or the D3 correction is computed at inference time using the torch-dftd [Takamoto et al., 2021] package with `dispersion_xc=pbe`, `dispersion_cutoff=40` Bohr, `damping=bj`.

A.1.3 Structural minimization

Our energy minimization optimizes atom positions and cell parameters simultaneously and is performed using the `FrechetCellFilter` and the `LBFGS` optimizer. All structures are relaxed until a force convergence criterion of 10^{-3} eV/Å or a maximum of 5,000 optimizer steps is reached. UFF/UFF4MOF simulations were performed using the LAMMPS package [Thompson et al., 2022]. Force field parameters were applied by running `lammps-interface` [Boyd et al., 2017] without `--fix-metal` and with `--fix-metal`. The minimizer style in LAMMPS was set to `cg`, all other parameters and convergence settings were kept the same as those in the ASE [Hjorth Larsen et al., 2017] calculations.

A.1.4 Molecular Dynamics Simulations

All simulations use a time step of 1 fs and are performed with the ase IsotropicMKTNPT driver [Hjorth Larsen et al., 2017].

Simulations start with atom position minimization using the `LBFGS` optimizer until a force convergence criterion of 10^{-3} eV/Å or 1,000 optimizer steps is reached. Velocities are then initialized from a Maxwell-Boltzmann distribution at 300 K, and adjusted for zero center-of-mass momentum and

zero total angular momentum. The structures are equilibrated in an NVT ensemble with Langevin dynamics at 300 K using a friction coefficient of $0.01\,\mathrm{fs}^{-1}$ for 1 ps. Subsequent *NpT* simulations use `tdamp` $= 100$ fs, `pdamp` $= 1000$ fs, and an external stress of 1 bar.

In our stability simulations, the NVT-equilibrated structures are simulated in an *NpT* ensemble at 300 K for 50 ps.

A.1.5 Bulk Modulus

All structures were optimized with atom position and cell minimization using the `LBFGS` optimizer until a force convergence criterion of 10^{-3} eV/Å or 1,000 optimizer steps was reached. We apply a volumetric strain of $\pm 4\,\%$ in 11 evenly spaced steps. The resulting structures were optimized using the `FIRE` optimizer until a force convergence criterion of 10^{-3} eV/Å or 1,000 optimizer steps was reached. The bulk modulus was computed from a fitted Birch-Murnaghan equation of state. Structures for which the EOS volume minimum deviates more than 1% from the volume minimum of the initial optimization procedure are considered failed and excluded. The EOS fits of these structures were unstable in our experiments, leading to wrongly predicted bulk moduli, often as extreme outliers. Such filtering does not rely on ground-truth data and can be applied to reduce uncertainty in the predictions.

The DFT bulk moduli were calculated using the CP2K ver.9.1 package [Kühne et al., 2020]. Structures were fully optimized with respect to atomic positions and cells. Given the high computational cost of the DFT calculations, a volumetric strain of $\pm 4\,\%$ was applied in 5 evenly spaced steps. Structures with strains were optimized with respect to atomic positions. All calculations were performed through the Automated Interactive Infrastructure and Database for Computational Science, AiiDA [Pizzi et al., 2016], employing the *Cp2KMultistageWorkChain* workflow from the *aiida-lsmo* plugin.

The Quickstep code [VandeVondele et al., 2005] was used in the CP2K calculation. The Perdew-Burke-Enzerhof (PBE) exchange-correlation functional [Perdew et al., 1996] was employed along with DFT-D3(BJ) dispersion corrections [Grimme et al., 2010]. The GTH pseudopotentials[Goedecker et al., 1996], DZVP-MOLOPT-SR basis sets, and Gaussian plane wave were used. The cutoff energy of plane waves was set to 800 Ry. The energy and force convergences in the self-consistent filed were $1\mathrm{E}{-8}$ Ry and $0.00015\,\mathrm{bohr}^{-1} \times \mathrm{hartree}$, respectively. All other settings were set to the default defined in the *Cp2KMultistageWorkChain*. These settings have been demonstrated to be robust enough to find energetically stable configurations by comparing with QMOF relaxed structures.

A.1.6 Heat Capacity

All structures were first optimized with atom position and cell minimization using the `LBFGS` optimizer until a force convergence criterion of 10^{-3} eV/Å or 1,000 optimizer steps was reached. The heat capacity of the obtained structure was computed using Phonopy [Togo et al., 2023, Togo, 2023]. Force constants are computed using no supercells and the finite difference method with a distance of 0.01 Å. The mesh sampling phonon calculation is performed with `mesh = 100`. The heat capacity is extracted for $T = 300$ K.

A.2 Additional figures

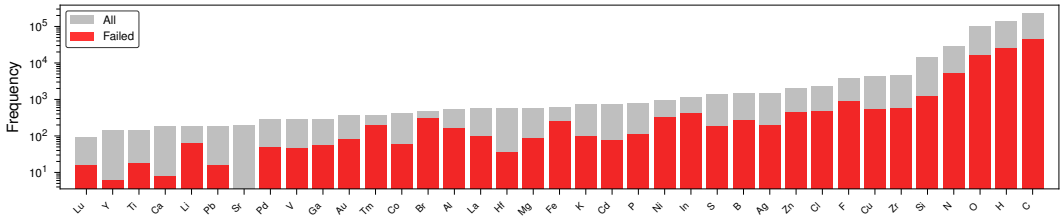

Fig. S3: **Histogram of elements present in outlier structures.** Number of elements present in structures with significant volume changes during structure minimization, shown with the total number of elements present in all structures.

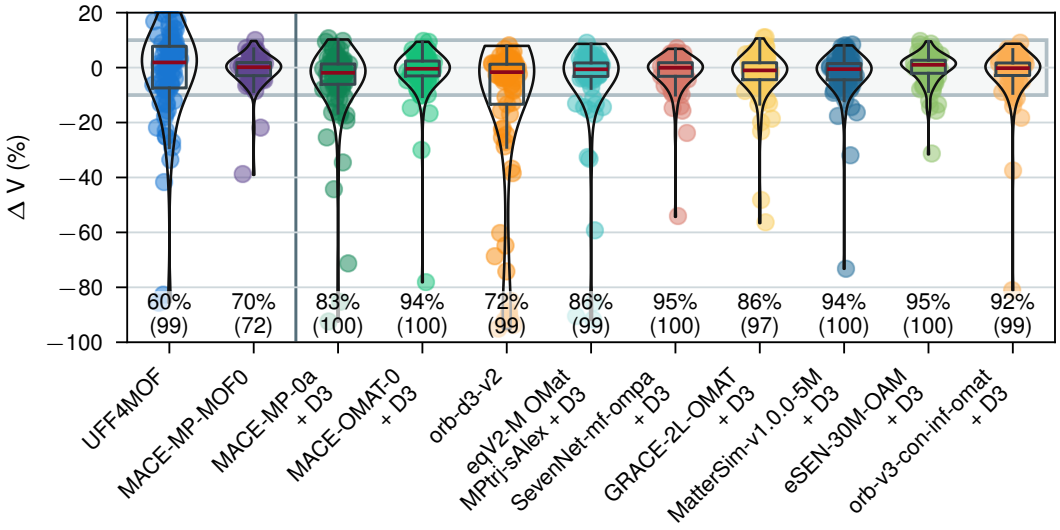

Fig. S4: Relative volume difference between the beginning and end of 50 ps long *NpT* simulations. Atom positions were first optimized, and the structures were equilibrated in the NVT ensemble. The grey shaded area indicates a threshold of ± 10 % for outliers. Numbers below the violin plots indicate the number of successfully computed structures with volume deviations of less than 10 % from the initial structure and the total number of successfully computed structures. GRACE-2L-OMAT encountered out-of-memory issues for three structures.