Benchmarking of Universal Machine Learning Interatomic Potentials for Structural Relaxation

Carmelo Gonzales
Intel Labs

Eric Fuemmeler
Department of Aerospace Engineering and Mechanics, University of Minnesota

Ellad Tadmor
Department of Aerospace Engineering and Mechanics, University of Minnesota

Stefano Martiniani
Center for Soft Matter Research, Department of Physics, New York University
Simons Center for Computational Physical Chemistry, Department of Chemistry, New York University
Courant Institute of Mathematical Sciences, New York University

Santiago Miret
Intel Labs

Abstract

The development of increasingly robust machine learning models for computational material science is escalating interest in integrating these models into real-world simulation workflows. Despite reporting strong model performance, the evaluation benchmarks typically only report a single error metric for the model's designated task. It is therefore difficult to predict how these models will perform in common workflows such as atomistic relaxations. Because of this, a more comprehensive set of testing benchmarks is needed to evaluate models performance on these dynamic tasks. A relaxation test is applied to three widely used models, namely: CHGNet, M3GNet, and MACE. The performance of these models showcase that although similar benchmark metrics are reported, models can exhibit significantly varied behavior in the relaxation test, even when trained on similar or identical datasets.

1 Introduction

Recent works have shown that evaluating the leading machine learning interatomic potential (MLIP) models solely on their performance of predicting energy and forces is not enough to provide insight into whether the models are usable for real-world tasks [1, 2]. It is therefore necessary to more thoroughly evaluate MLIPs with a wider range of tasks, tests, and evaluations to understand their capabilities for computational material science. Models trained to predict diverse material properties based on a given atomic structure can be useful as regression tools [3–6], however, those same models are often not practical for elementary downstream tasks, such as simulating the relaxation of a structure or predicting its elastic constants [1]. While some recent work [7–9] has claimed integration of MLIPs into real-world simulation workflows, much of that work remains difficult to verify and reproduce given the closed nature of the MLIPs and the associated benchmarks. Hence, there exists a

3rd Workshop for AI Accelerated Materials Design—NeurIPS 2024, Vancouver

clear need for an open-source, utilitarian benchmark to evaluate new and existing MLIPs based on a collection of elementary material science simulation tasks.

2 Background and Related Work

Duval et al. [10] provides a detailed review of how many MLIPs leverage geometric deep learning methods to encode concrete inductive biases based on geometric information. Many of the diverse set of architectures described in Duval et al. [10], however, have only been trained on energy and force prediction and often become unstable in materials simulation [1, 2]. While many deep learning architectures have been proposed for materials property prediction on various benchmarks [6, 5, 11], fewer have thoroughly evaluated the capabilities of MLIPs in real-world computational materials science tasks [1, 2]. While Bihani et al. [1] provides a valuable assessment of various deep learning models, the number of systems studied is limited compared to the broader availability of known materials systems [12, 13]. Some of the more promising MLIPs proposed [14, 7, 15, 16, 9] are geometric deep learning models trained on relaxation of bulk structure from the Materials Project (MP) [13, 14]. Given the importance of MP in training effective MLIPs, we leverage the framework in ColabFit [17] and OpenKIM [18] to create a testing benchmark for assessing the capabilities of MLIPs by performing atomistic relaxations based on crystals contained in MP.

3 Benchmarking Pretrained Models

3.1 Relaxation Test

OpenKIM provides a useful starting point for building out a benchmarking suite. It contains a robust set of tests for computing bulk, cluster, wall, line, and point properties of materials for >600 IPs. Nearly all IPs available on OpenKIM, however, are classical in nature, typically supporting just a few elements each. In addition, models must be made to be compatible with the KIM API [19], a standard developed to enhance portability of models across simulator platforms, *e.g.*, LAMMPS [20], ASE [21], DL_POLY [22], *etc*. Therefore, it is desirable to build upon this framework and provide an extensible, user-centric benchmarking toolkit for robust and reproducible testing of MLIPS.

As an initial proof of principle we focus on the relaxation of crystal structures using predicted MLIP atomic forces, with a particular emphasis on *convergence behavior*. We focus on this test as it is typically a dependency of further property calculations, and as such it is important to understand a model's performance here before moving to higher-level static and dynamic properties. The test itself is adapted from OpenKIM's EquilibriumCrystalStructure test driver [23] and uses the FIRE2 [24] optimization algorithm present in ASE [21] via a `Calculator` interface to perform symmetry-constrained relaxation of positional and cell degrees of freedom. Note, we do not attempt to fine-tune this procedure for each individual model. Rather, we select a general optimization framework that should behave reasonably well for all models considered here.

3.2 Experiment Setup

Three current and widely used models are chosen for evaluation, namely: CHGNet, M3GNet, and MACE. CHGNet from [14] is an invariant universal network trained on the Materials Project Trajectory Dataset, and is used to model the universal potential energy surface. M3GNet from [15] is a universal interatomic potential model with three-body interactions, trained on the collection of Materials Project structure relaxations. MACE from [9] is an equivariant message passing neural network model for interatomic potentials trained on the Materials Project Trajectory Dataset.

Table 1: Reported force validation error for the three pretrained models. Taken from Ref. 25

Model	Force MAE (eV/A)
CHGNet	0.063
M3GNet	0.075
MACE	0.057

For the following evaluations in 3.4, three different minimum force convergence thresholds for the structure relaxation, and a maximum number of 500 steps to reach the threshold are used. Convergence rates for a range of step numbers for each model at three different force convergence thresholds are reported in tables Table 2, Table 3, and Table 4. The model evaluations are run using double precision and typically using 4 CPU's with 12GB of memory available. Due to the large computational overhead of running over 10,000 evaluations for three models, some tests either do not get scheduled or fail to run. Only the set of structures that have test results for each model are presented. Furthermore, models which fail at some point during the test due to out of memory errors, computation errors in the simulation packages, or other convergence errors (typically large deformation gradient steps), are counted as failed runs. The percent of tests which converged, percent of tests which reached the max iteration limit, and percent of tests which fail are reported.

3.3 Benchmarking Data

All benchmarking data is taken from Materials Project, with every structure containing elastic properties downloaded, amounting to 10,773 structures at the time of download. This choice of structures was made as benchmarking of elastic properties will be conducted in the near future. It is important to note that the structures used to initialize relaxation are contained within the Materials Project Trajectory Dataset. However, other structures along the MLIP relaxation trajectories are unlikely to be *exactly* contained within the dataset. Similar analyses of MLIPs using data from MP have provided analogous "data similarity" disclaimers [9].

3.4 Results

A starting force convergence threshold of 0.1 eV/A with a maximum of 500 steps is used to first understand model performance in a low fidelity setting in which models can be expected to perform well based on their reported data. While not as helpful in practice, this initial benchmark is both faster to compute and insightful for quickly comparing ML models to each other. With this setup, a total of 4,937 total structures were evaluated by each model.

Despite both being trained on the MP trajectory dataset, MACE has far quicker convergence than CHGNet, and has a notably higher convergence after just one step of optimization. M3GNet and CHGNet both perform worse overall, with less test runs meeting the convergence threshold and taking more steps perform the relaxations.

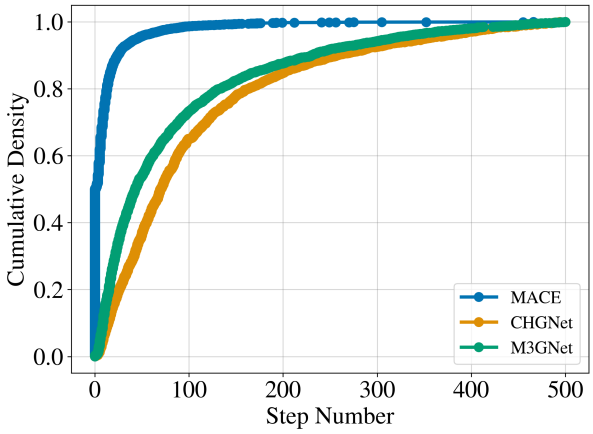


Figure 1: Cumulative density function using only the tests which reached the convergence threshold of 0.1 eV/A for each model. At this low force threshold, the model's performance is spread out, with MACE both relaxing the most structures successfully, and performing the relaxations in fewer steps.

An intermediate force threshold of 1e-3 eV/A is analyzed to get a sense of model performance in a more utilitarian regime. As machine learning and classical molecular dynamics simulation techniques begin to compliment each other, this convergence threshold represents a middle ground where the ML and classical techniques may be used together by first relaxing structures in the direction of a stable

Table 2: Model performance overview including percent of all tests which converged, reached the iteration limit, and failed. The cumulative density function section is built by only taking the tests which reached the convergence threshold of 0.1 eV/A. Notably, MACE is able to fully relax over 50% of the structures in only one step, and relaxes over 95% of all structures. CHGNet is only able to relax 40.9% of structures, and does so slower than MACE.

Model	%Converged	%Max Steps	%Failed	1	50	100	150	200	250	300	350	400	450	500
CHGNet	40.9%	4.7%	54.4%	0.001	0.372	0.650	0.780	0.848	0.894	0.926	0.949	0.971	0.989	1
M3GNet	62.6%	3.6%	33.8%	0.004	0.542	0.734	0.825	0.877	0.919	0.943	0.966	0.981	0.991	1
MACE	95.1%	0.1%	4.7%	0.502	0.958	0.987	0.994	0.998	0.999	0.999	0.999	1	1	1

configuration with the ML model, and then handing off the configuration to a classical model to finish out the simulation. With this threshold, a total of 10,700 structures are evaluated. MACE remains the most performant model, relaxing 84.6% of evaluated structures, while CHGNet and M3GNet both struggle to converge more than 26.2% and 38.5% of structures respectively.

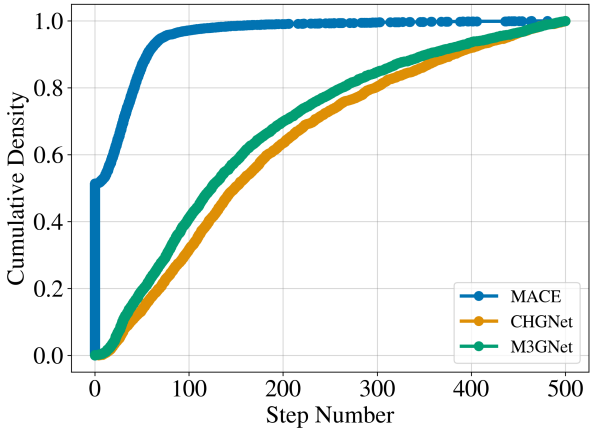
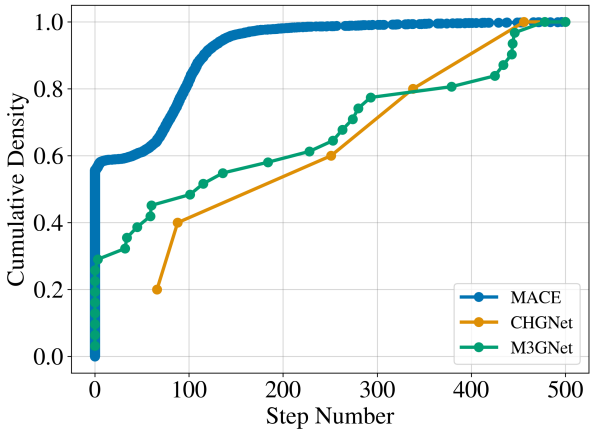


Figure 2: Only taking the tests which reached the convergence threshold of 1e-3 eV/A, the cumulative density function is shown across the range of allowed steps. The performance gap between models widens as M3GNet and CHGNet begin to relax fewer structures, and take more steps to do so. MACE still shows strong performance, and is able to relax 51.4% of structures after the first step.

Table 3: Model performance overview including percent of all tests which converged, reached the iteration limit, and failed. The cumulative density function section is built by only taking the tests which reached the convergence threshold of 1e-3 eV/A.

Model	%Converged	%Max Steps	%Failed	1	50	100	150	200	250	300	350	400	450	500
CHGNet	26.2%	13.3%	60.5%	0	0.139	0.319	0.506	0.636	0.733	0.804	0.864	0.920	0.964	1
M3GNet	38.5%	17.0%	44.4%	0.020	0.196	0.413	0.583	0.700	0.782	0.849	0.898	0.937	0.966	1
MACE	84.6%	2.0%	13.4%	0.514	0.871	0.972	0.986	0.991	0.993	0.996	0.997	0.998	0.999	1

A final force threshold of 1e-6 eV/A is used as a more practical benchmark value for the scenario where ML models are to be used to fully perform a relaxation. The maximum allowed steps is kept at 500, and all other hyper-parameters remain the same. In this setting, the differences in model performance becomes more distinct and provides deeper insight into where the top performing model, MACE, starts to struggle.A total of 10,665 structures are evaluated in all three models. For MACE, 55.8% of tests are still able to converge after the first step of optimization, however the convergence rate slows down as the model works to find the minimum force configuration. Similarly, the total number of converged tests for M3GNet and CHGNet are drastically reduced, with M3GNet able to relax 0.29% of structures, and CHGNet able to relax 0.05%.



(a)

Figure 3: The performance of M3GNet and CHGNet is drastically reduced, with both models relaxing less than 1% of structures. MACE is still able to relax 68.4% of structures, while still maintaining a convergence rate for relaxed structures of over 50% at step 1.

Table 4: Model performance overview including percent of all tests which converged, reached the iteration limit, and failed. The cumulative density function section is built by only taking the tests which reached the convergence threshold of 1e-6 eV/A. CHGNet and M3GNet performance is dramatically worse than the previous thresholds.

Model	%Converged	%Max Steps	%Failed	1	50	100	150	200	250	300	350	400	450	500
CHGNet	0.05%	27.5%	72.4%	0	0	0.400	0.400	0.400	0.400	0.600	0.800	0.800	0.800	1
M3GNet	0.29%	38.8%	60.9%	0.258	0.387	0.452	0.548	0.581	0.613	0.774	0.774	0.806	0.968	1
MACE	68.4%	11.7%	19.9%	0.558	0.612	0.829	0.963	0.981	0.987	0.991	0.995	0.997	0.998	1

To better understand the high convergence of structures in the first step for MACE, we relaxed a subset of the data (1000 structures) again, first perturbing the initial positions of each structure using ASE's `rattle` function (stdev=0.001, 0.1, 10, 1000). In all cases, including no perturbations to the same 1000 structures, the percentage of structures which converged after one step remained consistently near 37%. Despite all three models being trained on data that overlaps with the benchmarking structures, MACE appears to be much better fit near minima without being overfit to exactly the final relaxed structures.

4 Discussion

With an increase in robustness in ML model predictions for energy and forces, models begin to demonstrate the ability to serve as a supplement to downstream computational materials science tasks such as structure relaxation. Despite an increase in robustness, a static evaluation metric is not enough to quantify how well models may perform in practice, and thus a larger range of benchmark tasks must be employed to gather more performance data using these models. In this work, three models are evaluated against a relaxation test with varying levels of convergence fidelity. Comparing two models trained with the same dataset, CHGNet and MACE, the performance is drastically different, even at the lowest force convergence threshold. M3GNet, trained on a similar but not identical dataset, shows an intermediate level of performance despite having higher reported error metrics. The evaluation of MLIPs on the relaxation task reveals far more information about the model's capability than a static benchmarking number can show. This may lead to a deeper understanding of why certain models perform better, how best to perform training, and which hyper-parameters may be most important for learning to perform tasks which go beyond static predictions.

References

[1] Vaibhav Bihani, Sajid Mannan, Utkarsh Pratiush, Tao Du, Zhimin Chen, Santiago Miret, Matthieu Micoulaut, Morten M Smedskjaer, Sayan Ranu, and NM Anoop Krishnan. Egraffbench: evaluation of equivariant graph neural network force fields for atomistic simulations. *Digital Discovery*, 3(4):759–768, 2024.

[2] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi S Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*.

[3] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

[4] Santiago Miret, Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, and Matthew Spellings. The open matsci ML toolkit: A flexible framework for machine learning in materials science. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=QBMyDZsPMd.

[5] Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, Matthew Spellings, Mikhail Galkin, and Santiago Miret. Matsciml: A broad, multi-task benchmark for solid-state materials modeling. *arXiv preprint arXiv:2309.05934*, 2023.

[6] Janosh Riebesell, Rhys EA Goodall, Anubhav Jain, Philipp Benner, Kristin A Persson, and Alpha A Lee. Matbench discovery–an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920*, 2023.

[7] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

[8] Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

[9] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, William J. Baldwin, 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, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, 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, 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, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, 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. 2023.

[10] Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.

[11] Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.

[12] Mariette Hellenbrandt. The inorganic crystal structure database (icsd)—present and future. *Crystallography Reviews*, 10(1):17–22, 2004.

[13] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[14] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. Chgnet: Pretrained universal neural network potential for charge-informed atomistic modeling. 2023.

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

[16] Ilyes Batatia, David Peter Kovacs, Gregor N. C. Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=YPpSngE-ZU.

[17] Joshua A Vita, Eric G Fuemmeler, Amit Gupta, Gregory P Wolfe, Alexander Quanming Tao, Ryan S Elliott, Stefano Martiniani, and Ellad B Tadmor. Colabfit exchange: Open-access datasets for data-driven interatomic potentials. *The Journal of Chemical Physics*, 159(15), 2023.

[18] E. B. Tadmor, R. S. Elliott, J. P. Sethna, R. E. Miller, and C. A. Becker. The potential of atomistic simulations and the knowledgebase of interatomic models. *JOM*, 63(7):17–17, Jul 2011. ISSN 1543-1851.

[19] Ryan S. Elliott and Ellad B. Tadmor. Knowledgebase of Interatomic Models (KIM) application programming interface (API). https://openkim.org/kim-api, 2011.

[20] 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. *Comp. Phys. Comm.*, 271:108171, 2022. doi: 10.1016/j.cpc.2021.108171.

[21] 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, 2017. URL http://stacks.iop.org/0953-8984/29/i=27/a=273002.

[22] Ilian T. Todorov, William Smith, Kostya Trachenko, and Martin T. Dove. DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism. *J. Mater. Chem.*, 16: 1911–1918, 2006. doi: 10.1039/B517931A. URL http://dx.doi.org/10.1039/B517931A.

[23] I Nikiforov and Ellad B. Tadmor. Equilibrium structure and energy for a crystal structure at zero temperature and pressure v002. OpenKIM, https://doi.org/10.25950/2f2c4ad3, 2024.

[24] Julien Guénolé, Wolfram G. Nöhring, Aviral Vaid, Frédéric Houllé, Zhuocheng Xie, Aruna Prakash, and Erik Bitzek. Assessment and optimization of the fast inertial relaxation engine (fire) for energy minimization in atomistic simulations and its implementation in lammps. *Computational Materials Science*, 175:109584, 2020. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2020.109584. URL https://www.sciencedirect.com/science/article/pii/S0927025620300756.

[25] Matbench discovery. https://matbench-discovery.materialsproject.org. Accessed: 2024-09-01.