Benchmarking Machine Learning Potentials for Crystal Structure Relaxation

Kowen Woo[†,1,2,3,4] **Prashant Govindarajan**[1,2,3] **Sarath Chandar**[1,2,3]
[1]Chandar Research Lab [2]Mila – Quebec AI Institute
[3]Polytechnique Montréal [4]University of Alberta
Montreal, QC, Canada
{firstname.lastname}@mila.quebec

Abstract

High-throughput materials discovery workflows require rapid and accurate relaxation of crystal structures to identify thermodynamically stable phases among thousands to millions of candidate structures. We introduce a supplementary benchmark that evaluates state-of-the-art MLIPs and a one-shot relaxation model on structure relaxation with crystals generated via a reinforcement learning pipeline. We compare energy lowering and average maximum force computed via DFT, as well as relaxation runtime. We also contrast direct force-prediction strategies against conservative energy-differentiation approaches to determine which paradigm delivers superior relaxation performance. Our results indicate that there is a disconnect between MLIP energy prediction and force convergence in relaxation; however, we note that this is an ongoing study and further analysis is required.

1 Introduction

The discovery of new materials with desired properties presents an important challenge in modern materials science, with applications including energy storage, catalysis, and electronics. Central to these discovery pipelines is structure relaxation: the process of optimizing atomic positions and structural parameters to energetically stable configurations. For crystalline materials, property predictions are sensitive to their underlying structure, and having crystals in a stable, relaxed state is crucial [14]. High-throughput screening of materials routinely involves relaxation of thousands to millions of candidate structures, making the efficiency and accuracy of structure optimization a critical bottleneck in materials discovery pipelines. Crystal structure relaxation is generally performed with density functional theory (DFT), which provides a first-principles estimate of the energy and forces of a system. However, DFT's computational cost scales cubically with the system size[6], making it difficult to massively screen the vast chemical space in short time. This trade-off between accuracy and throughput has long restricted the acceleration of materials discovery with computers.

Machine learning interatomic potentials (MLIPs) have emerged as a promising solution to this computational bottleneck, offering the potential to achieve near-DFT accuracy at a fraction of the computational cost. MLIPs involve the mapping of atomic structures to the potential energy and forces, primarily via supervised learning architectures[5]. Recent advances in graph neural networks, equivariant architectures, and transformer-based models have demonstrated impressive performance on energy and force prediction tasks across diverse materials systems. MatBench Discovery[16] represents the most extensive benchmarking effort for MLIPs in stability prediction, evaluating both energies and structural relaxation. Our work provides a complementary perspective by evaluating MLIP performance on structures generated through RL-driven materials discovery pipelines, which may present different generalization challenges.

[†]Work done as part of an internship at Mila

Accepted at AI for Science workshop (NeurIPS 2025).

To address this gap, we evaluate the performance of state-of-the-art MLIPs on crystal structures generated through a reinforcement learning (RL) pipeline using CrystalGym[10]: CrystalGym is an open-source RL environment for crystal generation based on the Gymnasium framework for training policies with reward signals obtained directly from DFT. We evaluate 6 diverse architectures from MatBench Discovery including universal graph neural network potentials (M3GNet[3], CHGNet[4], MACE[2]), transformer-based approaches (EquiformerV2[13], eSEN-30M[8]), and direct structure prediction methods (DeepRelax[17]) on crystal structures obtained from trained CrystalGym RL policies. For evaluation, we compare the average formation energy difference and maximum atomic force obtained using DFT simulations. We also compare the differences in formation energies relative to the structure relaxed entirely with DFT. Overall, we aim to demonstrate that energy-based analyses can offer a more reliable measure of relaxation accuracy in addition to metrics such as RMSD, which quantify structural deviations from the reference relaxed state. We release the dataset of RL-generated crystals and their relaxed states obtained from DFT here: https://zenodo.org/records/16935035.

2 Background

2.1 MLIP Architectures

CHGNet[4] is a graph-based MLIP that integrates site-specific magnetic moments (as proxies for charge information) into its pretrained universal potential to capture both atomic positions and electronic orbital occupancy. MACE[2] enhances expressivity and efficiency using higher-body equivariant message passing, drastically reducing the number of layers needed while maintaining fast, accurate interatomic force predictions. M3GNet is a materials graph neural network that incorporates explicit three-body interactions, atomic coordinates, and full lattice tensors, enabling accurate tensorial predictions (forces and stresses) across the periodic table via auto-differentiation. Additionally, we include an iteration-free structure relaxation approach, DeepRelax, which is not an MLIP. DeepRelax is a generative model capable of directly predicting relaxed crystal structures without iterative energy minimization. Using a periodicity-aware equivariant GNN, it achieves 100-fold speed improvements over iterative models while maintaining competitive accuracy. CHGNet, MACE, and M3GNet have been trained on $\sim 1.3M$ structures from the MPTraj dataset, and are conservative models: atomic forces are obtained by differentiating the system energy. DeepRelax, in contrast, is trained on the X-MN-O dataset[12] derived from the MP database. EquiformerV2[13] presents a significant advancement in equivariant transformers, using eSCN convolutions to scale to higher degree representations along with attention re-normalization, separable S^2 activation, and separable layer normalization. eSEN[8] is a message-passing neural network that processes atomic structures through alternating edgewise and node-wise operations, with atoms embedded as multi-channel spherical harmonic representations, maintaining continuous representations throughout the network and significantly improving energy conservation. It is important to note that CHGnet, MACE, M3Gnet are conservative force prediction models while EquiformerV2 is a direct force prediction model and eSEN is pretrained on a direct force prediction model and fine-tuned with a conservative force prediction model.

2.2 Current Benchmarking Challenges

Evaluating ML models for materials applications presents unique challenges. Bartel et al.[1] demonstrated that accurate formation energy prediction does not necessarily translate to reliable stability predictions. In our work, we intend to evaluate whether the models optimized for accurate energy predictions can perform well in structural relaxation, particularly with new chemical systems. Together, we address a broader theme: that the utility of ML models in materials discovery depends not only on the energy estimation but also on their ability to capture the overall energy landscapes. Current MLIP benchmarks predominantly focus on energy-centric metrics, typically evaluating performance by comparing predicted energies to DFT reference calculations[7][11]. On the other hand, MatBench Discovery evaluates relaxation performance using RMSD between predicted and DFT-relaxed structures and compares the final relaxed energies against the DFT computed values for the WBM ground states. This is the current gold-standard for benchmarking MLIPs and our work aims to build upon it while exploring novel chemical spaces.

3 Methods

3.1 Experimental Setup and Dataset

We evaluate crystal structure relaxation performance across six state-of-the-art ML models. Five models (CHGNet, MACE, M3GNet, eSEN, and EquiformerV2 trained on the MPTraj dataset) utilize the Atomic Simulation Environment (ASE) library for structure optimization, while DeepRelax employs a direct structure prediction approach without iterative optimization. For ASE-based models, we maintained consistent use of the FIRE optimizer to ensure fair comparison as shown in the Appendix B.3. For DFT calculations, we use Quantum Espresso v7.3 [9], an open-source software suite for atomic simulations. We generated crystals by performing rollouts using different RL policies trained on the CrystalGym environment[10]. The original set of tasks of CrystalGym was to train policies to design crystal compositions on known crystal structures (from Materials Project) and optimize properties such as bulk modulus, band gap, and density. As structure optimization is not included as part of the original CrystalGym pipeline, the generated crystals are not relaxed. By choosing structures with unique compositions, we obtain around 1000 crystals. The distribution of formation energies of these crystals is shown in the Appendix 3. The presence of crystals with positive formation energies indicates that many of them are thermodynamically unfavorable.

3.2 Evaluation

We assess model performance using three complementary metrics that capture different aspects of relaxation quality and computational efficiency.

Formation energy reduction measures the difference in DFT-computed formation energies between initial and model-relaxed structures. This metric evaluates whether models successfully identify energetically favorable relaxation trajectories, with larger reductions indicating superior thermodynamic optimization performance. Additionally, we include the formation energy difference between relaxed structures from DFT and the MLIPs (Figure 1b) to highlight cases where MLIPs achieve better formation energy lowering than DFT.

Maximum atomic force represents the largest force magnitude across all atoms in the relaxed structure. This metric directly quantifies structural equilibrium quality, as well-relaxed crystals should exhibit near-zero atomic forces. Lower maximum forces indicate superior convergence to local energy minima.

Optimization runtime measures the computational time required for structure relaxation, providing essential insights into model efficiency for high-throughput materials discovery applications. This metric is particularly crucial for evaluating the practical deployment potential of different ML approaches as DFT proxies.

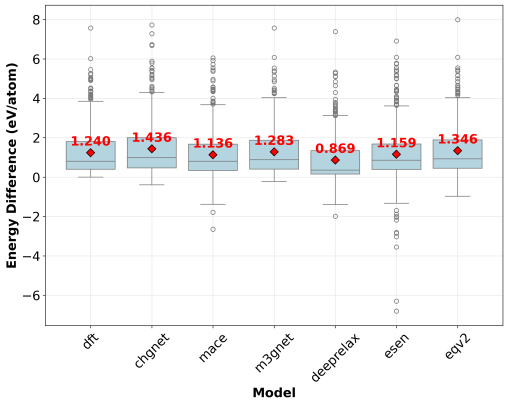
Our main analysis compares MLIP relaxation to a simplified DFT setting with looser relaxation criteria. In \Cref{sec:app-stronger}, we present a stronger DFT relaxation setting that we plan to investigate further as future work.

4 Results and Discussion

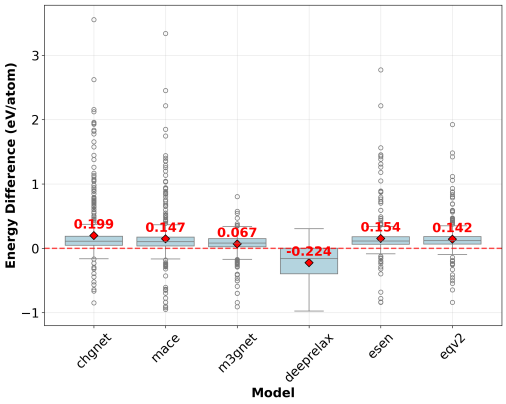
Our comprehensive evaluation reveals distinct performance profiles with clear trade-offs between computational efficiency and relaxation quality across three key metrics.

Formation Energy Reduction: While all models except for DeepRelax outperform DFT in formation energy reduction (Figure 1b), CHGNet and EquiformerV2 demonstrate superior thermodynamic optimization, achieving the most significant formation energy improvements compared to DFT (Figure 1a). For further analysis we also identified the compositions of structures that DeepRelax (our worst performer) achieved lower formation energies than DFT 1.

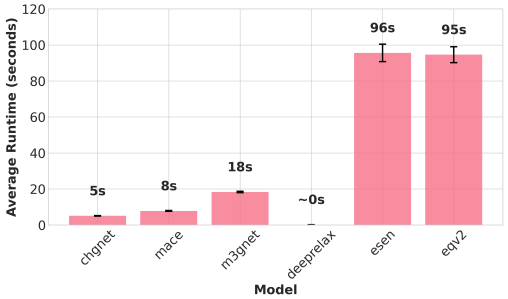
Force Convergence: eSEN and EquiformerV2 emerge as clear leaders, achieving force magnitudes >0.100 eV/Å, better than the next-best performer (M3GNet) and demonstrating superior ability to locate well-converged local minima (Figure 1d). Despite achieving the best energy statistics, CHGNet exhibits the second-worst force convergence, reinforcing that energy accuracy does not guarantee effective structural relaxation.



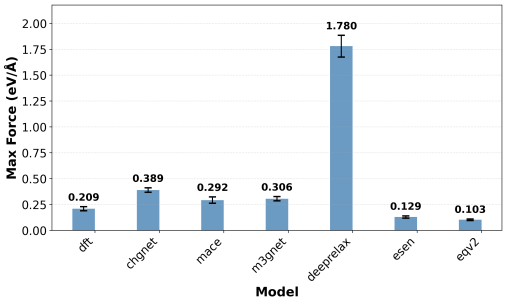
(a) E_{form} differences between unrelaxed and relaxed structures (↑)



(b) E_{form} difference w.r.t DFT relaxation (↑)



(c) Runtime Comparison (↓)



(d) Force Convergence (↓)

Figure 1: Evaluation of ML models for relaxation: **(a)** formation energy reduction achieved during relaxation (↑), **(b)** formation energy difference between MLIP and DFT relaxation (↑), **(c)** runtime of relaxation (↓), **(d)** maximum force magnitudes in relaxed structures (↓). All averages were taken across the intersection of each models outputs.

Computational Efficiency: DeepRelax offers relaxation times orders of magnitude faster than iterative approaches, but this speed comes at substantial accuracy costs in both formation energy and force convergence (Figure 1c). eSEN and EquiformerV2 provide superior accuracy but require significantly longer computation times with numerous outliers. All models perform significantly faster than DFT which has an average runtime of 184.40 seconds.

Overall Assessment: While CHGNet achieves the best formation energy, its force convergence is worse by a much larger margin compared to eSEN and eQV2. MACE emerges as an attractive middle ground, offering reasonable relaxation quality with moderate computational requirements, making it suitable for workflows requiring balanced throughput and accuracy, such as materials discovery pre-screening steps.

5 Conclusion

This work presents the first systematic evaluation of state-of-the-art MLIPs for crystal structure relaxation using unseen RL-generated materials, highlighting trade-offs between computational efficiency and relaxation quality. Our analysis reveals a critical disconnect between energy accuracy and relaxation performance: CHGNet achieves the best formation energy reduction but has the second-worst force convergence. This challenges energy-centric benchmarking and shows that energy prediction alone is insufficient for evaluating relaxation. Rather than identifying a single "best" model, our multi-metric evaluation emphasizes balancing runtime, formation energy reduction, and force convergence according to application needs. Future work should expand dataset size and model diversity to more comprehensively assess MLIPs as DFT proxies across chemical systems.

References

[1] Christopher J. Bartel, Amalie Trewartha, Qi Wang, Alexander Dunn, Anubhav Jain, and Gerbrand Ceder. A critical examination of compound stability predictions from machine-learned formation energies. *npj Computational Materials*, 6(1):97, Jul 2020.

[2] 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*, 35:11423–11436, 2022.

[3] Chi Chen and Shyue Ong. A universal graph deep learning interatomic potential for the periodic table. April 2022.

[4] 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, 2023.

[5] L. Fiedler, K. Shah, M. Bussmann, and A. Cangi. Deep dive into machine learning density functional theory for materials science and chemistry. *Phys. Rev. Mater.*, 6:040301, Apr 2022.

[6] Lenz Fiedler, Normand A. Modine, Steve Schmerler, Dayton J. Vogel, Gabriel A. Popoola, Aidan P. Thompson, Sivasankaran Rajamanickam, and Attila Cangi. Predicting electronic structures at any length scale with machine learning. *npj Computational Materials*, 9(1):115, Jun 2023.

[7] 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, Mar 2025.

[8] 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.

[9] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter*, 21(39):395502, 2009.

[10] Prashant Govindarajan, Mathieu Reymond, Antoine Clavaud, Mariano Phielipp, Santiago Miret, and Sarath Chandar. Crystalgym: A new benchmark for materials discovery using reinforcement learning. In *AI for Accelerated Materials Design - ICLR 2025*, 2025.

[11] Aaron D. Kaplan, Runze Liu, Ji Qi, Tsz Wai Ko, Bowen Deng, Janosh Riebesell, Gerbrand Ceder, Kristin A. Persson, and Shyue Ping Ong. A foundational potential energy surface dataset for materials. *arXiv preprint arXiv:2503.04070*, 2025.

[12] Sungwon Kim. A structure translation model for crystal compounds release for manuscript acceptance. 2023.

[13] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023.

[14] Artem R. Oganov, Chris J. Pickard, Qiang Zhu, and Richard J. Needs. Structure prediction drives materials discovery. *Nature Reviews Materials*, 4(5):331–348, May 2019.

[15] Gianluca Prandini, Antimo Marrazzo, Ivano E Castelli, Nicolas Mounet, and Nicola Marzari. Precision and efficiency in solid-state pseudopotential calculations. npj computational materials, 4 (1): 72, 2018.

[16] 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, Jun 2025.

[17] Ziduo Yang, Yi-Ming Zhao, Xian Wang, Xiaoqing Liu, Xiuying Zhang, Yifan Li, Qiujie Lv, Calvin Yu-Chian Chen, and Lei Shen. Scalable crystal structure relaxation using an iteration-free deep generative model with uncertainty quantification. March 2024.

A DFT parameters

We performed DFT single-point SCF simulations using Quantum Espresso v7.1 [9], which is fully open-source. Solid-state pseudopotentials from SSSP version 1.3.0 [15] were used for the calculations. The settings used are listed below.

1. `calculation`
   - `scf` for single-point calculations
   - `vc-relax` for relaxation
2. `nstep`: 50 (for relaxation)
3. `ecutwfc`: 50
4. `ecutrho`: 400
5. `occupations`: `smearing`
6. `degauss`: 0.001
7. `nspin`: 1
8. `electron_maxstep`: 300
9. `mixing_mode`: plain
10. `mixing_beta`: 0.7
11. `diagonalization`: david
12. `kpoints`: Chosen automatically from Kpoint density.

A.1 Stronger relaxation settings

We performed a stronger DFT relaxation with tighter convergence thresholds to determine if we can converge to structures that are more stable compared to simpler relaxation. The modifications in the setting used above are given below.

1. `forc_conv_thr`: 0.0001
2. `degauss`: 0.01
3. `electron_maxstep`: 500
4. `conv_thr`: $1e-8$
5. `mixing_mode`: local-TF
6. `mixing_beta`: 0.3
7. `ion_dynamics`: bfgs
8. `cell_dynamics`: bfgs
9. `press_conv_thr`: 0.05

A.2 Formation Energy Calculation

The formation energy per atom was calculated using the total energies of the crystals and their constituent elements. The total energies of the isolated elements (88 in the action space) were calculated by performing SCF calculations on the most stable elemental crystals (i.e., 0 formation energy) present in the Materials Project. For elements that do not have a stable elemental crystal (e.g. Lu) or those that have large number of atoms in the elemental crystal (e.g. P, Se), the total energies were calculated for a single atom inside a primary cubic cell of length 10. For a crystal with N atoms, the formation energy (per atom) calculation is defined as follows.

$$E_{form} = \left(\frac{E_{tot} - \sum_i \frac{N_i}{n_i} E^i_{tot}}{N} \right) \quad \text{(eV/atom)} \tag{1}$$

Here, N_i is the number of atoms of the constituent element i present in the crystal, n_i is the number of atoms (sites) of i in the elemental crystal, and E^i_{tot} is the total energy of i in the most stable elemental crystal form.

B Experimental Details

B.1 RL-Generated Crystals

The RL-generated crystals are generated in a dataset-independent manner using DFT-based reward signals starting from existing structures. While they do not necessarily resemble materials in Materials Project, some of the compositions (with unique structures) exist in the database.

B.2 Data Preprocessing and Filtering

Our evaluation dataset initially contained 1,080 crystal structures generated through reinforcement learning using CrystalGym. However, computing DFT forces for model-relaxed structures revealed that a subset of predicted structures failed DFT calculations across different models. To ensure robust statistical comparison, we applied intersection filtering, retaining only structures that successfully completed DFT force calculations for all evaluated models. This filtering process reduced our final analysis set to 831 structures, ensuring consistent evaluation across all models and metrics.

B.3 Relaxation Parameters

For DeepRelax we used the default checkpoint and parameters on the model GitHub. For ASE-based models, we maintained consistent optimization parameters to ensure fair comparison:

1. `optimizer`: Fast Inertial Relaxation Engine (FIRE)
2. `cell filter`: Unit Cell Filter
3. `max optimization steps`: 500
4. `fmax`: 0.01

C Compute Resources

For all tasks we used an Nvidia Quadro RTX 8000 to ensure fair comparison of runtimes.

D Supplementary Results

Table 1: Chemical formulas where DeepRelax achieved lower formation energy than DFT baseline ($n = 127$).

Sn_3C_2	Ca_2SnC_2	Ca_4Ge	$RbZnF_3$	NaP_2N	$SrSn_4$
$Mg(ZnSi)_2$	Zn_2SnGe_2	Sr_3S_2	$CaZn_3Ge$	Be_3Ni_2	NaSClO
$Mg(CBr)_2$	Sr_2GeNF	Ca_3SnO	$CaSn_3O$	Ca_3MgBe	$Ca_2MgBeNi$
Sr_4Mg_3S	Ca_4Zn	$SrLiBr_3$	$NaCa_5HC$	Ca_3C	Ca_2SO
$SrCa_2SO$	KH	Rb_4LiH_3	$SrCa_2Mg_2BeSO$	Na_2MgNCl	Na_2HCl
Li_2HCl	$KNa_3Li_2O_2$	$Rb_3NaSrCaN_2$	Na	K_2BeCl_2	$KLiBeF_2$
PCl_3	SrLiCaBrO	Rb_4Na_2PF	Na_3OF	Mg_3PBr	$CaMg_3Se$
$SrMg_3Se$	$MgBe_2BrCl$	Sr_3Be_2	K_6LiCl	$MgBeCl_2O$	$SrNCl_3$
$SrBeCl_2O$	Be_4C	$CaMgBe_2Se$	Sr_3MgS	Sr_2Mg_2Be	$SrCaC_3$
K_3P	KH_3	Ca_2CSeO_4	Ca_2Se_2S	Se_3S_2	$CaMgCS_2$
$CaMgSe_2S$	Ca_2MgO_2	Sr_3Mg_2	$SrMg_4$	CaMgCSeS	Ca_2Be_2C
LiPCO	Na_2HBr	$RbLi_2Cl$	Rb_2LiH	Li_3Cl	Ca_3O
$SrSeS_3$	$SrBe_3Se$	Na_3F	Se	Mg_3Cl_2	$RbCaH_2F$
$RbCaHF_2$	Na_3Br	H_3Br	Mg_3BeS	Mg_2Be_2S	Na_4Se
KRbSeSO	$SrMg_2O_2$	$SrMgO_3$	K_2HCl	K_2HBr	HBr
Rb_2HCl	$KLiBeCl_2$	$RbNa_2SF$	Rb_2NaSF	$NaMg_3F$	KRbHCl
Ca_2Mg_2O	Sr_3CaC	$SrCO_3$	KRb_3H_4	K_2NaSrF	$SrMgBe_2Se$
$SrMgBeSe_2$	$BeSe_4$	SrBr	$CaBr_3$	$ZnFe_2Sn$	Rb_2Mn_3
$KSeBr_2N$	KSe_2BrN	$MgCoSn_2$	Zn_3Cr_2	Zn_2Cu_3	Sr_2Zn_3
Rb_3NiBr	$MnFeCo_3$	Fe_4Co	$ZnCu_2Sn$	RbZnCuSn	Cu_2SeCl_2
Rb_4Se	KSrCaBr	Sn_3SeBr			

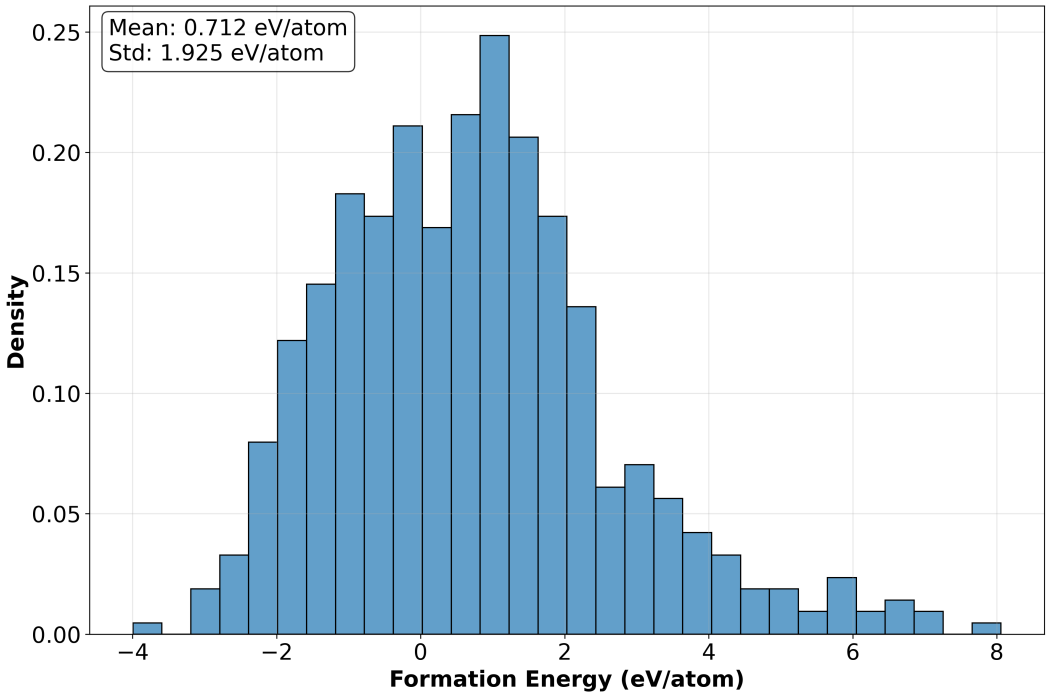


Figure 2: Histogram of formation energies from the unrelaxed crystals from our RL-generated dataset.

Model	Percentage
CHGNet	28.8%
MACE	22.2%
M3GNet	24.0%
DeepRelax	13.9%
eSEN	25.1%
eqV2	27.7%
DFT (simple)	21.8%
DFT (stronger)	29.3%

Figure 3: Percentage of structures with positive-to-negative formation energy transitions per model. Higher percentage means the model is relaxing unstable structures (positive energy) to stable states (negative energy). We include results from both old (simpler relaxation) and new DFT (stronger relaxation A.1) settings to contrast how changing hyperparameters can affect quality of relaxation.

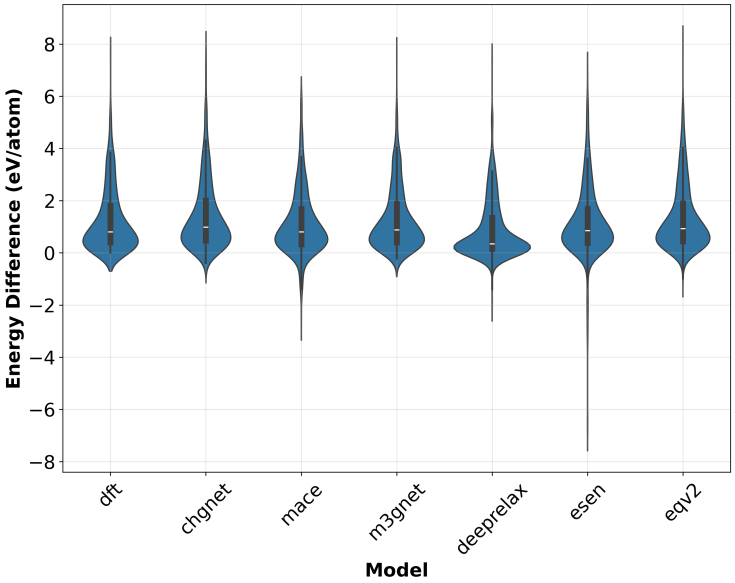


Figure 4: Violin plot of formation energy differences between relaxed and unrelaxed structures. Higher difference indicates better performance.

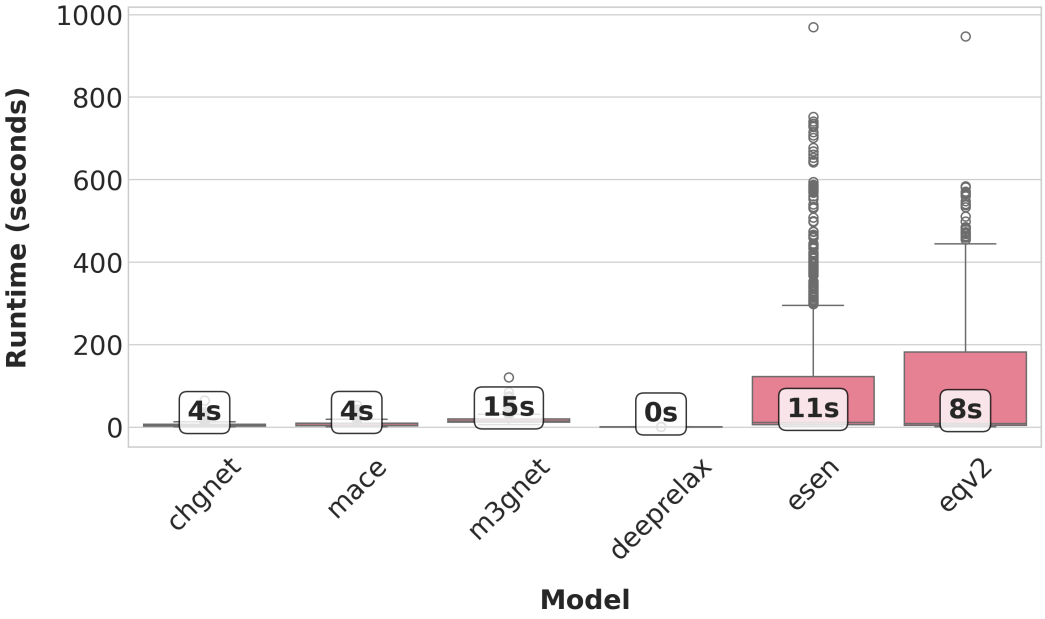


Figure 5: Box plot of runtimes for all models. The box label indicates median runtimes.

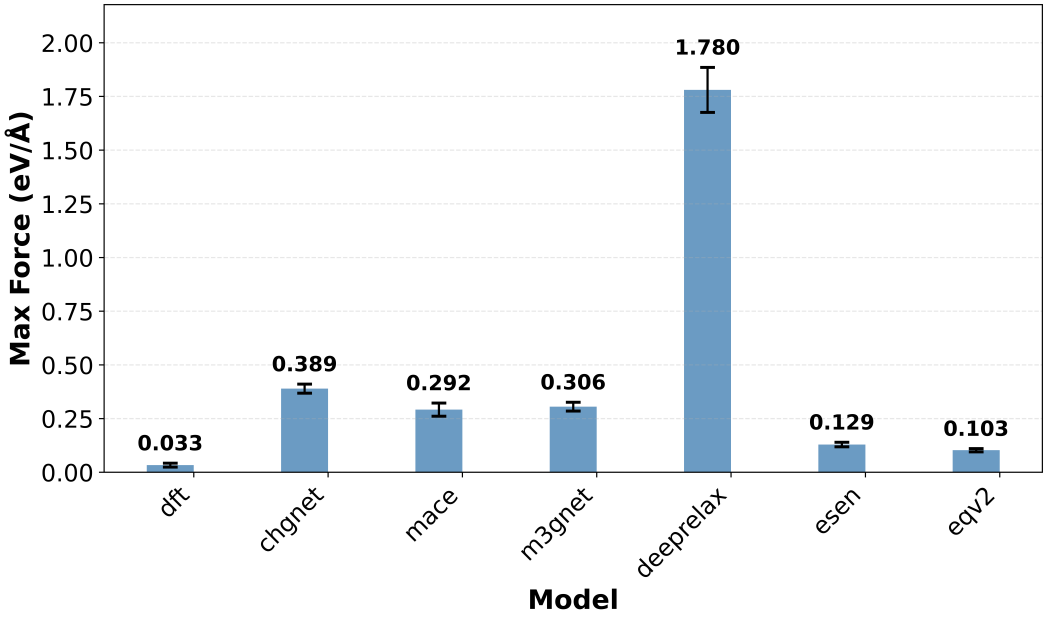


Figure 6: Maximum force magnitudes in relaxed structures (↓) for stronger DFT relaxation setting. Stronger DFT setting (Appendix A.1) achieved average maximum force of 0.200 eV per atom lower than old setting.

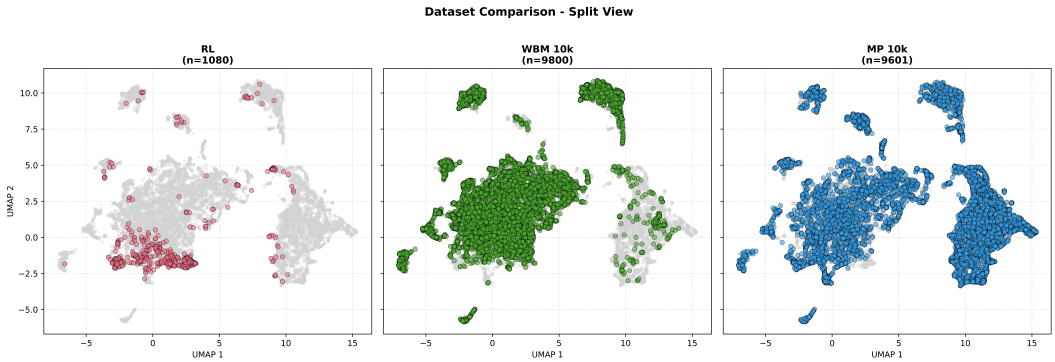


Figure 7: UMAP analysis of RL generated crystals. Embeddings generated with MACE [2]. We compare our dataset with the Materials Project (MP) and Wang Botti Marques (WBM) dataset. UMAP Parameters: n_neighbors=15, min_dist=0.1, n_components=2.

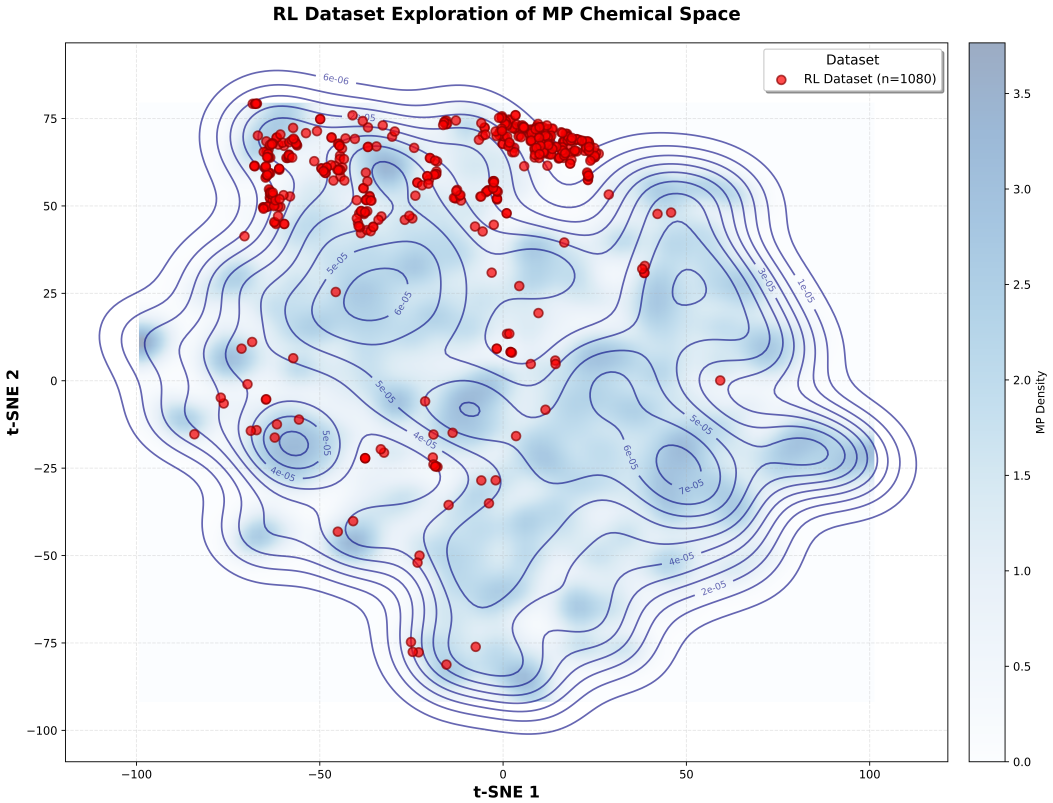


Figure 8: t-SNE analysis of RL generated crystals compared with Materials Project (MP). Embeddings from MACE [2]. t-SNE parameters: perplexity=30, max_iter=1000.