LeMat-Traj: A Scalable and Unified Dataset of Materials Trajectories for Atomistic Modeling

Anonymous Author(s)
Affiliation
Address
email

Abstract

The development of accurate machine learning interatomic potentials (MLIPs) is limited by the fragmented availability and inconsistent formatting of quantum mechanical trajectory datasets derived from Density Functional Theory (DFT). These datasets are expensive to generate yet difficult to combine due to variations in format, metadata, and accessibility. To address this, we introduce LeMat-Traj, a curated dataset comprising over 120 million atomic configurations aggregated from large-scale repositories, including the Materials Project, Alexandria, and OQMD. LeMat-Traj standardizes data representation, harmonizes results and filters for high-quality configurations across widely used DFT functionals (PBE, PBESol, SCAN, r2SCAN), significantly lowering the barrier for training transferrable and accurate MLIPs. LeMat-Traj spans both relaxed low-energy states and high-energy, high-force structures, complementing molecular dynamics and active learning datasets. By fine-tuning models pre-trained on high-force data with LeMat-Traj, we achieve a significant reduction in force prediction errors on relaxation tasks. We also present LeMaterial-Fetcher, a modular and extensible open-source library developed for this work, designed to provide a reproducible framework for the community to easily incorporate new data sources and ensure the continued evolution of large-scale materials datasets. LeMat-Traj and LeMaterial-Fetcher are publicly available at `https://huggingface.co/datasets/LeMaterial/LeMat-Traj` and `https://github.com/LeMaterial/lematerial-fetcher`.

1 Introduction

The discovery and design of novel materials are essential for technological advancement, offering solutions to pressing global challenges such as sustainable energy and climate change mitigation [31]. However, traditional lab experiments and computational approaches, particularly those involving Density Functional Theory (DFT), are resource-intensive [46]. Machine Learning Interatomic Potentials (MLIPs) have emerged as a promising alternative, offering DFT-level accuracy at a fraction of the computational cost. This acceleration is crucial for enabling large-scale molecular dynamics (MD) simulations over long timescales and rapid exploration of material properties [41, 12], potentially fast-tracking the development of materials for applications like carbon capture, improved batteries, or more efficient catalysts.

Graph Neural Networks (GNNs) have emerged as the most effective class of models for learning interatomic potentials, due to their ability to naturally represent atomic systems and to incorporate physical symmetries such as rotational and permutational equivariance [11]. As modern GNN architectures like EquiformerV2 [25] exhibit scaling laws behaviors [6], the need for even larger, more diverse, and consistently processed datasets becomes predominant. Despite several large-scale initiatives generating vast amounts of DFT data [17, 36], these datasets often remain siloed, employ distinct

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

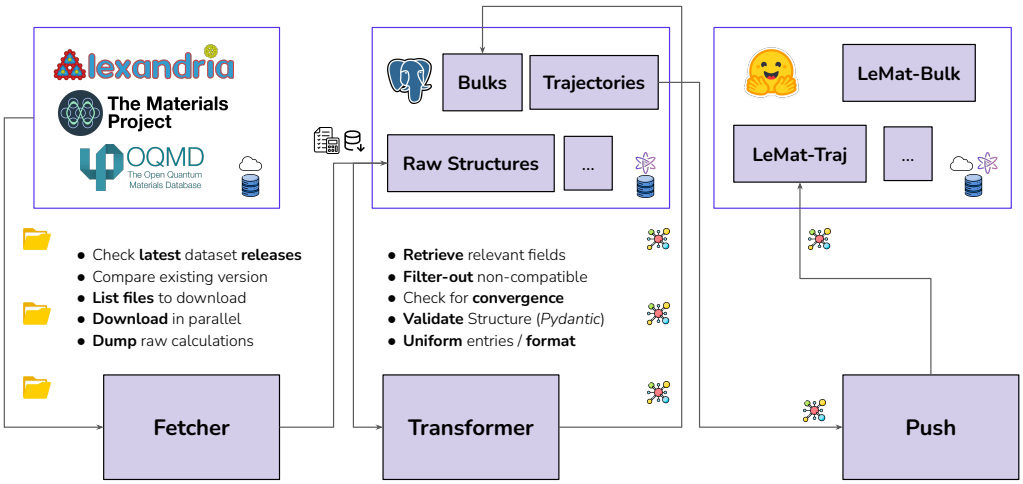


Figure 1: Data curation pipeline of LeMaterial-Fetcher. The library automates the process of fetching, transforming, validating, and harmonizing data from various sources, ensuring a consistent and reproducible dataset. The pipeline currently supports the continuous integration of fully relaxed bulk structures and full relaxation trajectories.

data formats, and use varying DFT parameters (e.g., functionals, parameters, pseudopotentials). This fragmentation poses a challenge for researchers aiming to leverage the full spectrum of available data, as combining these sources requires considerable preprocessing and harmonization efforts. Consequently, many MLIPs are trained on scattered and non-homogeneous datasets, potentially restricting their generalizability and predictive power, while introducing chemical bias due to the way the datasets are separately being used to train these models [36]. Moreover, large architectures—some now comprising over 30 million parameters—stand to benefit from the access to even bigger and more diverse datasets, as further scaling require proportionally more data to avoid overfitting and fully realize their expressive power [25].

To overcome these limitations, we introduce LeMat-Traj, a large-scale, aggregated dataset of materials trajectories. LeMat-Traj harmonizes data from three prominent sources: Materials Project [17, 18], Alexandria [36], and OQMD [34], into a unified format, encompassing calculations performed with various DFT functionals (PBE, PBESol, SCAN, and r2SCAN). Furthermore, we introduce `LeMaterial-Fetcher`, an open-source Python library designed for the systematic and reproducible curation of materials science datasets.

Our contributions can be summarised as follows:

1. We release LeMat-Traj, to our knowledge one of the largest publicly available datasets of crystalline materials trajectories (120 million configurations). LeMat-Traj provides dense, high-quality coverage of near-equilibrium and low-force states—an underrepresented but crucial regime for accurate geometry optimization.
2. We empirically demonstrate the value of this data philosophy through extensive benchmarks. We show that by fine-tuning a MACE model with LeMat-Traj, we can reduce force prediction errors on relaxation tasks by over 36% and improve performance on the Matbench Discovery stability benchmark by 10%.
3. We introduce LeMaterial-Fetcher, a modular and extensible open-source library used to create LeMat-Traj. It provides a reproducible platform for community-driven curation, extension, and combination of large-scale materials datasets, enabling future research in multi-dataset and curriculum learning strategies.

We believe LeMat-Traj and LeMaterial-Fetcher will serve as a versatile foundation for the community, supporting not only the training of MLIPs but also a wide range of downstream tasks with crystalline materials, including benchmarking, subsampling strategies, self-supervised pretraining, and curriculum learning.

2 Related Work

The development of MLIPs has been closely correlated with the availability of suitable training datasets [7, 18, 22]. These datasets typically consist of sequences of atomic configurations, along with their corresponding energies and forces, generated from quantum mechanical simulations. Such sequences, often referred to as *trajectories*, can originate from various simulation types, including geometry optimizations (tracing paths to energy minima) or molecular dynamics (MD, exploring configurations at specific thermodynamic conditions). Large-scale computational materials science initiatives like the Materials Project [17, 18], Alexandria [35, 36], and the Open Quantum Materials Database (OQMD) [34], along with resources like AFLOW [13], NOMAD [10], and ColabFit [43], have provided invaluable data to the community.

While MLIPs are frequently trained using data derived from these sources such as MPtrj [8] which curates relaxation trajectories from the Materials Project and has been used in models like CHGNet [8], MACE [4] and subsequent architectures, practitioners frequently encounter challenges [28]. Data from these diverse sources may employ different DFT parameters (e.g., functionals, k-points, pseudopotentials), varying data formats, and inconsistent preprocessing methodologies [33]. This fragmentation means that combining data requires considerable, often repetitive, data engineering efforts [45], potentially limiting the generalizability and predictive power of the resulting MLIPs, and can introduce chemical biases depending on how individual datasets are leveraged [28].

In parallel, two main philosophies of dataset design for training such MLIPs have emerged. One emphasizes broad exploration of the potential energy surface through high-force sampling, as in OMat24 [3], MatPES [19] or MP-ALOE [20], and active-learning datasets like ANI-1x [39]. These are well suited for pretraining robust models and capturing diverse regions of the configuration space [14]. The other focuses on dense, near-equilibrium sampling from DFT geometry optimization trajectories, which provide clean, structured data in the low-force regime critical for accurate geometry optimization and stability prediction with datasets like Alexandria [36]. Since machine learning force fields often display varying accuracy across the potential energy surface, with near-equilibrium and high-force regions posing different challenges [42, 26], these two philosophies of dataset design reflect complementary but compatible strategies to address that imbalance.

Recent work has underscored the need for large, harmonized, and extensible datasets that bridge these philosophies and mitigate fragmentation [19, 36]. Our contribution follows this direction by introducing a systematically curated dataset of DFT trajectories together with an open-source pipeline to ensure reproducibility and extensibility. This places our work in line with ongoing efforts toward foundational datasets in materials science that can serve pretraining, benchmarking, and fine-tuning across a wide range of downstream MLIP applications.

Our work aims to address the data fragmentation challenge by providing not only a large, aggregated dataset but also a transparent, reproducible curation pipeline with LeMaterial-Fetcher. This aligns with the increasing need for foundational datasets in materials science [19] that are large-scale, internally coherent, and extensible, facilitating pretraining, benchmarking, and fine-tuning across a wide range of downstream MLIP applications.

3 Methodology

LeMat-Traj is constructed by aggregating and processing data primarily from three major materials databases: Materials Project, Alexandria and OQMD (Open Quantum Materials Database). The core challenge lies in developing a scalable and reproducible methodology to handle the existing heterogeneity of these sources into a single and unified dataset.

3.1 Unified Generation Pipeline

To address this, we developed LeMaterial-Fetcher, a highly parallelized Python-based open-source library described in Figure 1. It provides a unified and automated framework for:

- **Fetching**: Interfacing with open APIs and direct downloads from various data sources.
- **Transformation**: Converting diverse input formats and attributes into a consistent schema. This includes standardizing atomic structure representations, energy units, and force compo-

nents. It also handles the extraction and organization of metadata related to DFT calculations. All of this is done by allowing to interface with powerful atomistic modelling tools like `Pymatgen` [30], `Matminer` [44].

- **Validation**: Implement checks to ensure data quality and integrity, such as verifying physical plausibility or consistency across reported values.
- **Harmonization**: Aligning DFT calculation parameters where possible and creating separate splits of data based on key parameters like the DFT functional.
- **Push**: Exporting the curated dataset in a user-friendly and efficient format, for direct use with libraries like HuggingFace's `Datasets` [23]. This allows for easy integration with existing ML frameworks and tools, because they can adapt to limited computational resources, but also data versioning and metadata tracking as outlined in [10].

LeMaterial-Fetcher is designed to be modular, extensible but also scalable and fast, allowing for the easy integration of new data sources (e.g., future integration of quantum calculations sources) and adaptation to more materials science domains such catalysis, experiments, defects. This framework ensures the reproducibility of LeMat-Traj and facilitates continuous integration of new DFT calculations as they become available from the source databases. Mainly, this eliminates the need to have to manually go through every dataset one by one, download it, and then apply the updates before releasing new versions. Additional details on the pipeline design are provided in Appendix E.

3.2 Data Sources and Harmonization

LeMat-Traj specifically extracts geometry optimization trajectories from DFT calculations. A key aspect of our curation is the harmonization of data across different exchange–correlation functionals. We categorize trajectories based on the reported functional, primarily focusing on PBE, PBESol, SCAN, and r2SCAN, allowing users to train functional-specific models or to explore multi-fidelity learning across levels of theory (section 5). Table 1 gives a full summary of the dataset repartition.

The dataset follows the OPTIMADE specification [2], enabling interoperability with other datasets that follow the same standard. We introduce a slight adaptation to accommodate trajectory data: each entry in the database corresponds to an individual atomistic configuration, which is part of a trajectory and is associated with energy and force information. Full optimization trajectories can be reconstructed by grouping entries by a shared trajectory identifier. This design choice facilitates seamless integration into machine learning interatomic potential (MLIP) training pipelines, where per-frame forces and energies are required.

To support trajectory-specific use cases, two new fields are introduced into the schema:

1. **Relaxation Step**: An integer indicating the step number of the structure within a given geometry optimization sequence.
2. **Relaxation Number**: An identifier that distinguishes different optimization runs for the same initial structure. This is particularly useful in high-throughput settings, where structures may undergo coarse relaxations before being re-relaxed with tighter thresholds or more accurate methods.

Table 1: Number of trajectories and atomic configurations per source database and functional.

Functional	Database	Number of Trajectories	Number of configurations
PBE	Materials Project	195,721	3,649,785
	Alexandria	3,414,074	110,804,226
	OQMD	135,966	264,782
PBESol	Materials Project	39,981	309,873
	Alexandria	252,791	6,099,623
SCAN	Materials Project	7,756	180,528
r2SCAN	Materials Project	37,888	516,576

3.3 Data Filtering

Our data filtering strategy prioritizes retaining a large volume of diverse configurations while establishing essential quality control. To this end, several criteria were applied: First, any atomic configuration lacking either energy or atomic force data was discarded. Second, entire trajectories were removed if the energy difference between the penultimate and final optimization step exceeded a threshold of 2×10^{-2} eV, a criterion adapted from MPtrj [8] to ensure reasonable convergence. Third, trajectories were also excluded if the maximum atomic force norm in the final configuration surpassed 0.2 eV/Å, i.e. the structure is not fully relaxed. While this force threshold is relatively high, it allows the inclusion of structures that, despite not being fully relaxed, still provide valuable information about the potential energy surface far from equilibrium, enriching the dataset for training robust force fields. Finally, all configurations were validated against the OPTIMADE format specifications, and any entry failing these schema checks or other implemented validation tests was removed.

4 Coverage of Chemical and Configurational Space

LeMat-Traj comprises approximately 120 million atomic structures derived from geometry optimization trajectories. The dataset is partitioned based on the DFT functional used for the calculations: PBE, PBESol, SCAN, and r2SCAN. This partitioning facilitates targeted model training and research into multi-fidelity approaches.

4.1 Chemical and Structural Diversity

We compare the elemental and structural diversity of LeMat-Traj with other popular datasets such as MPtrj [8] and MatPES [19]. LeMat-Traj aims to offer a broader coverage by combining multiple sources as illustrated in Figure 2. While MPtrj primarily focuses on Materials Project data, LeMat-Traj's explicit harmonization and inclusion of OQMD and Alexandria data offer a unique combination of scale and more balanced distribution.

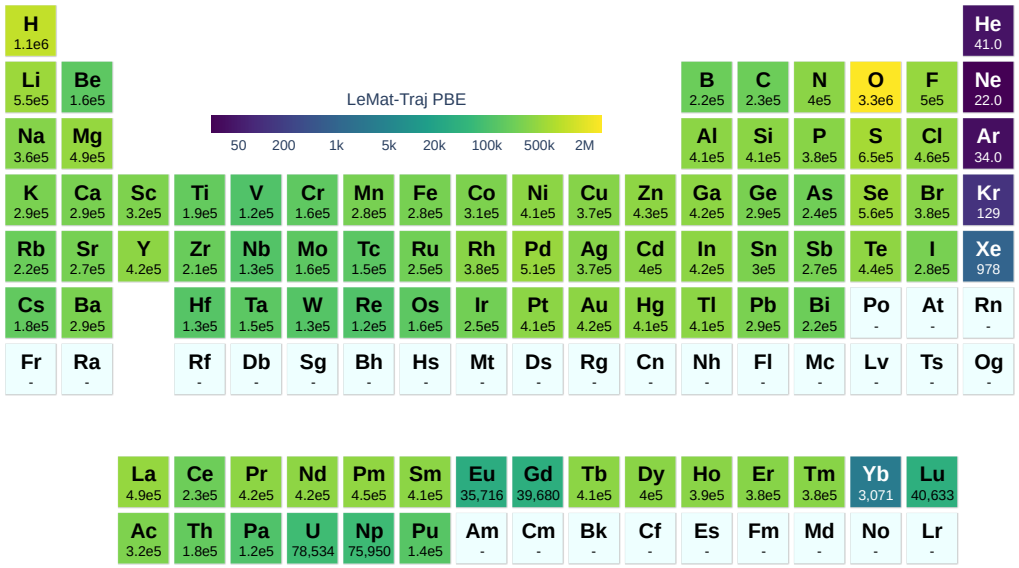


Figure 2: Chemical distribution in number of trajectories for the PBE split of LeMat-Traj using `Pymatviz` [32].

While the Alexandria dataset constitutes the majority of the PBE split by volume (approx. 92%), the inclusion of data from Materials Project and OQMD is critical for diversity. First, it enriches the chemical space; Materials Project contains a higher concentration of oxides and battery materials, balancing the bi-metallic bias present in Alexandria. Second, it diversifies the force distribution; the average maximum force norm in Materials Project trajectories is significantly higher (593 meV/Å)

than in the rest of the dataset (110 meV/Å), providing crucial high-force examples that help prevent models from under-estimating forces during relaxation.

Inclusion of Equilibrium Structures. A notable feature of LeMat-Traj is the inclusion of equilibrium structures from OQMD, which is rarely leveraged by ML practitioners when training machine-learned interatomic potentials (MLIPs). These configurations, characterized by near-zero atomic forces, serve as valuable reference points for MLIPs, particularly in capturing energy minima accurately. While relaxation trajectories naturally include low-force structures near convergence, the explicit addition of a large and diverse set of OQMD equilibrium configurations enhances the dataset's richness. Although these single-point structures may be underrepresented compared to the total number of frames in full trajectories, they can be strategically leveraged by models focused on accurately learning stable configurations.

4.2 Trajectory Analysis

Trajectory Length. Figure 3 shows the distribution of trajectory lengths in LeMat-Traj. LeMat-Traj exhibits a broad distribution, with many trajectories across all length scales. It uniquely features a long tail with a significant number of trajectories extending beyond 100 frames, and even exceeding 1000 frames. In contrast, MPtrj is predominantly characterized by shorter trajectories, with the majority having fewer than 50 frames and a pronounced spikiness in its distribution at very short lengths. MatPES shows a broader distribution than MPtrj, with more medium-length trajectories (up to 100-200 frames), but still lacks the extensive representation of very long trajectories seen in LeMat-Traj. These longer trajectories are not indicative of optimization issues but are rather a feature of the highly stringent convergence criteria used in the source calculations, representing valid, but slow, convergence paths to energy minima.

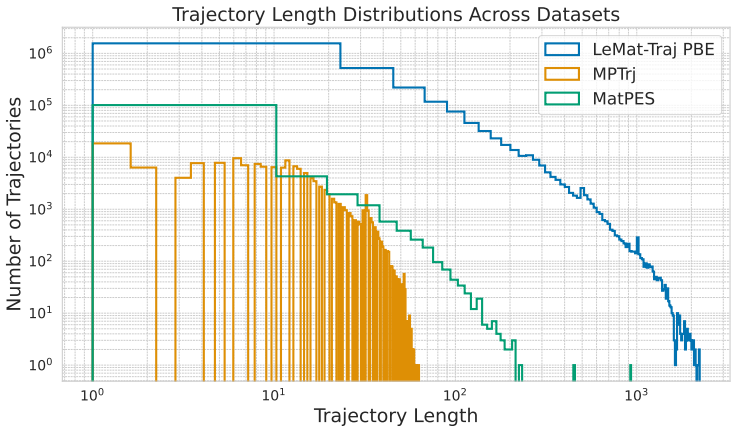
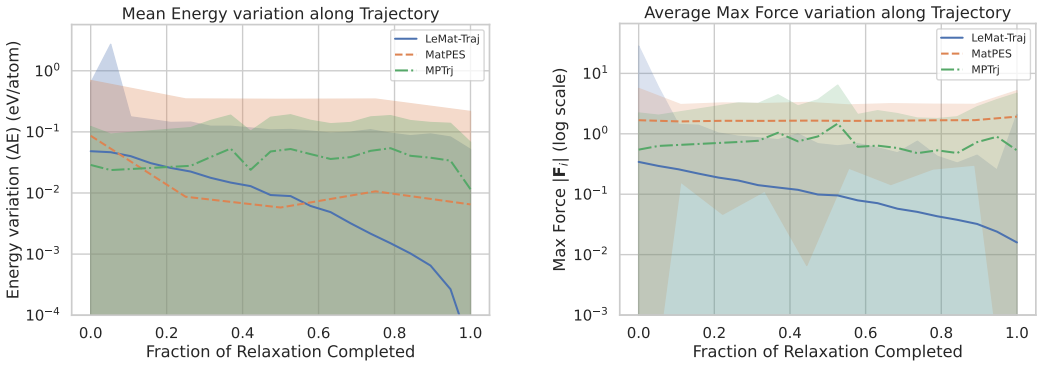


Figure 3: Comparison of trajectory length distributions for LeMat-Traj (PBE split), MPtrj, and MatPES, on a log-log scale. For every trajectory, the number of configurations associated is computed. LeMat-Traj exhibits a broader range of trajectory lengths.

Targets spread along trajectories. Figure 4 illustrates the evolution of mean energy variation (ΔE relative to the final relaxed state) and average maximum atomic forces norm throughout the relaxation trajectories of LeMat-Traj, MatPES, and MPtrj. LeMat-Traj uniquely demonstrates comprehensive sampling across the entire relaxation pathway. At the initial stages (low fraction of relaxation completed), it encompasses a wide distribution of high-energy and high-force configurations, with mean ΔE around 0.05 eV/atom (and variance extending >1 eV/atom from structures that are very far from their relaxed states iniially) and mean maximum forces around 0.3-0.4 eV/Å (variance extending >1 eV/Å). Crucially, as relaxations progress towards completion, LeMat-Traj systematically converges to very low ΔE (approaching $10^{-3} - 10^{-4}$ eV/atom) and near-zero maximum forces (mean 0.01-0.02 eV/Å, with significant density below 10^{-3} eV/Å). This shows a robust sampling both far-from-equilibrium states and accurately representing near-equilibrium energy minima and low-force structures, making LeMat-Traj well-suited for training versatile MLIPs capable of both high accuracy for stable configurations and robustness across diverse energy landscapes.



(a) $\Delta E = E^t - E^T$ (b) Maximum Force Norm

Figure 4: Evolution of mean energy variation ($\Delta E = E^t - E^T$, where E^t is current step energy and E_T is final relaxed energy) per atom (a) and average maximum atomic force (b) as a function of the fraction of relaxation completed. Trajectories from LeMat-Traj, MatPES, and MPtrj are binned by their normalized progress. Solid lines represent the mean values, and shaded areas depict one standard deviation, both on a logarithmic y-axis. LeMat-Traj demonstrates comprehensive sampling from high-energy/high-force initial states to well-converged, low-energy/low-force final states.

5 Results

To empirically validate the utility of LeMat-Traj, we conduct a series of benchmark experiments using the MACE architecture [4], a well-established and performant equivariant model. These experiments are designed to demonstrate the dataset's value for improving model accuracy on relaxation-focused tasks, both through fine-tuning and in downstream applications. A key hypothesis of our work is that LeMat-Traj's dense sampling of near-equilibrium states is complementary to datasets focused on high-force configurations. While high-force data, such as in OMat24, provides strong gradients that facilitate stable initial training and learning of the general energy landscape, LeMat-Traj is designed to refine model accuracy in the low-force regime critical for geometry optimization.

5.1 Complementary Value for Fine-Tuning

To test our hypothesis, we evaluate the performance of a MACE model pre-trained on the general-purpose OMat24 dataset and then fine-tune it on LeMat-Traj. As shown in Table 2, while the OMat24-trained model serves as a strong baseline, fine-tuning on LeMat-Traj reduces the Forces MAE on our held-out test set of relaxation trajectories by over 36%. This result provides direct evidence that LeMat-Traj contains critical information for achieving high fidelity in force predictions near energy minima, a crucial capability for accurate geometry optimization.

Table 2: Performance of MACE on the LeMat-Traj PBE 10K held-out test set. Fine-tuning a model pre-trained on OMat24 with LeMat-Traj significantly reduces prediction errors, demonstrating the complementary nature of the datasets.

MACE Training Dataset	Energy MAE (meV) ↓	Forces MAE (meV/Å) ↓	Forces Cos ↑
OMat24	59.5	42.7	0.29
MPtrj	49.8	81.7	0.23
MatPES PBE	316.4	88.7	0.16
LeMat-Traj only	25.3	50.8	0.23
OMat24 + ft LeMat-Traj	**18.8**	**27.2**	**0.30**

5.2 Downstream Performance on Matbench Discovery

To assess practical utility, we evaluate our models on a subset of the Matbench Discovery benchmark, which measures a model's ability to predict the stability of novel crystalline materials. This task relies heavily on accurate structural relaxation. Table 3 shows that the MACE model trained on a split of LeMat-Traj with left-out matching protocol from Matbench Discovery following the method in Barroso-Luque et al. [3] significantly outperforms the same model architecture trained on OMat24 or MPtrj alone, achieving a 10% higher F1 score. The best performance is achieved by the model pre-trained on OMat24 and fine-tuned on LeMat-Traj, reinforcing the value of combining high-force and near-equilibrium data,

Table 3: Matbench Discovery benchmark results on a 50k uniform subset. Models incorporating LeMat-Traj data achieve superior performance in predicting material stability.

Model (Training Set)	**F1 Score ↑**	**MAE (meV) ↓**	**RMSE (meV) ↓**
MACE (OMat24)	0.575	87.8	172.8
MACE (MPtrj)	0.694	47.2	83.9
MACE (LeMat-Traj Full)	0.768	37.2	69.0
MACE (OMat24 + ft-LeMat-Traj)	**0.772**	**33.4**	**67.8**

5.3 Multi-Fidelity Learning.

A notable challenge in materials modeling is the transferability of MLIPs trained on data from one level of theory (e.g., a specific DFT functional) to another. LeMat-Traj, with standardized formats of its different splits for PBE, PBESol, SCAN, and r2SCAN, provides a natural testbed for multi-fidelity learning strategies. We conduct experiments to assess how well models trained on one functional (e.g., PBE) can be fine-tuned or adapted for tasks involving another functional (e.g., PBESol and r2SCAN). For each of the PBESol and r2SCAN datasets, we use the subset described in Appendix D.1 during the experiments.

1. We train a MACE model from scratch (using the same number of parameters as `MACE-MPA-0` [5]). The training procedure is done in two stages (similar to how the foundation model is trained from scratch). During the first stage, the forces' weight in the loss computation is way higher than the other predicted targets, then during the second stage, we match the energy weight to that of the forces weight.
2. We fine-tune that same model separately on the split.

Evaluation results on the test set are reported in Table 4. LeMat-Traj helps facilitate effective transfer learning across functionals, especially when data or computational resources are limited, and can help in the development and research of general cross-atomic data source learning methods like Shoghi et al. [37], Huang et al. [16]. Results show that using a model pre-trained on one functional helps transferring to another functional more easily and in fewer steps.

Table 4: Performance of pre-trained MACE and ORB Models on Different DFT Functionals split when fine-tuning on a functional split (referred to with `<split>`) and after fine-tuning (`-<split>-ft`). Energy MAE is reported in meV/atom, Force MAE in meV/Å, Stress MAE in meV/Å^3, and Cosine Similarity is averaged over the forces vectors. All measures are across the test split described in Appendix D.

Model	PBESol				r2SCAN		
	Energy MAE	Force MAE	Stress MAE	Cosine Sim.	Energy MAE	Force MAE	Cosine Sim.
`MACE-MPA-0`	370.9	101	14.7	0.13	9204.9	111	0.15
`MACE-PBESol`	51.2	33	2.1	0.04	/	/	/
`MACE-MPA-0-PBESol-ft`	18.0	27	1.6	0.19	/	/	/
`MACE-r2SCAN`	/	/	/	/	141.7	36	0.09
`MACE-MPA-0-r2SCAN-ft`	/	/	/	/	96.3	28	0.22

5.4 Limitations and Future Work

While LeMat-Traj and LeMaterial-Fetcher mark substantial advancements, several areas offer opportunities for improvement. The current dataset primarily consists of DFT geometry optimization trajectories, and does not include molecular dynamics (MD) trajectories, which could enhance modeling of dynamic properties. Additionally, although the dataset is chemically diverse, the PBE split is largely drawn from the Alexandria database, potentially introducing some data source bias. Future work should aim to incorporate MD trajectories and correctly identify them to diversify data origins, while ensuring compatibility and avoid incorporating noisy data points. This initial release primarily focuses on dataset construction and characterization; comprehensive benchmarking of MLIPs trained on LeMat-Traj is planned to fully demonstrate its utility (preliminary results in Appendix D). Finally, the pipeline in LeMaterial-Fetcher is designed to gather detailed DFT calculation parameters if available from the source (e.g., k-point meshes, pseudopotentials). While not fully exploited in the current version of LeMat-Traj for all entries, this capability can help introduce future MLIP architectures that explicitly embed these parameters as inputs, leading to more versatile multi-fidelity models, enabling LeMat-Traj to continually evolve as a richer resource for the community. Aggregating data from sources using different underlying DFT parameters (e.g., k-point grids, pseudopotentials) without explicit harmonization risks introducing noise. While we ensure pseudopotential compatibility for included elements following the method in Siron et al. [38], a deeper quantitative analysis of these potential cross-database biases is an important area for future investigation. We note that LeMaterial-Fetcher's provenance tracking is a first step, enabling researchers to isolate and study these effects.

6 Conclusion

In this work, we introduced LeMat-Traj, a scalable, high-quality and unified dataset comprising over 120 million atomic configurations from DFT relaxation trajectories, and LeMaterial-Fetcher, the open-source library enabling its creation and continued evolution. By aggregating, standardizing, and harmonizing data from prominent repositories across multiple DFT functionals, LeMat-Traj lowers the barrier for training robust, transferable, and accurate MLIPs, which are essential technical bricks of accelerated materials discovery. Our analysis demonstrates its comprehensive sampling of the potential energy surface along relaxation pathways, capturing both high-energy structures and near-equilibrium states, making it a valuable resource for researchers to develop next-generation interatomic potentials, explore multi-fidelity learning, and advance self-supervised learning techniques in materials science.

While LeMat-Traj currently focuses on geometric optimization trajectories, the modularity of LeMaterial-Fetcher enables future expansions. With the incorporation of compatible molecular dynamics simulations, diversifying data sources further, and implementing dataset-level sampling strategies for more coherent fine-tuning datasets. Integrating LeMaterial-Fetcher with automated active learning and DFT calculation workflows can enable the continuous enrichment of LeMat-Traj with high-fidelity data. We believe LeMat-Traj and LeMaterial-Fetcher represent a step towards democratizing access to high-quality, curated training data, fostering community collaboration, and ultimately accelerating the pace of data-driven materials discovery

References

[1] Brandon Amos. Tutorial on amortized optimization. *arXiv preprint arXiv: 2202.00665*, 2022.

[2] Casper W Andersen, Rickard Armiento, Evgeny Blokhin, Gareth J Conduit, Shyam Dwaraknath, Matthew L Evans, Ádám Fekete, Abhijith Gopakumar, Saulius Gražulis, Andrius Merkys, et al. Optimade, an api for exchanging materials data. *Scientific data*, 8(1):217, 2021.

[3] 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. *arXiv preprint arXiv: 2410.12771*, 2024.

[4] 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. *arXiv preprint arXiv: 2206.07697*, 2022.

[5] 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. *arXiv preprint arXiv: 2401.00096*, 2023.

[6] Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale?, 2024. URL `https://arxiv.org/abs/2410.23179`.

[7] L. Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, M. Rivière, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. L. Zitnick, and Zachary W. Ulissi. The open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 2020. doi: 10.1021/acscatal.0c04525.

[8] 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. *arXiv preprint arXiv: 2302.14231*, 2023.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL `https://arxiv.org/abs/1810.04805`.

[10] Claudia Draxl and Matthias Scheffler. The nomad laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials*, 2(3):036001, may 2019. doi: 10.1088/2515-7639/ab13bb. URL `https://dx.doi.org/10.1088/2515-7639/ab13bb`.

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

[12] Alexandre Agm Duval, Victor Schmidt, Alex Hernández-Garcıa, Santiago Miret, Fragkiskos D Malliaros, Yoshua Bengio, and David Rolnick. Faenet: Frame averaging equivariant gnn for materials modeling. In *International Conference on Machine Learning*, pages 9013–9033. PMLR, 2023.

[13] Hagen Eckert, Simon Divilov, Michael J Mehl, David Hicks, Adam C Zettel, Marco Esters, Xiomara Campilongo, and Stefano Curtarolo. The aflow library of crystallographic prototypes: Part 4. *Computational Materials Science*, 240:112988, 2024.

[14] 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. *arXiv preprint arXiv:2210.07237*, 2022.

[15] Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2019.106949. URL `https://www.sciencedirect.com/science/article/pii/S0010465519303042`.

[16] Xu Huang, Bowen Deng, Peichen Zhong, Aaron D. Kaplan, Kristin A. Persson, and Gerbrand Ceder. Cross-functional transferability in universal machine learning interatomic potentials. *arXiv preprint arXiv: 2504.05565*, 2025.

[17] 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.

[18] Anubhav Jain, Joseph Montoya, Shyam Dwaraknath, Nils ER Zimmermann, John Dagdelen, Matthew Horton, Patrick Huck, Donny Winston, Shreyas Cholia, Shyue Ping Ong, et al. The materials project: Accelerating materials design through theory-driven data and tools. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pages 1751–1784, 2020.

[19] 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.

[20] Matthew C Kuner, Aaron D Kaplan, Kristin A Persson, Mark Asta, and Daryl C Chrzan. Mp-aloe: An r2scan dataset for universal machine learning interatomic potentials. *arXiv preprint arXiv:2507.05559*, 2025.

[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] Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The open molecules 2025 (omol25) dataset, evaluations, and models. *arXiv preprint arXiv: 2505.08762*, 2025.

[23] Quentin Lhoest, Albert Villanova Del Moral, Yacine Jernite, Abhishek Thakur, Patrick Von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*, 2021.

[24] Yi-Lun Liao, Tess Smidt, Muhammed Shuaibi, and Abhishek Das. Generalizing denoising to non-equilibrium structures improves equivariant force fields. *arXiv preprint arXiv: 2403.09549*, 2024.

[25] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations, March 2024. URL `http://arxiv.org/abs/2306.12059`. arXiv:2306.12059 [physics] version: 3.

[26] Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Universal machine learning interatomic potentials are ready for phonons. *npj Computational Materials*, 11(1):178, 2025.

[27] Santiago Miret, Kin Long Kelvin Lee, Carmelo Gonzales, Sajid Mannan, and N. M. Anoop Krishnan. Energy force regression on dft trajectories is not enough for universal machine learning interatomic potentials. *arXiv preprint arXiv: 2502.03660*, 2025.

[28] David Montes de Oca Zapiain, Mitchell A Wood, Nicholas Lubbers, Carlos Z Pereyra, Aidan P Thompson, and Danny Perez. Training data selection for accuracy and transferability of interatomic potentials. *npj Computational Materials*, 8(1):189, 2022.

[29] Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential. *arXiv preprint arXiv: 2410.22570*, 2024.

[30] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2012.10.028. URL https://www.sciencedirect.com/science/article/pii/S0927025612006295.

[31] Edward O Pyzer-Knapp, Jed W Pitera, Peter WJ Staar, Seiji Takeda, Teodoro Laino, Daniel P Sanders, James Sexton, John R Smith, and Alessandro Curioni. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(1):84, 2022.

[32] Janosh Riebesell, Haoyu Yang, Rhys Goodall, and Sterling G. Baird. Pymatviz: visualization toolkit for materials informatics, 2022. URL https://github.com/janosh/pymatviz. 10.5281/zenodo.7486816 - https://github.com/janosh/pymatviz.

[33] Elliot Rossomme, Leonardo A Cunha, Wanlu Li, Kaixuan Chen, Alexandra R McIsaac, Teresa Head-Gordon, and Martin Head-Gordon. The good, the bad, and the ugly: pseudopotential inconsistency errors in molecular applications of density functional theory. *Journal of Chemical Theory and Computation*, 19(10):2827–2841, 2023.

[34] James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65:1501–1509, 2013.

[35] Jonathan Schmidt, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel A. L. Marques. Crystal graph attention networks for the prediction of stable materials. *Science Advances*, 7(49): eabi7948, 2021. doi: 10.1126/sciadv.abi7948. URL https://www.science.org/doi/abs/10.1126/sciadv.abi7948.

[36] Jonathan Schmidt, Tiago FT Cerqueira, Aldo H Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024.

[37] Nima Shoghi, Adeesh Kolluru, John R. Kitchin, Zachary W. Ulissi, C. Lawrence Zitnick, and Brandon M. Wood. From molecules to materials: Pre-training large generalizable models for atomic property prediction. *arXiv preprint arXiv: 2310.16802*, 2023.

[38] Martin Siron, Inel DJAFAR, Etienne du Fayet, Amandine Rossello, Ali Ramlaoui, and Alexandre Duval. Lemat-bulk: aggregating, and de-duplicating quantum chemistry materials databases. In *AI for Accelerated Materials Design - ICLR 2025*, 2025. URL https://openreview.net/forum?id=w0AsJpgwKq.

[39] Justin S Smith, Roman Zubatyuk, Benjamin Nebgen, Nicholas Lubbers, Kipton Barros, Adrian E Roitberg, Olexandr Isayev, and Sergei Tretiak. The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific data*, 7(1):134, 2020.

[40] Atsushi Togo, Kohei Shinohara, and Isao Tanaka. Spglib: a software library for crystal symmetry search, 2024. URL https://arxiv.org/abs/1808.01590.

[41] Oliver T. Unke, Stefan Chmiela, Michael Gastegger, Kristof T. Schütt, Huziel E. Sauceda, and Klaus-Robert Müller. Spookynet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature Communications*, 12(1), December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-27504-0. URL http://dx.doi.org/10.1038/s41467-021-27504-0.

[42] Valentin Vassilev-Galindo, Gregory Fonseca, Igor Poltavsky, and Alexandre Tkatchenko. Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules. *The Journal of Chemical Physics*, 154(9), 2021.

[43] 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. *arXiv preprint arXiv: 2306.11071*, 2023.

[44] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils E.R. Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kyle Chard, Mark Asta, Kristin A. Persson, G. Jeffrey Snyder, Ian Foster, and Anubhav Jain. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2018.05.018. URL `https://www.sciencedirect.com/science/article/pii/S0927025618303252`.

[45] Brandon M Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R Kitchin, Daniel S Levine, et al. Uma: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

[46] C. Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, Muhammed Shuaibi, Anuroop Sriram, Kevin Tran, Brandon Wood, Junwoong Yoon, Devi Parikh, and Zachary Ulissi. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv: 2010.09435*, 2020.

A Data Availability and Licensing

LeMat-Traj is publicly available at `https://huggingface.co/datasets/LeMaterial/LeMat-Traj` and is distributed under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. LeMaterial-Fetcher library, developed for the curation of LeMat-Traj, is open-source and available on GitHub at `https://github.com/LeMaterial/lematerial-fetcher`. LeMaterial-Fetcher is distributed under the Apache License 2.0.

LeMat-Traj aggregates, filters and standardizes data from the following publicly available repositories:

- **The Materials Project** [17, 18]
- **Alexandria** [35, 36]
- **The Open Quantum Materials Database (OQMD)** [34]

All data retrieved from these original sources for inclusion in LeMat-Traj are distributed under licenses compatible with CC-BY 4.0, primarily their own CC-BY 4.0 licenses. Specifically, for data originating from the Materials Project, care was taken to ensure that only structures and calculations designated under the CC-BY 4.0 license were included. We gratefully acknowledge the original creators and maintainers of these foundational datasets for making their valuable work publicly accessible.

B Distribution Analysis

Chemical diversity. To highlight the chemical diversity of the dataset, Figure 5 and 2 present periodic table heatmaps of the number of trajectories involving each element for the LeMat-Traj dataset, separately for the PBE and PBESol splits. The distribution spans nearly the entire periodic table, with particularly high representation of elements such as transition metals (e.g., Fe, Ni, Co), light elements (e.g., H, C, O, N), and main group elements (e.g., Si, Al, S). Besides oxides dominating and actinides being under-represented, the distribution is well-balanced. This ensures that the dataset is suitable for training universal machine-learned interatomic potentials that generalize across diverse chemistries and bonding environments.

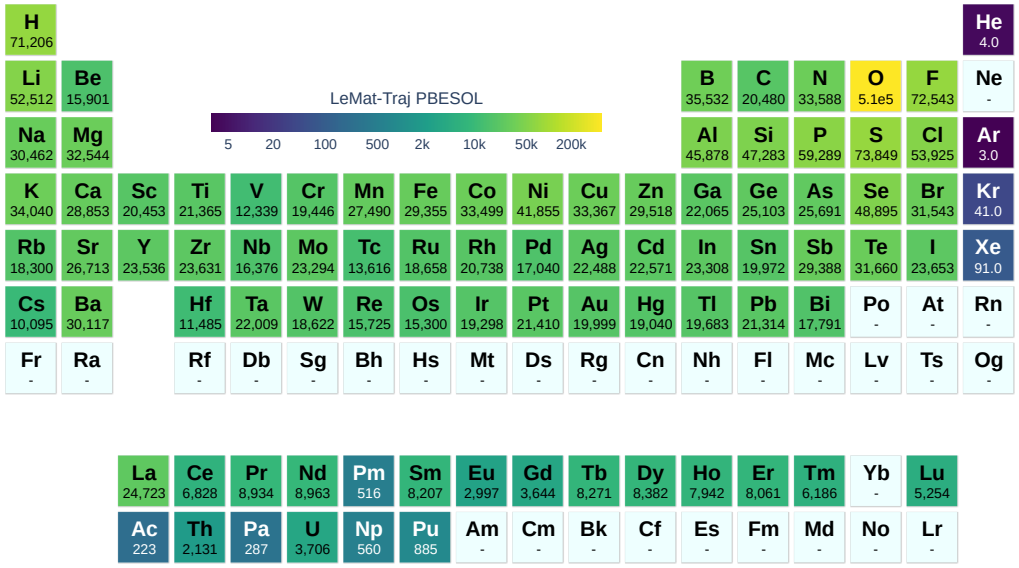


Figure 5: Chemical distribution in number of trajectories for the PBESol split.

Max Forces. Figure 6 displays the distribution of maximum atomic force norms, revealing LeMat-Traj's (PBE split) extensive coverage. It contains substantially more configurations spanning a wider range of force magnitudes (from approximately 10 to 10^3 eV/Å) compared to MPTrj and MatPES, indicating comprehensive sampling from near-equilibrium to high-force states.

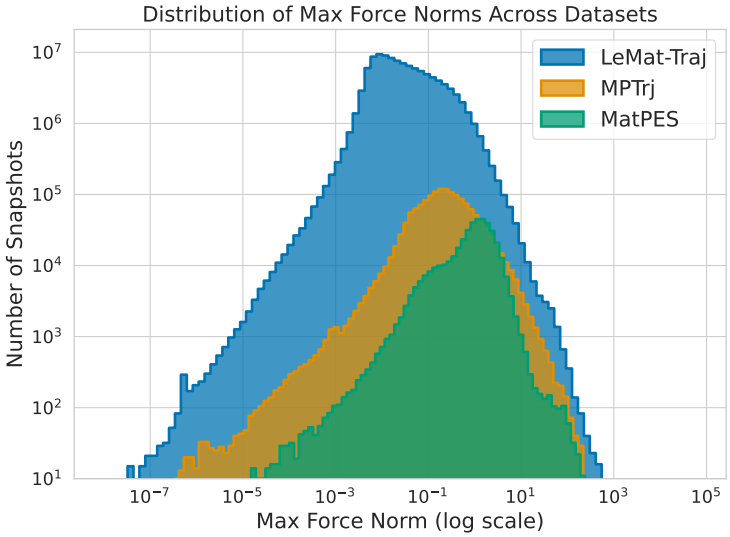


Figure 6: Coverage in log-log scale of the maximum norm of the force vector on every atomic configurations in LeMat-Traj (PBE split), MPtrj and MatPES.

Space Group diversity. To assess the structural diversity of the dataset, we analyzed the distribution of crystallographic space groups for the LeMat-Traj PBE subset. The space groups of the 120M structures were computed during the dataset creation using `moyo` a faster alternative to `Spglib` [40] in `LeMaterial-Fetcher`. The strict default parameters for space group identification (`symprec` 10^{-4}) were used in the dataset, allowing for a unified space group description across all the structures. As shown in Figure 7, the dataset spans the full range of crystal systems, including triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic groups. More than 200 unique space groups are represented, with a significant number of entries in low-symmetry systems (e.g., triclinic and monoclinic), which can be explained by the strict tolerance. This symmetry diversity is essential for training machine learning interatomic potentials (MLIPs) that generalize across materials with varying spatial constraints and bonding environments. It is also worth noting that 98% of the trajectories are assigned the same space group label at the first step of the relaxation and the last one showing the symmetry conservation during the geometric optimization calculations.

Relaxation Steps. Figure 8 illustrates the distribution of the number of geometry optimization steps performed across the first, second, and third relaxation stages within LeMat-Traj as described in section 3.1. The plots reveal that the first relaxation generally involves a broader and more varied distribution of steps, often exceeding 50 or even 100 steps for more complex or strained initial structures. In contrast, the second and third relaxations show sharply peaked distributions concentrated at lower step counts, reflecting incremental refinements of already partially relaxed geometries. This progression highlights the effectiveness of multi-stage relaxation strategies in achieving convergence, while also emphasizing that the dataset captures a wide range of relaxation behaviors—from flat minima to deep, multi-step optimization paths.

C Alternative training tasks

The trajectory data and associated metadata in LeMat-Traj support the exploration of training tasks beyond standard force and energy prediction.

Direct Structure-to-Property Prediction and Amortized Optimization. LeMat-Traj is suitable for Initial Structure to Relaxed Structure/Energy (IS2RE/IS2RS) tasks [7], as each trajectory contains the initial unrelaxed configuration, the final relaxed state, and its energy. This data structure can be used for developing *amortized optimization* methods for crystal structure relaxation [1]. In contrast to MLIPs that provide forces for an external optimizer, amortized methods attempt to learn the direct mapping from an initial structure to its relaxed state by utilizing the DFT optimization paths within the

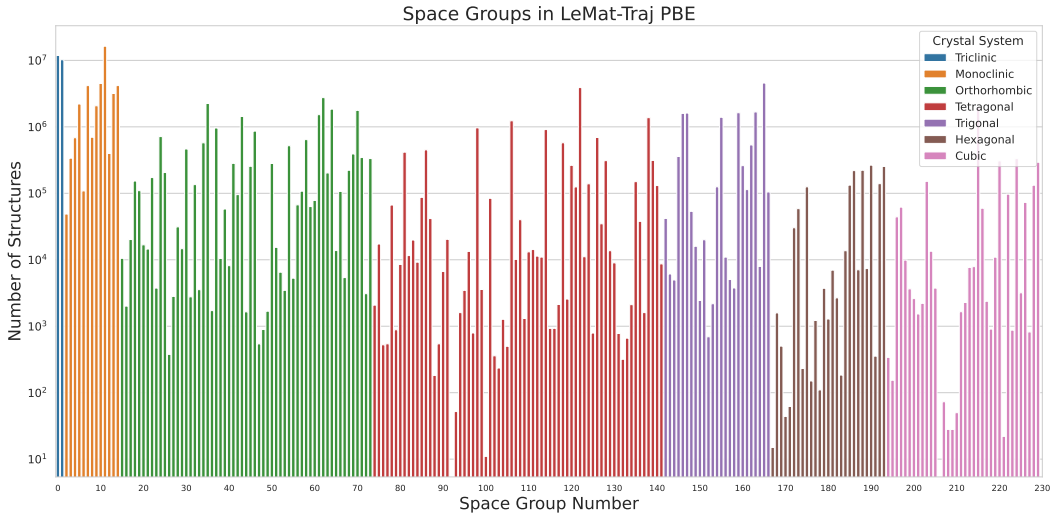


Figure 7: Distribution of space groups in LeMat-Traj (PBE subset), categorized by crystal system. The figure illustrates the number of structures for each space group on a logarithmic scale, highlighting the dataset's broad coverage of crystallographic symmetries. All seven crystal systems are represented, spanning over 200 distinct space groups.

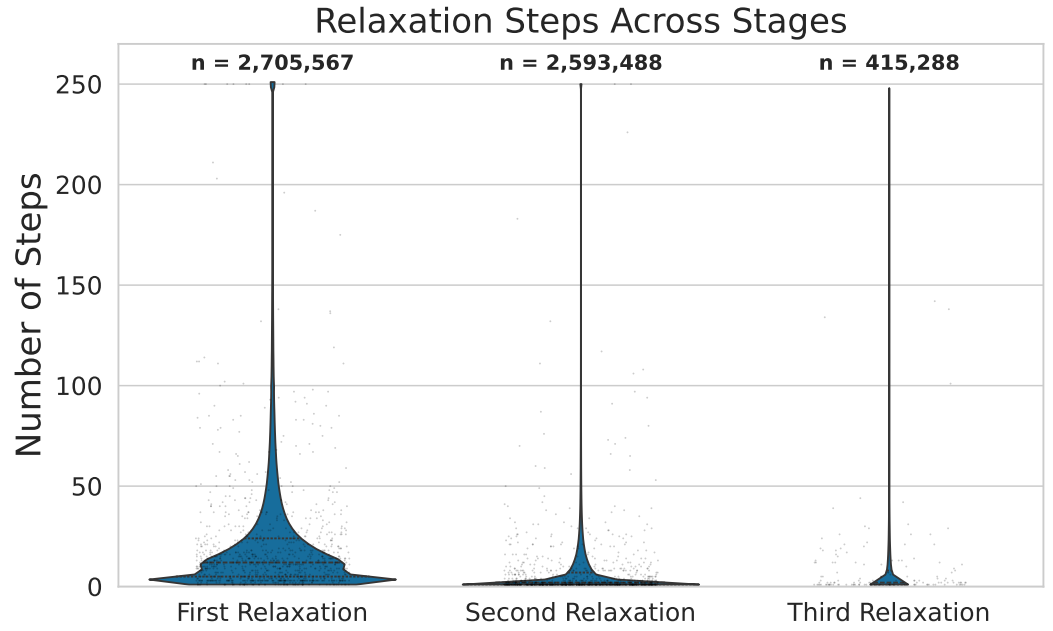


Figure 8: Number of geometry optimization steps across the first, second, and third relaxation stages in LeMat-Traj. The density of number of steps for each stage, with the total number of trajectories (n) labeled above are represented. While the first relaxation often involves more extensive structural changes, subsequent stages typically require fewer steps, indicating convergence toward optimized geometries.

dataset. Such approaches may be beneficial for applications requiring rapid structure prediction, for example, in high-throughput screening or for large systems where conventional relaxation methods can be computationally demanding [21]. While not impossible with MPtrj and Alexandria, the raw format of these datasets makes this task difficult. In contrast, the relaxation step number associated to each trajectory, and the name of the trajectory it belongs can be easily leveraged for this specific task on LeMat-Traj.

Self-Supervised Learning (SSL) for Representation Learning. The scale and diversity of LeMat-Traj also make it a relevant dataset for pre-training models using self-supervised learning (SSL) techniques [27]. The sequential information in trajectories, the relationships between different configurations along a relaxation path, and the large number of atomic configurations can serve as signals for SSL. For example, methods based on contrastive learning (e.g. DeNS [24]), masked atom or coordinate prediction, or generative pre-training (such as diffusion models, e.g. ORB [29]) could be applied. Learning to predict masked information or reconstruct parts of the input structures can help models develop general atomic representations. These representations could then be used as a starting point for fine-tuning on specific downstream tasks, potentially aiding sample efficiency and generalization, analogous to approaches in other domains like natural language processing [9]. The consistent formatting of LeMat-Traj facilitates the application of these SSL methods.

The unified format produced by LeMaterial-Fetcher allows for the distribution of LeMat-Traj via platforms like HuggingFace Datasets, providing access to the data for these training approaches.

D Experiments on LeMat-Traj

D.1 Subsets of LeMat-Traj.

In this section, we provide additional details on the way the subsets of LeMat-Traj were created and splitted for the small experiments. Due to the dataset's size, we focus on measuring performances on a few selected subsets of the dataset. The splits are available at `https://huggingface.co/datasets/LeMaterial/LeMat-Traj-subset` and can be used on more limited computational resources. Each entry represents an atomic configuration within a trajectory. To avoid data leakage, subsampling and splitting are performed at the trajectory level, ensuring all configurations from a given trajectory appear exclusively in either the training or test set. Splits are stratified based on the one-hot encoding of chemical elements present in the trajectory. This ensures no atomic species in the test set are unseen during training—essential for model generalizability. To ensure balance between the different sources for all subsets, we keep the same 10% MP, 10% OQMD and 80% Alexandria balance across all splits and all functionals, as long as the data source provides data for the functional. For SCAN and r2SCAN where the only provenance source is Materials Project, we keep all the data from the original dataset in these subset because they are small enough for these experiments and split the train and test split with a stratified 80-20% separation of the trajectories.

D.2 Cross-Dataset Generalization

The benchmarks in Section 5 highlight that combining high-force data (OMat24) with near-equilibrium data (LeMat-Traj) yields the best performance. To further explore this, we conducted a cross-dataset evaluation, testing models trained on one dataset against the test sets of others. As shown in Tables 5, 6, and 7, models consistently perform best on their in-distribution test data. For example, the model trained on OMat24 achieves the lowest errors on the OMat24 test set, but performs poorly on the LeMat-Traj test set (Table 2), and vice-versa. This reinforces our central argument: different data generation strategies (MD/active learning vs. geometry optimization) capture distinct but complementary regions of the potential energy surface. A single data source is often insufficient for creating a truly general-purpose potential. Our results demonstrate that LeMat-Traj is a crucial resource for specializing models in the low-force regime essential for accurate relaxations, complementing existing high-force datasets.

D.3 Model Training.

We report in Table 8 the hyperparameters used for training MACE. Experiments were all conducted on a single A100-40GB GPU.

Table 5: Evaluation on the MatPES PBE 10K held-out test set.

Training Dataset	Energy MAE (meV) ↓	Forces MAE (meV/Å) ↓	Forces Cos ↑
OMat24	193.8	**123.5**	0.77
MPtrj	250.2	187.5	0.70
MatPES PBE	**56.6**	127.1	**0.78**
LeMat-Traj only	245.8	217.9	0.68
OMat24 + ft LeMat-Traj	249.1	203.9	0.75

Table 6: Evaluation on the OMat24 Validation 10K test set.

Training Dataset	Energy MAE (meV) ↓	Forces MAE (meV/Å) ↓	Forces Cos ↑
OMat24	**17.9**	**103.4**	**0.99**
MPtrj	156.4	404.5	0.94
MatPES PBE	312.3	358.8	0.96
LeMat-Traj only	153.6	598.3	0.95
OMat24 + ft LeMat-Traj	218.5	395.8	0.97

E LeMaterial-Fetcher

As described in section 3.1, the pipeline to download and process the datasets is made to be both extremely customizable but also highly parallel and scalable. By default, LeMaterial-Fetcher uses `PostgreSQL` as a backend to dump the raw downloaded datasets but also to process the transformed structures before pushing them to HuggingFace. Other backends are supported and easy to integrate in the library, with for example `MySQL` being used for OQMD (the source dataset from their website is a full database with scattered tables). One of the main challenges with writing this pipeline was allowing for full parallelization to decrease the time from download to pushing the unified dataset. Indeed, having multiple connections opened for both fetching data from a table and pushing them to the other one with database cursors is prone to high memory usage and leakage. Naive implementations of parallelism do not allow to fully take advantage of high compute machines. To that end, we designed the library to be very memory-efficient. For LeMat-Traj, it was possible to take advantage of 128 cores with 256GB without any issue. The entire pipeline to create LeMat-Traj took around 16 hours to create the 120M rows and upload them on HuggingFace running with 12 workers on an AMD Ryzen 5600G. This time gets significantly reduced when running on larger machine on which we are able to max-out the usage.
For the dataset curation process, we follow the same procedure as [38] with the exception that we pick Ytterbium (Yb) containing samples from Materials Project rather than Alexandria because of the non-compatibility between their pseudo-potentials.

Materials Project. For the Materials Project data transformation process, we look through every single task available (around 1.5M at the latest release during the first LeMat-Traj version), and then only keep the non-deprecated tasks. To ensure accurate sampling of the PES, we pick all the trajectories for a given material as long as they pass the data filtering described in 3.3.

Alexandria. All samples from Alexandria were used except for the ones containing Yb.

OQMD. OQMD trajectories are obtained by going through all the entries of the OQMD database, gathering their associated calculations from *relaxation*, *coarse relaxation* and *fine relaxation* for every relaxation stage. The input structures and output structures are then processed, provided they contain the targets expected in the right format.

F Potential Energy Surfaces

To visualize the coverage of the potential energy surface (PES) by LeMat-Traj, we projected atomic configurations onto a lower-dimensional space derived from Smooth Overlap of Atomic Positions

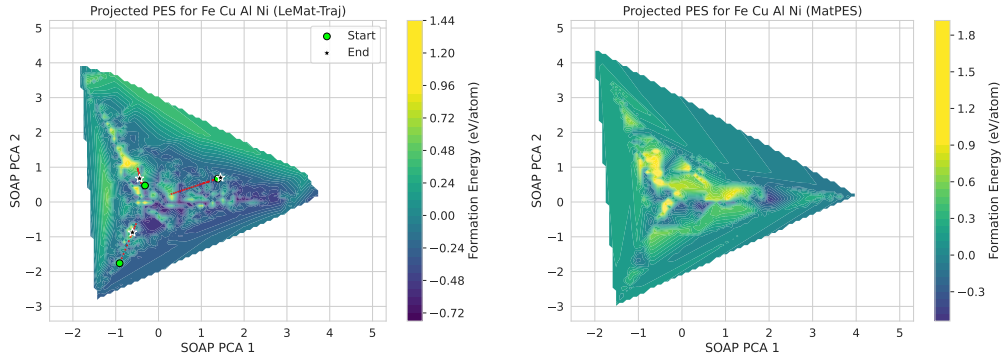
Table 7: Evaluation on the MPtrj 10k held-out test set.

Training Dataset	Energy MAE (meV) ↓	Forces MAE (meV/Å) ↓	Forces Cos ↑
OMat24	58.7	68.7	**0.54**
MatPES PBE	237.6	114.6	0.36
LeMat-Traj only	**20.2**	**63.3**	0.52
OMat24 + ft LeMat-Traj	37.3	73.4	0.52

Table 8: Hyperparameters used to train MACE on the subsets of LeMat-Traj.

Hyperparameter	Training Stage 1	Training Stage 2	Fine-tuning
Learning Rate	8e-4	8e-4	8e-4
Scheduler	Constant	Constant	Constant
Batch Size	128	128	128
Energy Weight	1	100	1
Force Weight	10	100	100
Stress Weight	1	1	1

(SOAP) descriptors [15]. Figure 9 illustrates this for the systems in the metallic Fe-Cu-Al-Ni hull within the PBE functional subset of LeMat-Traj, contrasting it with a similar projection for the MatPES dataset. LeMat-Traj projection (9(a)) reveals a broad exploration of the PES, with example trajectories (red lines) originating from diverse initial high-energy states (green circles) and converging towards distinct low-energy minima (black stars). The gradient energy gradient is clearly visible in the line levels far from the very high energy regions. This visualization is also very similar with the MatPES projection (9(b)) which, while also covering a significant area, appears to have a different structural sampling emphasis, with less granularity around maxima, revealing a smaller number of saddle points. Further details on the visualization methodology are provided in Appendix F.

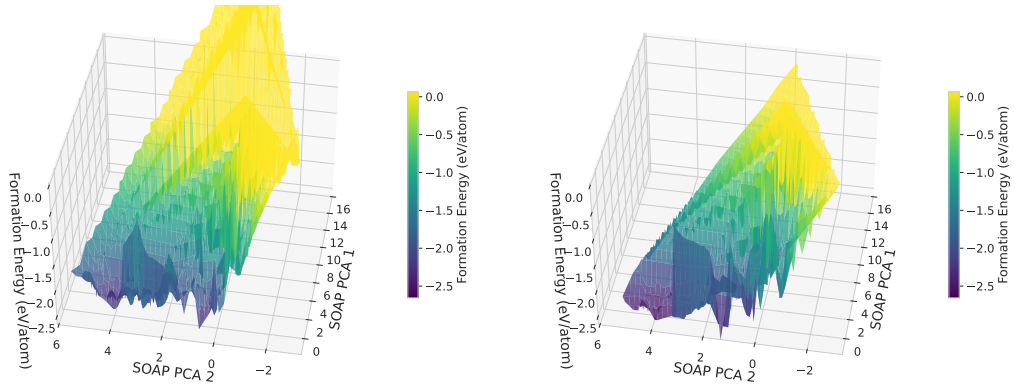


(a) LeMat-Traj PBE (b) MatPES

Figure 9: Projected Potential Energy Surfaces (PES) for the metallic Fe-Cu-Al-Ni systems. Atomic configurations are featurized using SOAP descriptors [15] and projected onto their first two principal components. The PCA 1 and PCA 2 axes are qualitative representations of structural similarity and do not have a direct physical interpretation. Color indicates formation energy (eV/atom). **(a)** PES derived from the LeMat-Traj PBE dataset. Green circles and black stars mark initial and final structures of example trajectories (red lines). The visualization highlights LeMat-Traj's dense, high-frequency sampling of the PES, which is crucial for resolving fine details near energy minima. **(b)** PES derived from the MatPES dataset, showing a broader but sparser sampling of the overall landscape.

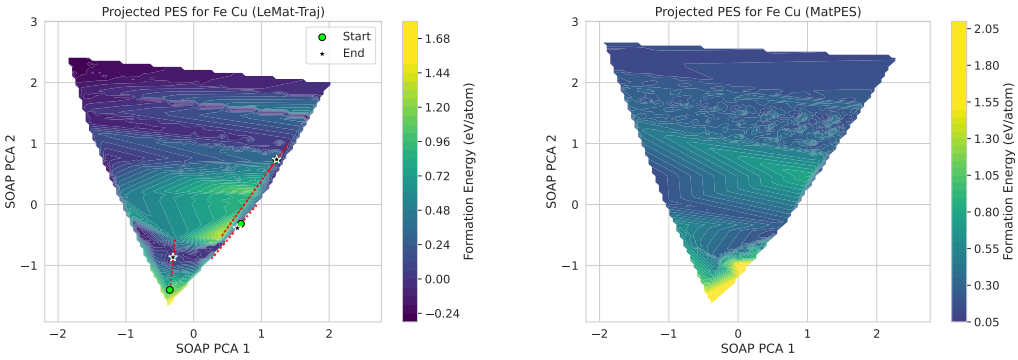
To allow for easier interpretability we limit the analysis to specific coherent subsets of chemical elements (metallic or ionic). For every dataset, all the atomic configurations whose chemical formula is a subset of the chosen elements are gathered. Then SOAP descriptors are computed for all these

configurations with the same hyperparameters (`r_cut` = 5.0, `n_max` = 8 and `l_max` = 6, with `outer` averaging to get a vector for every structure). All of these SOAP vectors are used to fit a PCA and the formation energy per atom (eV/atom) is computed. Because the sampling of atomic configurations is scattered across the PCA space and not continuous, we use a linear interpolation of the convex hull to get this visual description. Figure 10 illustrates the PES of a different chemical subset, highlighting the close similarity between LeMat-Traj and MPtrj. Indeed, since MPtrj is contained in LeMat-Traj, the PES of the latter describes local minima and transition pathways with a higher resolution. Additionally, when only limiting the sampling to two elements systems with Fe-Cu, we notice the advantages of having a larger structural configuration sampling to better describe the entire PES. Although having a smaller dataset may result in a smoother landscape that might help models converge faster and more easily, it is not enough to completely capture the large number of local energy minima that exist in the complex DFT force field.



(a) LeMat-Traj PBE (b) MPtrj

Figure 10: Projected Potential Energy Surfaces (PES) for the ionic Na-Cl-O systems for LeMat-Traj and the MPtrj datasets, similar to Figure 9 in 3D projection.



(a) LeMat-Traj PBE (b) MatPES

Figure 11: Projected Potential Energy Surfaces (PES) for the subset Fe-Cu systems for LeMat-Traj and the MPtrj datasets, similar to Figure 9.