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.

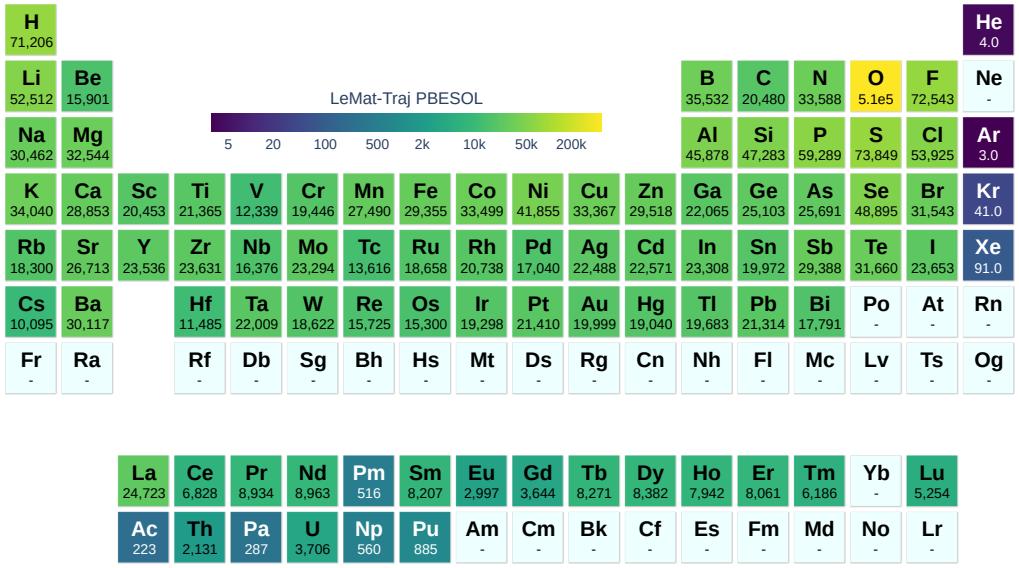


Figure 5: Chemical distribution in number of trajectories for the PBESol split.

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^{-7} to 10^{3} eV/Å) compared to MPTrj and MatPES, indicating comprehensive sampling from near-equilibrium to high-force states.

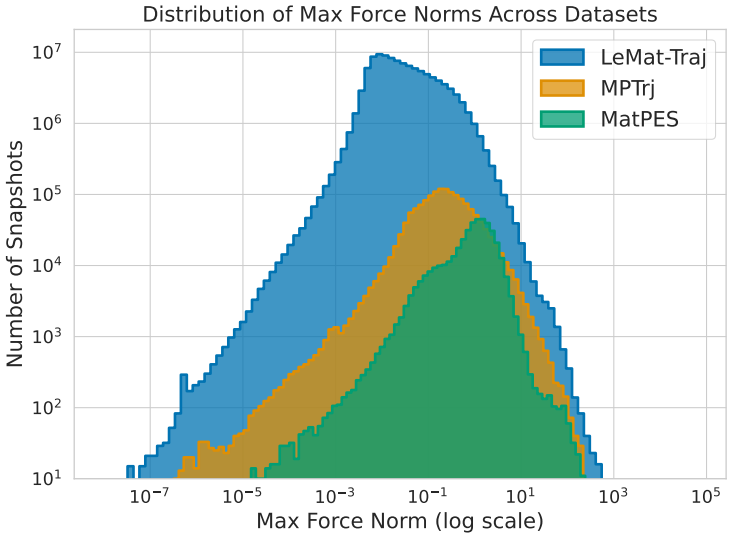


Figure 6: Coverage in log-log scale of the maximum norm of the force vector on every atomic configurations in LeMat-Traj (PBE split), MPtrj and MatPES.

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

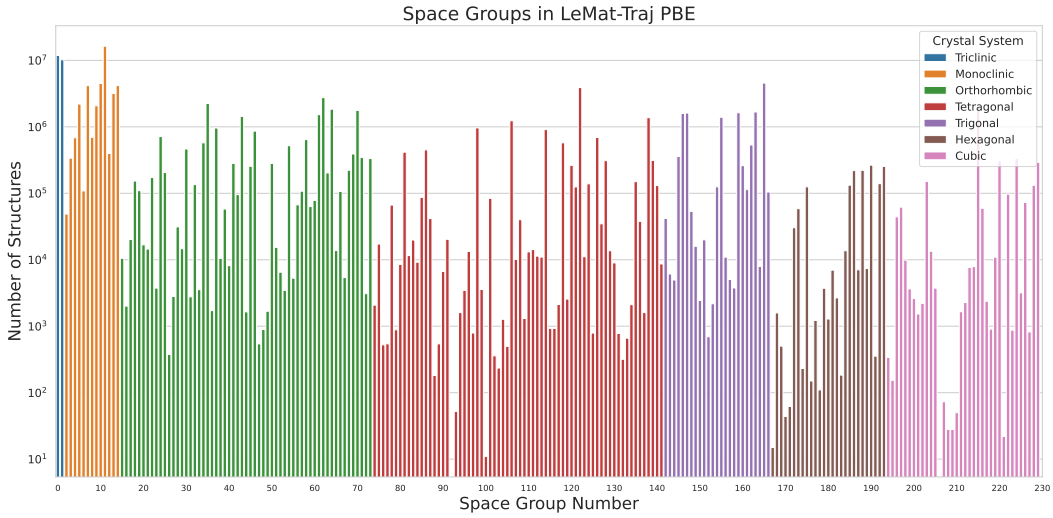


Figure 7: Distribution of space groups in LeMat-Traj (PBE subset), categorized by crystal system. The figure illustrates the number of structures for each space group on a logarithmic scale, highlighting the dataset's broad coverage of crystallographic symmetries. All seven crystal systems are represented,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

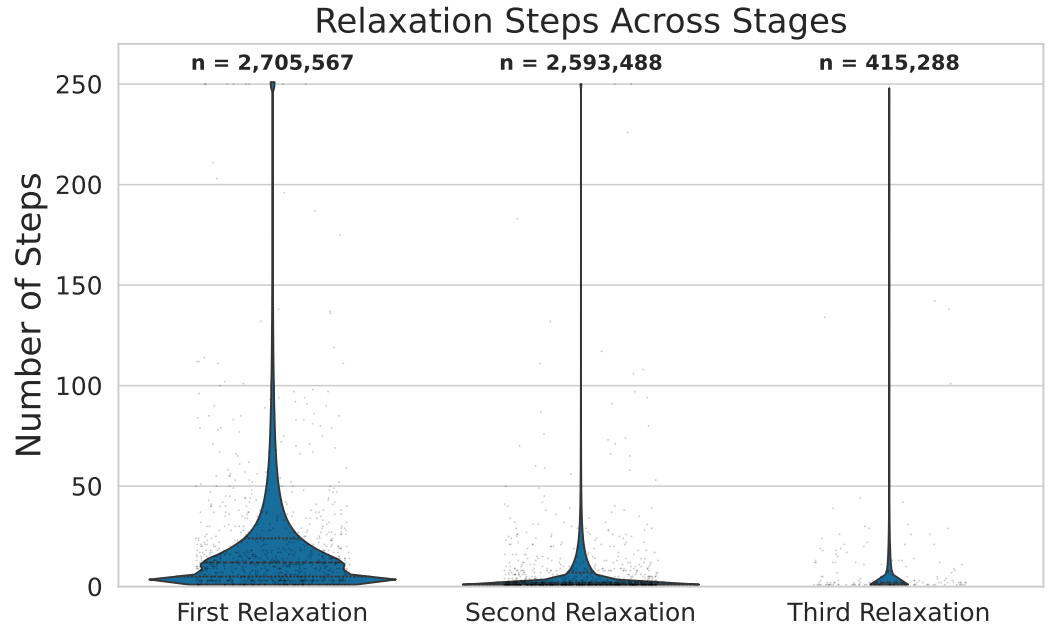


Figure 8: Number of geometry optimization steps across the first, second, and third relaxation stages in LeMat-Traj. The density of number of steps for each stage, with the total number of trajectories (n) labeled above are represented. While the first relaxation often involves more extensive structural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

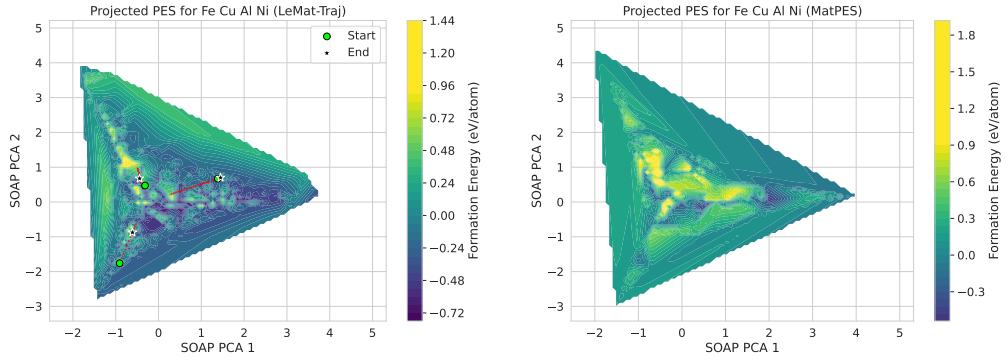
Table 7: Evaluation on the MPtrj 10k held-out test set.

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.

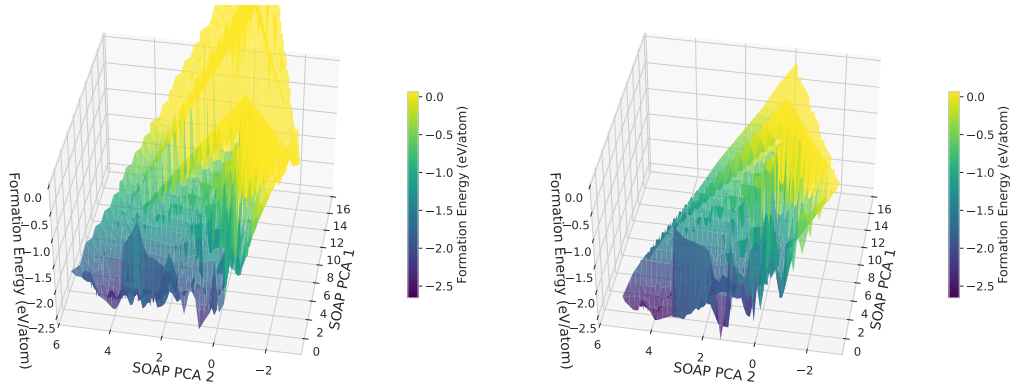


(a) LeMat-Traj PBE (b) MatPES

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.

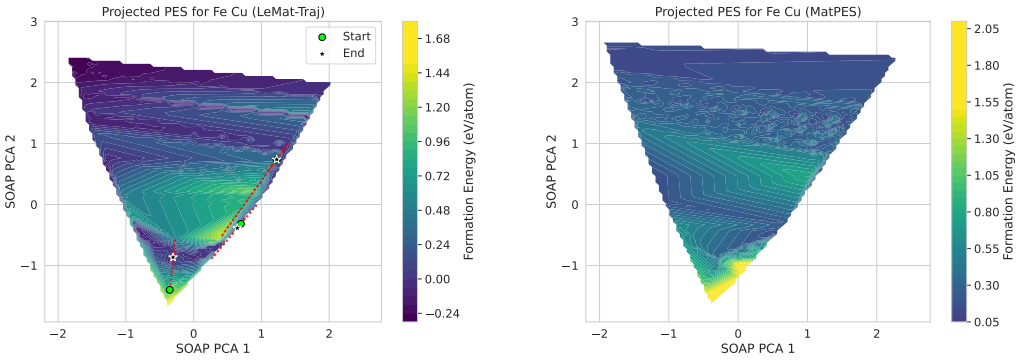
To allow for easier interpretability we limit the analysis to specific coherent subsets of chemical elements (metallic or ionic). For every dataset, all the atomic configurations whose chemical formula is a subset of the chosen elements are gathered. Then SOAP descriptors are computed for all these

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.