Pre-training, Fine-tuning, and Distillation (PFD): Automatically Generating Machine Learning Force Fields from Universal Models

Ruoyu Wang[1,2], Yuxiang Gao[1,2], Hongyu Wu[3], Zhicheng Zhong[1,2,3,*]

[1]School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei 230026, China

[2]Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou 215123, China

[3]Suzhou Lab, Suzhou 215123, China

[*]zczhong@ustc.edu.cn

I. DETAILED METHODS

A. Entropy filtering

QUESTS method

The descriptor-based QUESTS (Quick Uncertainty and Entropy from STructural Similarity) method[1] is employed to evaluate the similarity between local atomic environments encoded by a simple vector descriptor. The descriptor for atom i consists of two parts, the first part $X_i^{(1)}$ captures bond information,

$$X_i^{(1)} = [\frac{\omega(r_{i1})}{r_{i1}}...\frac{\omega(r_{ik})}{r_{ik}}]^T, r_{ij} \leq r_{i(j+1)} \tag{1}$$

where $\omega(r) = [1 - (\frac{r}{r_c})^2]^2 \quad (0 \leq r \leq r_c) \quad \text{or} \quad 0 \quad (r > r_c)$ is a smooth function. The second part $X_i^{(2)}$ supplements $X_i^{(1)}$ by accounting for bond angle information. The n^{th} element of $X_i^{(2)}$ is,

$$X_{in}^{(2)} = \langle \frac{\sqrt{\omega(r_{ij})\omega(r_{il})}}{r_{jl}} \rangle_n, j, l \in N(i), X_{in} \geq X_{i(n+1)} \tag{2}$$

where $\langle\cdot\rangle_n$ represents the average for the n^{th} element of $X_{ij}^{(2)'}$,

$$X_{ij}^{(2)'} = (\frac{\sqrt{\omega(r_{ij})\omega(r_{i1})}}{r_{j1}}, ..., \frac{\sqrt{\omega(r_{ij})\omega(r_{ik})}}{r_{jk}}) \tag{3}$$

The final descriptor X_i is constructed by concatenating $X_i^{(1)}$ and $X_i^{(2)}$. The structural diversity of a dataset is measured by its entropy,

$$\mathcal{H}(\{X\}) = -\frac{1}{n}\sum_{i=1}^{n} \log[\frac{1}{n}\sum_{j=1}^{n} K_h(X_i, X_j)], \tag{4}$$

where K_h is a Gaussian kernel function,

$$K_h(X_i, X_j) = \exp(\frac{-||X_i - X_j||^2}{2h^2}) \tag{5}$$

and the h effectively controls the smearing bandwidth. The maximum entropy of a given data set with n atomic environments is log$n$ when $K_h(X_i, X_j) = \delta_{ij}$. The entropy would be zero if $K_h(X_i, X_j) = 1, \forall i, j$. Similarly, a differential entropy (dH) can be defined that quantifies how much new structural information a test frame Y contains relative to the fine-tuning dataset $\{X_i\}$,

$$d\mathcal{H}(Y|\{X\}) = -\log[\sum_{i=1}^{n} K_h(X_i, X_j)]. \tag{6}$$

In PFD workflow, a default Gaussian smearing bandwidth of 0.015 Å is set as in the original QUESTS paper.

Iterative algorithm for frame selection

Building on the QUESTS approach, we developed an iterative algorithm to select new frames that are maximally dissimilar from previously sampled data. The pseudo-code for this algorithm is presented below. When entropy-based filtering is employed, the algorithm is applied at each iteration of the fine-tuning process. Figure S1 compares the efficiency of the two selection strategies: the original random selection and the proposed entropy-based selection.

Algorithm 1 Iterative Frame Selection by Differential Entropy

Require: F ▷ set of all candidate frames of the i^{th} iteration

Require: D_0 ▷ initial dataset

Require: chunk_size ▷ frames selected per iteration

Require: N_{target} ▷ total number of frames to select

Ensure: D ▷ final selected frames

$D \leftarrow \varnothing$

while $|D| < N_{\text{target}}$ **do**

  $R \leftarrow F \setminus D$ ▷ remaining frames to evaluate

  for all $f \in R$ **do**

    $dH(f) \leftarrow \text{DifferentialEntropy}(f, D)$

  end for

  $S \leftarrow$ top chunk_size frames with largest $dH(f)$

  $D' \leftarrow D \cup S$ ▷ updated dataset after selection

  $D \leftarrow D'$

end while

return D

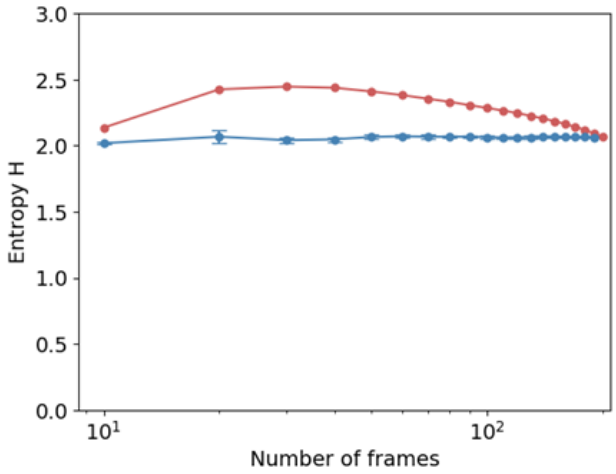

FIG. S1. Comparison of selection methods. We performed a short MD simulation of diamond Si at 500 K and 1 GPa for 4 ps (2000 steps, 2 fs timestep), sampling frames every 10 steps. The figure compares the atomic entropy of frames selected randomly versus those chosen using the entropy filtering algorithm implemented in PFD-kit. The results demonstrate that the algorithm selects frames with more diverse atomic environments, improving the effectiveness of the selection process.

II. COMPUTATION DETAILS

A. Hardware specifications

All the tests are carried out using the Bohrium online platform. The type of GPU devices used for model training and molecular dynamics (MD) simulation with DPA-2 and DeePMD model is Nvidia V100 card of 32 GB memory; the CPU nodes used for DFT calculations have 32 CPU cores. The time consumption of the PFD workflow is estimated using the above hardware specifications and assuming parallel execution of DFT and MD simulations.

B. DFT calculation

All first-principle density functional theory (DFT) data are generated within the VASP package using the GGA (generalized gradient approximation) functional of the Perdew-Burke-Ernzerhof (PBE) form and plane-wave basis with energy cutoff of 500 eV. k-point mesh with 0.2 $Å^{-1}$ spacing is constructed for all calculations, excluding the Γ-point. An energy convergence criteria of $1x10^{-5}$ eV is selected, along with a Gaussian smearing of 0.05 eV to aid convergence.

C. Model training

The "2.3.1-rc0-medium" version of pre-trained DPA-2 model is selected as the foundation model. The DPA-2 descriptor combines a local part "repinit" and several layers of message-passing-like "repformer". The descriptor is first initialized by the repinit module, which has a cutoff of 6 Å. Then six repformer layers, each with a cutoff of 4 Å, iteratively update the descriptor in a message-passing manner. In fine-tuning, the descriptor is initialized by the pre-trained model weight file. The predicting head is randomly initialized, which is a MLP network with three hidden layers. During fine-tuning, the pre-trained DPA-2 model is refined for 20,000 steps, with a small learning rate of 0.001 that decays to 3.51×10^{-8} at an interval of 200 steps.

The distilled model is standard DeePMD model with the "se_atten_v2" local descriptor. No attention layer is activated, and a cutoff of 6.5 Å is selected. All distilled models are trained for 1,000,000 steps using a learning rate decaying from 0.01 to 1×10^{-8} at an interval

of 2000 steps. The batch size is set as 1.

D. Phonon calculation

All phonon dispersions are calculated by the finite displacement method using the phonopy package[2, 3]. To calculate the second force constant and its Fourier transform, $3 \times 3 \times 3$ supercells with displacements are generated by the phonopy, and the atomic forces subsequently calculated with DFT or machine learning force fields.

III. SUPPLEMENTARY RESULTS

A. Data efficiency

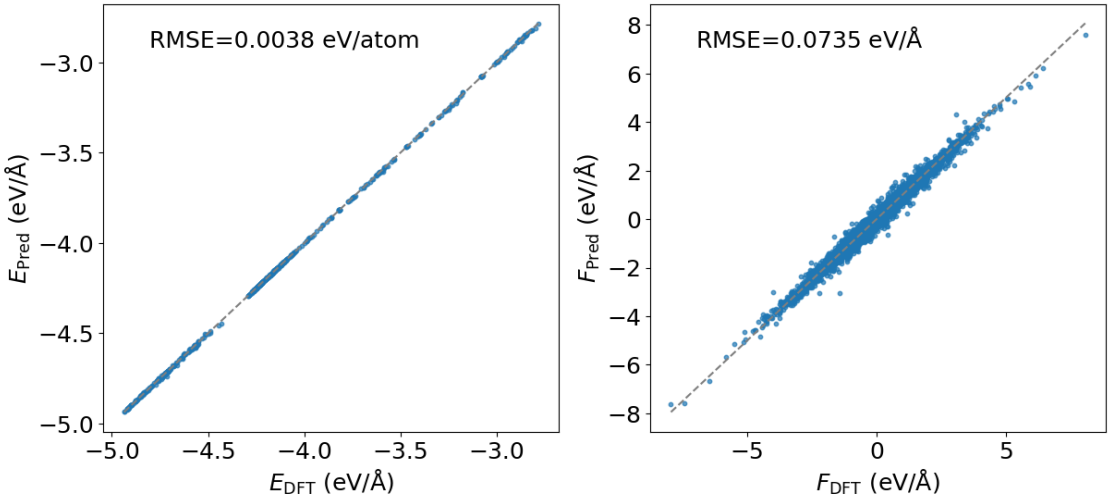

FIG. S2. The energy and force prediction accuracy of the fine-tuned model for argyrodite electrolyte Li_6PS_5X (X=Cl,Br,I) and its subsystems using 500 data frames randomly extracted from the dataset.

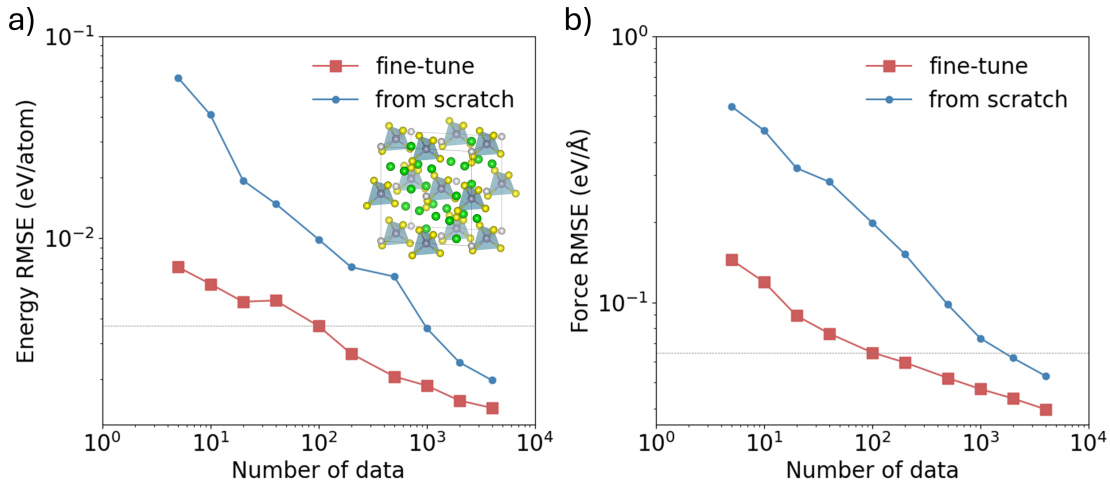

FIG. S3. The energy and force prediction accuracy of the fine-tuned model specifically for Li_6PS_5Cl.

B. Crystal Si

The exploration for fine-tuning data was carried out in three sequential stages. At each stage, new configurations are explored by MD trajectories at 500 K and pressures of 0.001 and 1 GPa. Each MD exploration task runs for 2000 steps with 2 fs timestep, and candidate configurations are extracted from the MD trajectories every 100 steps. Table S1 lists the number of data collected from each stage. In total, the composition of the final fine-tuning dataset is shown in Table S2.

TABLE S1. Fine-tuning exploration stages

	Si phases	Num. frames	Num. iterations
Init	dia., hex. dia., FCC, BCC, HCP, β-Sn and dia. vac.	52	—
Stage 1	dia., hex. dia.	80	2
Stage 2	FCC, BCC, HCP, β-Sn	346	9
Stage 3	dia. vac.	37	2

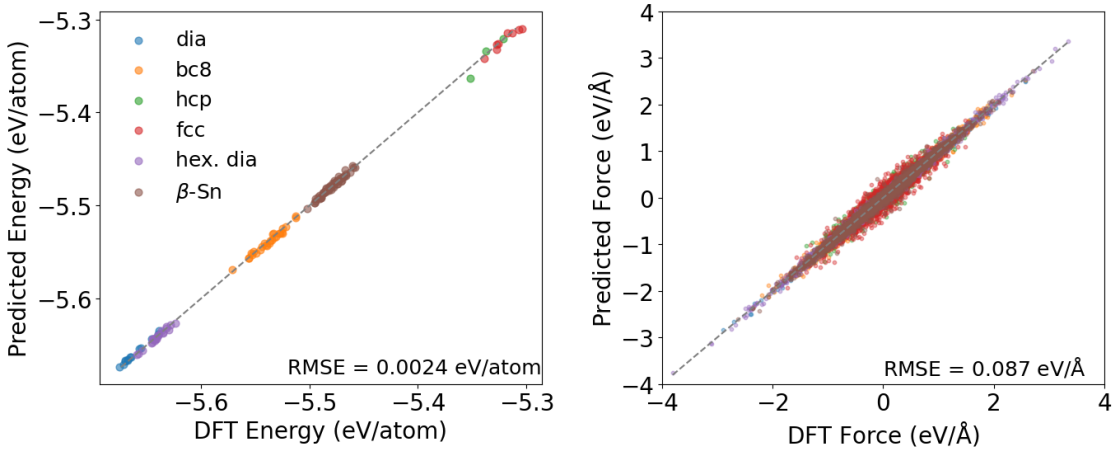

FIG. S4. The energy and force prediction accuracy of the distilled Si model.

The vacancy defect formation energy is calculated by extracting a single Si atom from the supercell of various silicon phases. The energy of diamond Si is used as a reference for Si chemical potential. The atom numbers of the supercells are 32 (diamond), 128 (BC8), 64

TABLE S2. Composition of the final crystal Si fine-tuning data

Phases	Number of frames	Number of atoms
Diamond	43	32
	3	8
Hexagonal diamond	49	32
	3	4
FCC	101	32
	3	4
BC8	70	32
	41	16
HCP	92	16
	3	2
β-Sn	61	32
	3	4
Diamond vacancy	43	31
Total	515	31

(β-Sn), 108 (FCC), 108 (hexagonal diamond) and 64 (HCP), respectively.

C. LATP solid electrolyte

The training data for $Li_{1+x}Al_xTi_{2-x}(PO_4)_3$ (LATP) is constructed from $2 \times 2 \times 1$ supercell of the $LiTi_2(PO_4)_3$ unit cell (space group $R\bar{3}C$) of 36 atoms. LATP structures of various Li compositions and Al substitution are built using the Pymatgen package.

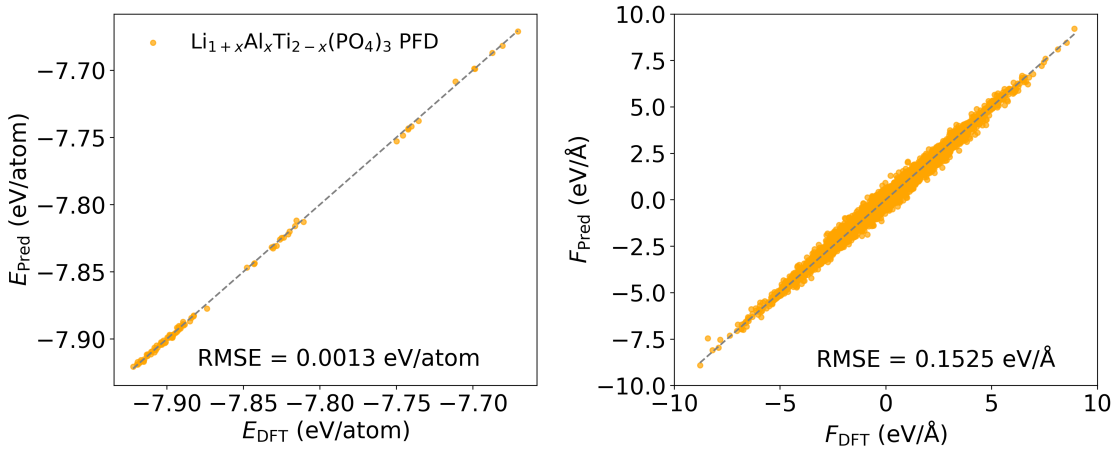

FIG. S5. Energy and force prediction the distill model for $Li_{1+x}Al_xTi_{2-x}(PO_4)_3$ solid electrolyte.

The mean square displacement (MSD) of lithium ions after time period t can be calculated as $\mathrm{MSD}(t) = \frac{1}{N}\sum_i^N |r_i(t) - r_i(0)|^2$, where N is the number of lithium ions. Assuming Brownian motion, the diffusion coefficient D of lithium ions in 3-dimensional space can be deduced as $D = \lim_{t\to\infty} \frac{\mathrm{MSD}(t)}{6t}$. Ion conduction is a thermally activated process, and the activation energy E_a can be extrapolated by fitting the modified Arrhenius relation to the D at various temperatures, $D(T) = D_0 \exp(\frac{-E_a}{k_B T})$.

The small simulation cells for $Li_{1.3}Al_{0.3}Ti_{1.7}(PO_4)_3$ are constructed from the $2 \times 2 \times 1$ supercell of $LiTi_2(PO_4)_3$, which contains 110 atoms. This supercell is further enlarged by $8 \times 8 \times 2$ to construct a large simulation cell containing 3520 atoms.

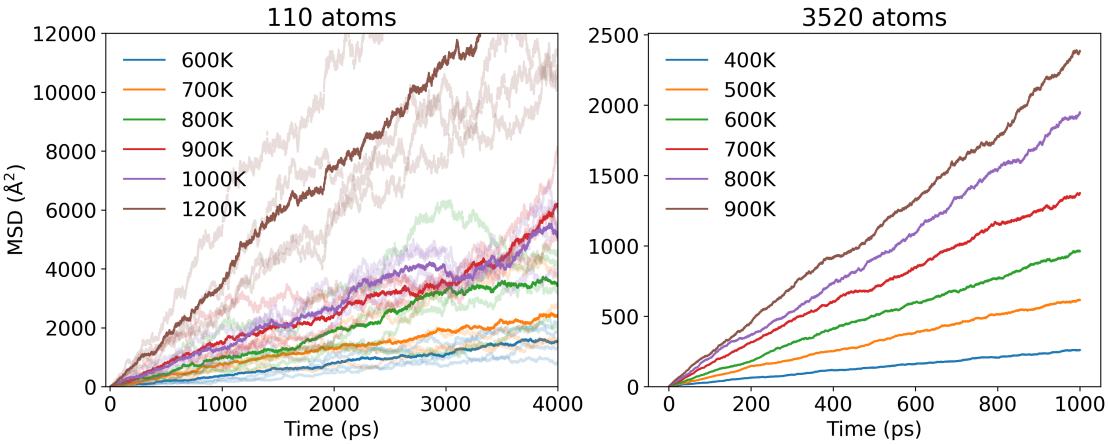

FIG. S6. Li-ion mean square displacement (MSD) trajectory for $Li_{1+x}Al_xTi_{2-x}(PO_4)_3$ (LATP) simulation system with a) 110 and b) 3520 atoms, respectively. In the case of 110 atom cell, the MSD trajectory is averaged on four independent simulations, which are represented by the translucent lines in the image.

D. Molecular chains

For the example of 1,4-polyisoprene molecular chains, each chain is made up of carbon and oxygen atoms with a total number of atoms ranging from 40 ($C_{20}H_{20}$) up to 180 ($C_{60}H_{120}$). The test set is constructed by clusters of several molecular chains. All DFT calculations in this example include the vdW dispersion energy-correction term in the DFT-D3 form to account for the van-der Waals interaction.

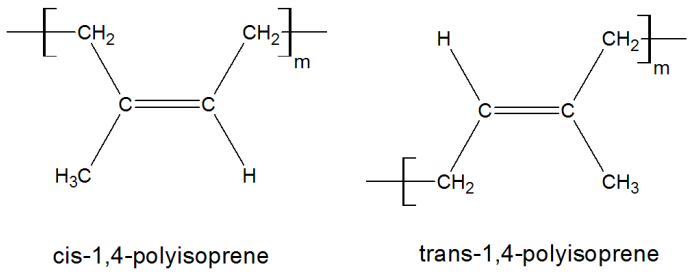

FIG. S7. The chemical formula of cis- and trans-1,4-polyisoprene chains.

E. Amorphous carbon

The amorphous carbon structures with sp^3 hybridization are randomly generated using the RG^2 package[4], and the number of carbon atoms in the unit cell ranges from 8 up to 84.

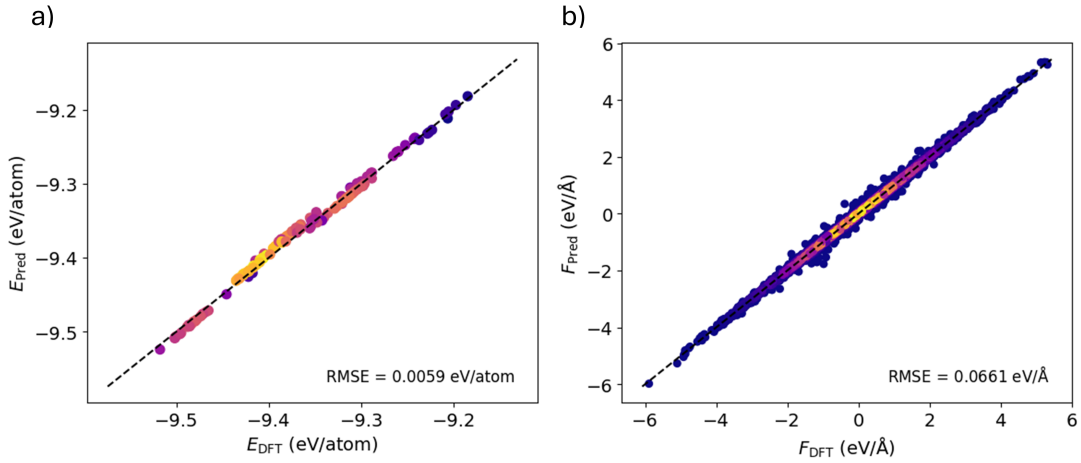

FIG. S8. **Fine-tuned model for amorphous carbon generated using PFD workflow.** The **a)** energy and **b)** force prediction error of the fine-tuned model on amorphous carbon of sp^3 tetrahedral local geometry.

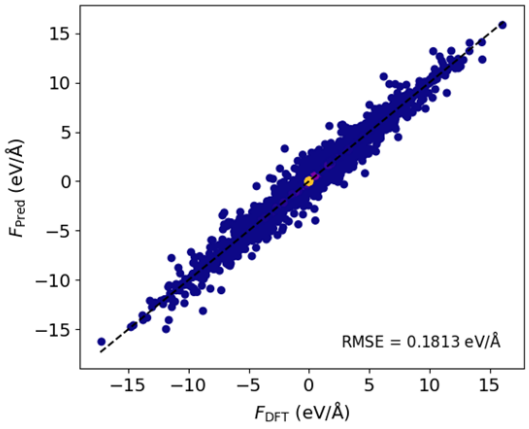

FIG. S9. Force prediction error of the amorphous carbon PF model .

F. Heavily doped perovskite

The training set for high entropy perovskite $Ba_{0.5}Na_{0.25}Bi_{0.25}Ti_{0.875}Zr_{0.125}O_3$ is constructed by random substitution in a $2 \times 2 \times 2$ supercell of cubic $BaTiO_3$ of 40 atoms.

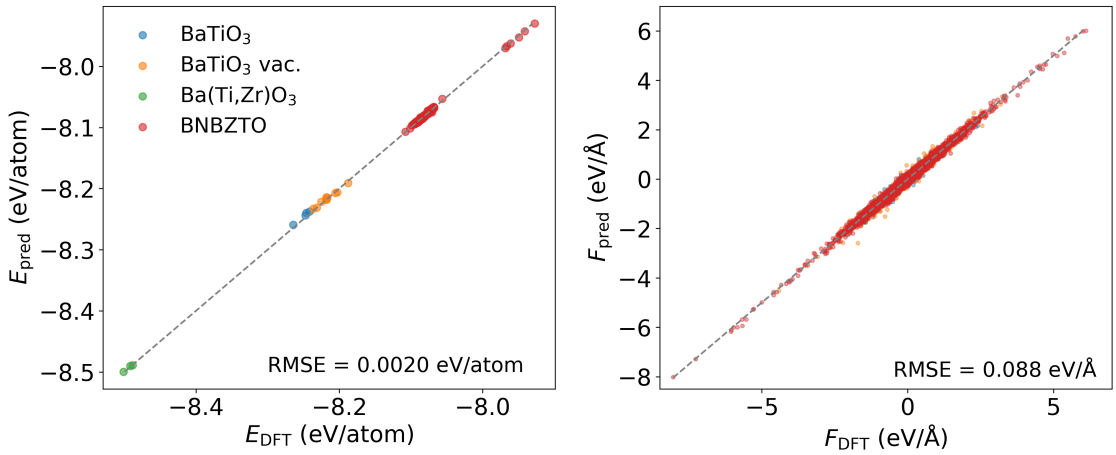

FIG. S10. Energy and force prediction accuracy of the PFD model for the various $BaTiO_3$ perovskite phases as well as the heavily doped $Ba_{0.5}Na_{0.25}Bi_{0.25}Ti_{0.875}Zr_{0.125}O_3$.

The oxygen vacancy defect formation energy is calculated by extracting a single O atom from the $2 \times 2 \times 2$ supercell of various perovskite phases. The energy of O_2 is used as a reference for O chemical potential. The atom numbers for perovskite supercells are 40 (cubic), 80 (orthorhombic), 40 (tetragonal) and 80 (rhombohedral), respectively.

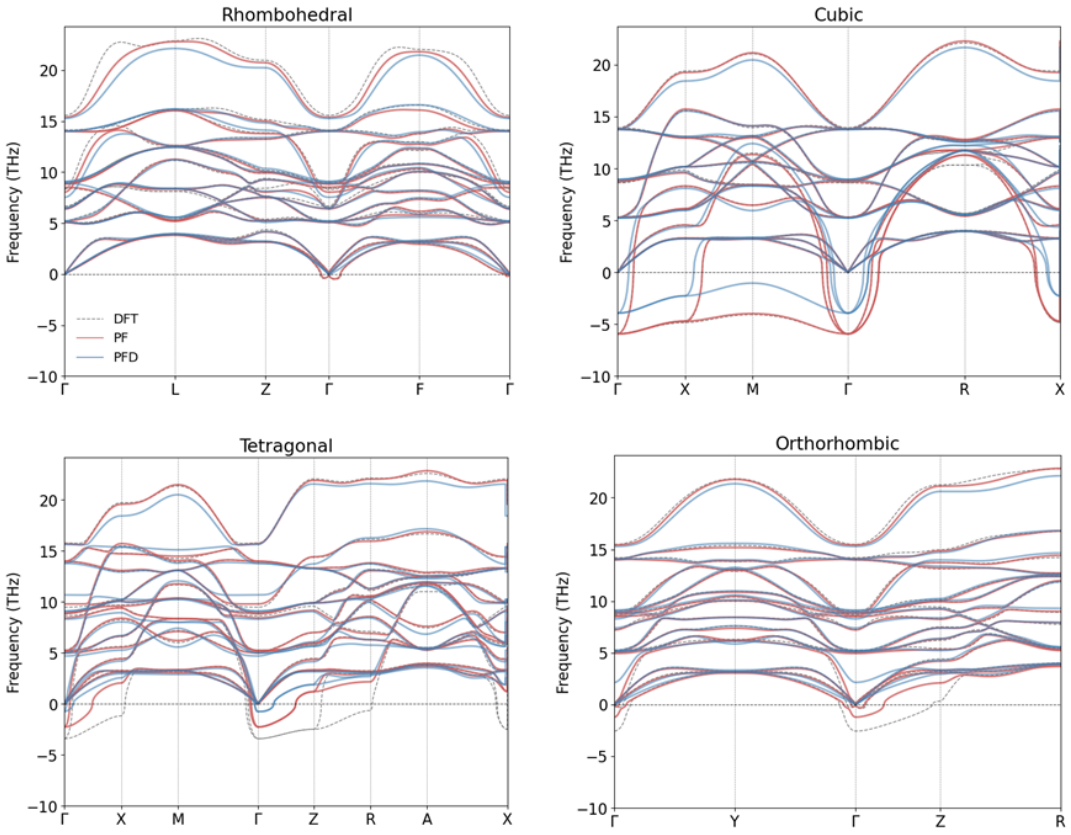

FIG. S11. Phonon dispersion of various $BaTiO_3$ perovskite phases.

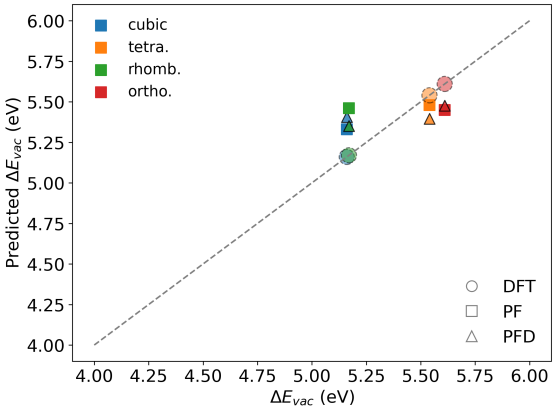

FIG. S12. The vacancy formation energy of various perovskite phases calculated with the PF and PFD models.

G. The Li/Li_6PS_5Cl interface model

The initial structures of the Li_6PSCl_5/Li interfaces are made by first concatenating the Li crystal and the bulk Li_6PSCl_5 structure and then running structural optimization. The final fine-tuning dataset, with a total of 409 data frames, consists of 96 frames of metal Li each with 16 atoms, 85 frames of Li_6PSCl_5 each with 52 atoms and 228 frames of Li_6PSCl_5/Li interfaces each with 106 atoms.

The QUESTS (Quick Uncertainty and Entropy from STructural Similarity) method[**?**] is employed to evaluate the similarity between the fine-tuning and test datasets for the Li/Li_6PS_5Cl interface model, with a Gaussian smearing bandwidth of 0.015 Å as in the original QUESTS paper. Figure S13 illustrates the correlation between dH and atomic force prediction errors for each atom in the test dataset. Most atomic environments in the Li/Li/Li_6PS_5Cl interface test data are similar to those in the fine-tuning dataset, as indicated by low dH values. Notably, many environments with high dH (>2) still exhibit force errors below 0.1 eV/Å. This suggests that the foundation model's knowledge, transferred during fine-tuning, enhances the model's ability to generalize to less familiar atomic configurations, contributing to the robust description of the amorphous interface.

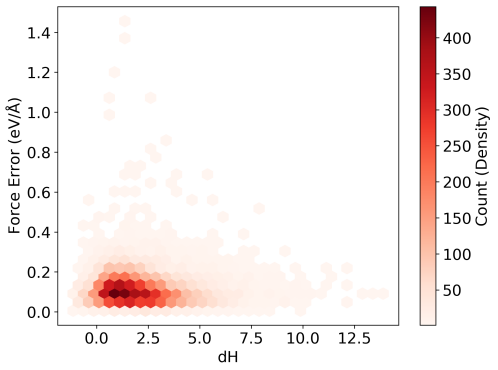

FIG. S13. Correlation between the differential entropy and the force prediction error by the PF model for each atonic environment in the Li/Li_6PS_5Cl interface test data.

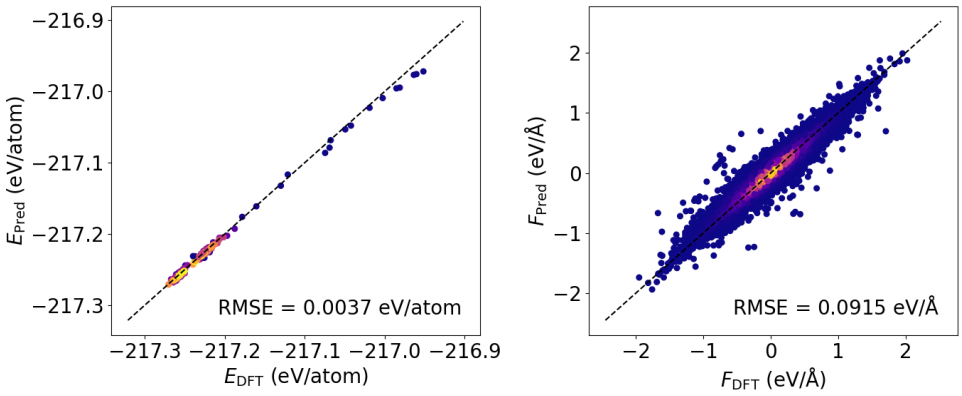

FIG. S14. Energy and force accuracy of the distilled model for the Li_6PSCl_5/Li interface model.

[1] Daniel Schwalbe-Koda, Sebastien Hamel, Babak Sadigh, Fei Zhou, and Vincenzo Lordi. Model-free estimation of completeness, uncertainties, and outliers in atomistic machine learning using information theory. *Nature Communications*, 16(1):4014, April 2025.

[2] Atsushi Togo and Isao Tanaka. First principles phonon calculations in materials science. *Scripta Materialia*, 108:1–5, November 2015.

[3] Atsushi Togo, Laurent Chaput, Terumasa Tadano, and Isao Tanaka. Implementation strategies in phonopy and phono3py. *Journal of Physics: Condensed Matter*, 35(35):353001, September 2023. Publisher: IOP Publishing.

[4] Xizhi Shi, Chaoyu He, Chris J. Pickard, Chao Tang, and Jianxin Zhong. Stochastic generation of complex crystal structures combining group and graph theory with application to carbon. *Physical Review B*, 97(1):014104, January 2018.