Supplementary Information 1 for: LightPFP: A Lightweight Route to Ab Initio Accuracy at Scale

Wenwen Li,[1, a)] Nontawat Charoenphakdee,[1, b)] Yong-Bin Zhuang,[1] Ryuhei Okuno,[1] Yuta Tsuboi,[1] So Takamoto,[1] Junichi Ishida,[2] and Ju Li[3, 4]

[1)] *Preferred Networks Inc., Tokyo, Japan.*

[2)] *Matlantis Corporation, Tokyo, Japan.*

[3)] *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

[4)] *Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*

(Dated: 26 November 2025)

[a)]Electronic mail: wenwenli@preferred.jp

[b)]Electronic mail: nontawat@preferred.jp

CONTENTS

Supplementary Note 1. Discussion about Error transfer in the DFT→PFP→LightPFP Pipeline 3

Supplementary Note 2. Details of data efficiency evaluation 4

Supplementary Note 3. Details of high-entropy alloy example 6

Supplementary Note 4. Details of SiO2 dry etching MD simulations 13

Supplementary Note 5. Pretrained student model 14

Supplementary Note 6. Training structure sampling methods 18

A. Molecular dynamics sampling 19

B. Uniform compression/stretch sampling 19

C. Deformation sampling 19

D. Displacement sampling 19

E. Rattle sampling 20

F. Vacancy sampling 20

G. Surface sampling 20

H. Substitution sampling 20

Supplementary Note 7. Hyperparameters of LightPFP models 22

Supplementary Note 8. 3-stages training method 24

Supplementary Note 9. Active learning method 26

References 29

Supplementary Note 1. DISCUSSION ABOUT ERROR TRANSFER IN THE DFT→PFP→LIGHTPFP PIPELINE

As mentioned in the main text, the dominant error arises from formally exact → DFT due to DFT's intrinsic limitations; by contrast, DFT → PFP transfer error is already small (with the costly training completed), and PFP → LightPFP is even smaller with fast, overnight training. Independent sources of error, e.g. e_1, e_2, ..., e_m, typically do not add up linearly but rater quadratically if statistically uncorrelated due to different "physics":

$$e = \sqrt{e_1^2 + e_2^2 + ... + e_m^2}, \tag{S1}$$

if $|e_1| \sim 100$ meV/atom dominates over $|e_2|, ..., |e_m|$, then the leading-order contributions of $|e_2|, ..., |e_m|$ to the total error would be even smaller than it seems, based on Taylor expansion:

$$e \approx e_1 + \frac{e_2^2 + ... + e_m^2}{2e_1}, \tag{S2}$$

and likely become practically negligible. In other words, if the DFT→PFP and PFP→LightPFP neural network trainings are done well, LightPFP may represent DFT much better than how DFT reflects reality.

Supplementary Note 2. DETAILS OF DATA EFFICIENCY EVALUATION

A full dataset of 1529 structures was constructed by exhaustively sampling the relevant configuration space using the PFP. Sampling combined both static and dynamic approaches. The static method involved sampling structures by applying lattice compression, deformation, and atomic displacements. The dynamic sampling was performed using molecular dynamics (MD) simulations, starting from initial configurations of defect-free and defective bulk structures, as well as surface structures.

To evaluate data efficiency, smaller datasets ranging from 100 to 850 structures were generated via two strategies: subsampling and direct sampling. Subsampling involves randomly picking a certain number of structures from the full dataset, while direct sampling is performed from scratch with a reduced number of MD steps. Each dataset size was created five times to estimate error uncertainties. The subsampled datasets exhibit a broader distribution across the configuration space, while the direct sampling approach more closely mimics common practical scenarios where users typically shorten MD trajectories.

The data efficiency of pre-trained students is demonstrated by the comparing the performance of LightPFP models trained from different size of datasets. The energy, force, phonon frequency and surface energy have been shown in the main text, and here more detailed results, including, stress, lattice length, elastic tensor, elastic modulus, and vacancy formation energy are presented in Fig S1.

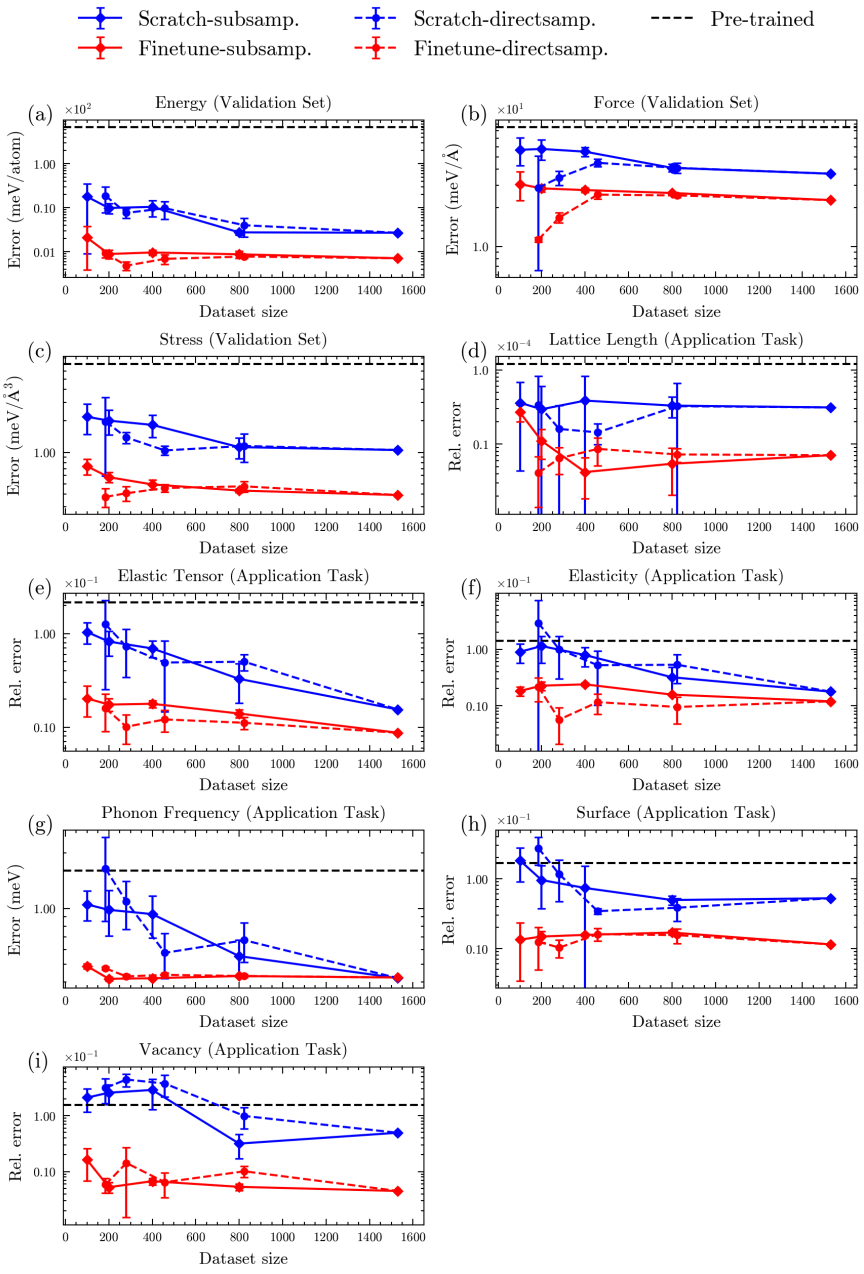

FIG. S1. Detailed comparison of data efficiency between fine-tuned pretrained and scratch-trained student models.

Supplementary Note 3. DETAILS OF HIGH-ENTROPY ALLOY EXAMPLE

This example focuses on high entropy alloys (HEAs), specifically the Cantor alloy with a face-centered cubic (FCC) lattice. The composition is 20% each of Al, Co, Cr, Fe, and Ni. HEAs have attracted significant attention due to their exceptional mechanical properties. However, their complex multi-element nature poses challenges for training MLIPs. In the following, we train MLIPs applicable not only to bulk HEA but also to interfaces and grain boundaries.

As in the previous example, we evaluate the same four MLIP usage strategies—PFP, MACE, LightPFP, and MTP-DFT—with the same meanings as defined in the main text. For LightPFP and MTP-DFT, we construct training datasets using an identical sampling workflow. Because equiatomic AlCoCrFeNi high-entropy alloys are substitutional solid solutions without a unique ordered configuration, each lattice site experiences a wide variety of local chemical environments. To efficiently sample this diversity, we adopt a random-substitution protocol: starting from an fcc Al host, each lattice site is independently assigned one of Al, Co, Cr, Fe, Ni with equal probability ($\approx$ 20 at.% per element), and the resulting structures are sampled using PFP-driven molecular dynamics. This procedure is repeated across multiple initial cells to diversify the dataset. The initial pool includes fcc bulk crystals, surface slabs with Miller indices less than 4, and coincidence-site-lattice (CSL) grain boundaries with low Σ ($<$10). For LightPFP, PFP-driven MD sampling takes 26 hours to generate 9,638 structures (1,356,616 atoms), followed by 1 hour of model training (27 hours total). For the DFT-based baseline (MTP-DFT), we use the same PFP-driven sampling strategy but label a smaller set—1,012 configurations (60,360 atoms), including surfaces and grain boundaries relevant to the intended application—by single-point DFT calculations. Some configurations (e.g., the (3 1 1) slab with at least 144 atoms) require relatively large cells, making DFT labeling expensive due to the nominal cubic scaling of Kohn–Sham DFT. The DFT calculations took 637 hours on a single GPU; by simple extrapolation, fully *ab initio* MD sampling would require on the order of 60,000 hours, i.e., more than three orders of magnitude slower than the LightPFP route. Table S2 summarizes the datasets used for the training of LightPFP and MTP-DFT models. These results again highlight the advantage of using a universal potential for rapid, low-cost data collection.

Runtime benchmarks on an NVIDIA V100 (16 GB) show that LightPFP and MTP-DFT

achieve an inference speed of 9.8×10^{-7} s/step/atom—66x faster than PFP (6.5×10^{-5} s/step/atom) and 249x faster than MACE (2.4×10^{-4} s/step/atom), see Figure S2. The maximum system size that fits on a single GPU is 716,800 atoms for LightPFP/MTP-DFT, compared to 13,824 for PFP (52x smaller) and 1,792 for MACE (400x smaller). When construction cost is considered, LightPFP offers the best overall trade-off: it pairs the fastest inference with a 27-hour build, which is orders of magnitude cheaper than the 60,000 hours required for MTP-DFT.

Using DFT forces as ground truth on a held-out test set, the force MAEs follow the same ordering observed previously: PFP (0.103 eV/Å) < MTP-DFT (0.123 eV/Å) < LightPFP (0.134 eV/Å) < MACE (0.184 eV/Å). This again shows that the distilled LightPFP incurs a modest accuracy penalty relative to its teacher and a DFT-trained baseline, yet retains substantially higher efficiency.

We assess the accuracy of the four MLIPs on key properties of AlCoCrFeNi, using DFT as the reference: the equation of state (EOS), elastic constants, surface formation energies, and grain-boundary (GB) formation energies. Unless otherwise noted, results are averaged over multiple random elemental arrangements to account for chemical disorder, and numerical comparisons are summarized in Table S1.

We began with the equation of state. Starting from a relaxed 256-atom bulk cell, we varied the lattice constant by ±5%, relaxed atomic positions at fixed volume, and fitted the resulting energy–volume data with a Birch–Murnaghan EOS to obtain the equilibrium volume and bulk modulus. As shown in Fig S3, PFP and LightPFP closely reproduce the DFT energy–volume curve. MACE also follows the DFT curve but exhibits small systematic deviations in the fitted parameters. By contrast, MTP-DFT underestimates the equilibrium volume by approximately 2.5%, which may reflect limited coverage of relevant local environments in its DFT-labeled training set.

Then, the elastic tensor, bulk, Young's, and shear moduli are computed with the stress–strain methodology[1] using the same bulk structure. PFP provides the closest agreement with DFT with average error of 7.2 GPa. LightPFP (10.65 GPa) tracks PFP closely. MTP-DFT (12.55 GPa) generally remains comparable to LightPFP for these mechanical properties, while MACE shows more pronounced deviations, 23.35 GPa. Overall, the spread among PFP, LightPFP, and MTP-DFT is modest for elasticity, whereas MACE underperforms on this task.

TABLE S1. Comparison of DFT and MLIPs on properties of AlCoCrFeNi high-entropy alloy

Property	**DFT**	**PFP**	**LightPFP**	**MACE**	**MTP-DFT**
Equation of State					
Volume (Å^3/atom)	11.58	**11.51**	*11.51*	11.48	11.29
Bulk modulus (GPa)	165.64	**165.66**	*164.35*	159.18	162.27
Mechanical Properties (GPa)					
C11	195.2	202.5	196.3	177.2	197.2
C22	211.4	206.9	203.3	183.5	202.7
C33	197.5	206.7	204.3	182.7	203.1
C12	140.9	145.9	151.7	145.9	153.3
C13	142.9	152.9	156.6	148.3	157.6
C23	131.1	137.9	144.9	141.3	148.2
C44	116.5	109.4	106.2	80.2	103.7
C55	124.0	114.2	110.6	84.6	107.1
C66	120.3	112.9	109.9	83.9	106.8
Bulk modulus	159.23	165.45	167.81	157.14	169.02
Shear modulus	69.99	65.79	60.42	45.05	58.54
Young's modulus	183.14	174.27	161.84	123.36	157.44
Average Error	–	**7.20**	*10.65*	23.35	12.55
Surface Energy (eV/Å^2)					
(4, 1, 0)	0.127	0.136	0.133	0.121	0.126
(4, 1, 1)	0.170	0.171	0.165	0.167	0.168
(4, 2, 1)	0.142	0.149	0.148	0.134	0.145
(4, 3, 0)	0.139	0.144	0.143	0.137	0.143
(4, 3, 2)	0.137	0.143	0.145	0.126	0.142
(4, 4, 1)	0.148	0.153	0.153	0.146	0.154
(4, 4, 3)	0.171	0.178	0.175	0.174	0.176
Average Error	–	0.0058	0.0053	*0.0052*	**0.0036**
Grain Boundary Energy (eV/Å^2)					
Σ13 22.62/[1 0 0]	0.0559	0.0621	0.0578	0.0424	0.0523
Σ15 48.19/[1 2 0]	0.0794	0.0809	0.0787	0.0602	0.0825
Σ13 147.80/[1 1 1]	0.0378	0.0300	0.0294	0.0206	0.0268
Σ13 67.38/[1 0 0]	0.0584	0.0617	0.0563	0.0332	0.0504

Since the low-index surfaces were included in training dataset, we evaluated higher-index surfaces with Miller index > 3 to probe the performance of MLIPs in surface formation energy calculation. The surface formation energy was computed as:

$$\gamma_{\text{surf}} = \frac{E_{\text{surf}} - \frac{n_{\text{surf}}}{n_{\text{bulk}}} E_{\text{bulk}}}{2A_{\text{surf}}} \tag{S3}$$

where E_{surf} is the energy of a slab with two surfaces, E_{bulk} is the energy of the bulk HEA, n_{surf} and n_{bulk} are the atom counts in the surface and bulk structures, and A_{surf} is the surface area. All four MLIPs achieve high accuracy, with average absolute errors below 0.006 eV/Å^2 relative to DFT. On this task the inter-model differences of average error are very small among PFP, LightPFP and MACE (0.0052-0.0058 eV/Å^2); while MTP-DFT (0.0036 eV/Å^2) is marginally closer to DFT.

Several CSL grain boundaries with $\Sigma > 10$ are selected for testing the MLIPs in GB formation energy. The GB formation energy was computed as:

$$\gamma_{\text{GB}} = \frac{E_{\text{GB}} - \frac{n_{\text{GB}}}{n_{\text{bulk}}} E_{\text{bulk}}}{2A_{\text{GB}}} \tag{S4}$$

where E_{GB} and E_{bulk} are the energy of GB and bulk structures, n_{GB} and n_{bulk} are their atoms counts, and A_{GB} is the grain boundary area. LightPFP, PFP and MTP-DFT reproduce the GB formation energy with modest accuracy with an average error < 0.01 eV/Å^2, whereas MACE shows larger deviations.

Across EOS, elasticity, surface energies, and GB energies, the overall spread among PFP, LightPFP, and MTP-DFT is small, and no single model dominates all properties. Importantly, despite its slightly larger force MAE relative to PFP and MTP-DFT, LightPFP does not exhibit a clear disadvantage in property-level predictions for this materials. This mirrors the earlier example, Li_6PS_5Cl: modest differences in force MAE do not necessarily translate into large discrepancies in computing materials properties, which can be comparably influenced by factors such as finite-size effects, and simulation settings. Together with its substantially lower construction cost and faster inference, these results support model distillation from a strong universal potential as a practical and accurate route for property calculations in complex, chemically disordered materials.

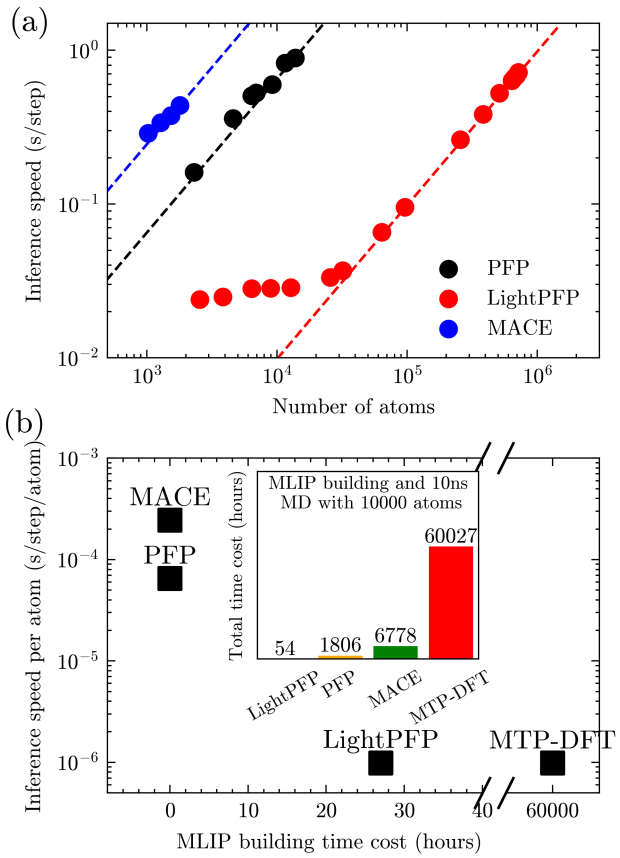

FIG. S2. (a) Molecular dynamics (MD) computational speed with AlCoCrFeNi high-entropy alloy as a function of number of atoms for three MLIPs: PFP, LightPFP (MTP), and MACE. (b) Trade-off between the overall time spent on MLIP building for AlCoCrFeNi high-entropy alloy, including data collection and model training, and MD computational speed for PFP, LightPFP, MACE, and MTP. Inset: the total time cost to complete both MLIP building and a 10 ns MD simulation of a 10,000-atom system With PFP, LightPFP, MACE, and MTP.

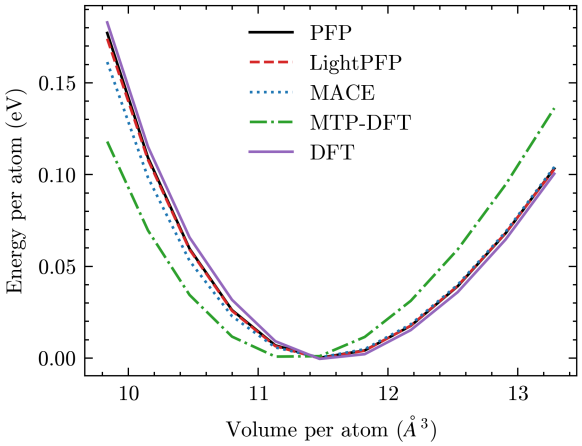

FIG. S3. Equation of states of AlCoCrFeNi high-entropy alloy calculated by DFT, PFP, LightPFP, MACE and MTP

TABLE S2. Composition of the AlCoCrFeNi high-entropy alloy dataset.

Type of structure	**Sampling method**	**Number of structures**	**Number of atoms**
LightPFP Dataset (labeled by PFP)			
crystal	substitution+MD	2040	206040
boundary	substitution+MD	6200	1083760
slab	substitution+MD	1398	66816
Total		**9638**	**1356616**
MTP Dataset (labeled by DFT)			
crystal	substitution+MD	531	42484
boundary	substitution+MD	286	9152
slab	substitution+MD	195	8724
Total		**1012**	**60360**

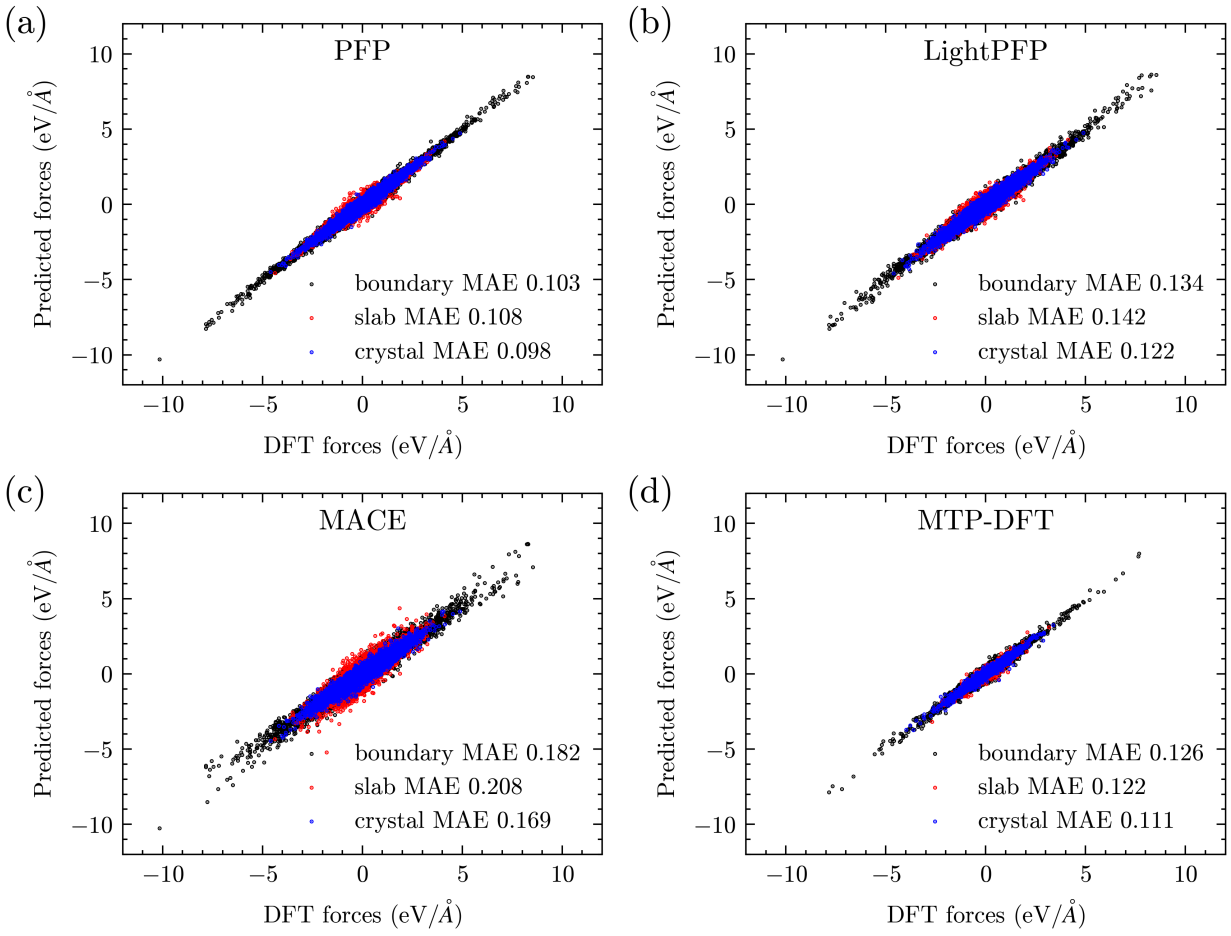

FIG. S4. Parity plot of DFT forces againes predicted forces by different MLIPs, (a) PFP; (b) LightPFP; (c) MACE and (d) MTP

Supplementary Note 4. DETAILS OF SIO2 DRY ETCHING MD SIMULATIONS

In the example of SiO_2 dry etching, the reactive MD simulation is performed in active learning, etching yeilding rate validation and large-scale simulation. The details of MD simulation settings are described as below. In each insertion cycle a single HF molecule was introduced above the SiO_2 surface with its velocity vector oriented normal to the surface. Its kinetic energy is specified as following: randomly sampled in 20-80 eV in active learning; 20, 40 and 60 eV in etching yielding rate evaluation; 40 eV in large-scale simulation. The positions of HF molecules in the plane are random, except in large-scale simulations where we restrict them to a 2x2nm square region to simulate the actual etching process of semiconductor devices. To capture the short-time, high-energy collision events the trajectories were first propagated in the microcanonical (NVE) ensemble for 200 fs using a 0.2 fs time step; this fine time resolution ensures accurate integration of the forces during energetic impacts. Following the collision stage, the system was relaxed in the canonical (NVT) ensemble for 1,000 fs with a 1 fs time step to thermalize the surface back to 300 K.

Task index	Dataset description
1	Single substance structures
2	Bulk structures with two elements
3	Crystal structures from materials project database
4	Structures with random atomic position
5	Bulk structures with defects
6	Single molecules
7	Structures composed by multiple molecules
8	Molecules with element substitution
9	Adsorption structures of surfaces and molecules
10	Random combinations of surfaces and molecules
11	Slabs
12	Clusters

TABLE S3. Task definitions for Reptile meta learning algorithm based on datasets trained for PFP[2].

Supplementary Note 5. PRETRAINED STUDENT MODEL

a. All-elements pretrained student model We employed a large dataset which is used to train PFP. The dataset comprises 86 different elements, covering nearly the entire periodic table and encompassing both equilibrium structures and numerous disordered structures that deviate from equilibrium states. The dataset includes not only bulk phases but also complex structures such as surfaces, adsorption configurations, and clusters. This comprehensive coverage is the fundamental reason why PFP exhibits broad applicability across diverse materials simulations. For dataset details, please refer to Takamoto *et al.*[2].

However, compared to PFP[2], moment tensor potentials (MTPs) are compact models with limited parameters and constrained expressive power, typically applicable only to single materials systems. Consequently, using MTPs to fit all datasets simultaneously presents significant challenges. Therefore, our MTP pretraining strategy aims to optimize the model to facilitate subsequent fine-tuning for individual tasks, instead of maximizing accuracy across all datasets. To achieve this objective, we employed the Reptile meta-learning algorithm[3].

The Reptile algorithm operates by iteratively sampling tasks from a task distribution and

updating model parameters to enhance the model's ability to rapidly adapt to new tasks. In our implementation, we partitioned the complete dataset into 12 specific tasks based on structural types, as detailed in Table S3. During each inner loop iteration, we select a task (i.e., a dataset containing specific structural types such as single molecules) to train the MTP model. Given the substantial size of each task's dataset, we limit training to one epoch per inner loop before proceeding to the parameter update. The model parameters are then updated according to the following formula:

$$\delta\theta = \theta_i - \theta,$$

$$\theta \leftarrow \theta + \beta\delta\theta,$$

where θ represents the MTP parameters, θ_i denotes the parameters after the i-th inner loop, and β is a hyperparameter in the Reptile algorithm that controls the magnitude of the meta-update step during training. In our implementation, β is set to 0.5. We iteratively repeat the task sampling and inner-loop/meta-update procedures for 100 iterations until convergence of energy, forces, and stress is observed across all datasets.

We employed the Adam optimization method with a learning rate of 1×10^{-3}. The model was trained for 1 epochs with a batch size of 256. Total pretraining time was approximately 100 hours.

For example, pretrained student model with hyperparameter (levmax=8, C_μ=1, C_ν=1, n_q=16) contains 86×86×4×16 training parameters for the radial function c and additional 27 coefficients for the basis functions ξ. The modular structure of MTP enables selective parameter extraction during inference or fine-tuning, significantly enhancing computational efficiency. The extraction procedure is straightforward, depending on elements used for the task. For example, when handling a material containing only H and O elements, we can extract the relevant subset of the radial function parameter tensor—specifically a 2×2×4×16 matrix corresponding to these elements, while maintaining the coefficients of the basis function unchanged. Consequently, although the pretrained model may contain numerous parameters, it automatically reduces to a compact, element-specific model equivalent in size to those trained from scratch for the particular material system.

b. Specific type pretrained student model In addition to the pretrained LightPFP model that covers almost all materials, we also tried pretrained LightPFP models for special types of materials. As an illustrative example, we consider our organic pretrained student model,

which is specifically designed for organic molecular systems. The training process begins with dataset construction. We randomly sample molecular information, such as SMILES representations, from PubChem[4] and generate corresponding three-dimensional conformers. Several molecules are then randomly placed into a simulation box, ensuring that the overall density falls within an appropriate range. An optimization algorithm is employed to adjust atomic positions without breaking chemical bonds, thereby minimizing atomic overlap between molecules. From these initial configurations, we perform molecular dynamics (MD) simulations using PFP at temperatures randomly selected between 300 and 3000 K. Each simulation runs for 1000 steps, and configurations are sampled every 100 steps. This procedure is repeated many times to obtain a diverse collection of molecular configurations. The resulting dataset is subsequently used for training a Moment Tensor Potential (MTP) model, yielding a pretrained MTP tailored for organic systems.

We observe that when the model is restricted to a specific class of materials, the pretrained MTP demonstrates a notable capability for direct application without fine-tuning. As shown in Fig. S5, the pretrained model accurately predicts the densities of various organic molecules, exhibiting strong agreement with experimental results despite the absence of these molecules from the training dataset. This finding suggests a promising new direction: by constraining the material domain, one can develop lightweight machine-learned interatomic potentials (MLIPs) with reduced generalization compared to universal MLIPs (uMLIPs), yet capable of fast inference and requiring no additional training. For instance, pretrained student models can be constructed for specific material classes such as alloys, oxides, perovskites, and metal–organic frameworks (MOFs).

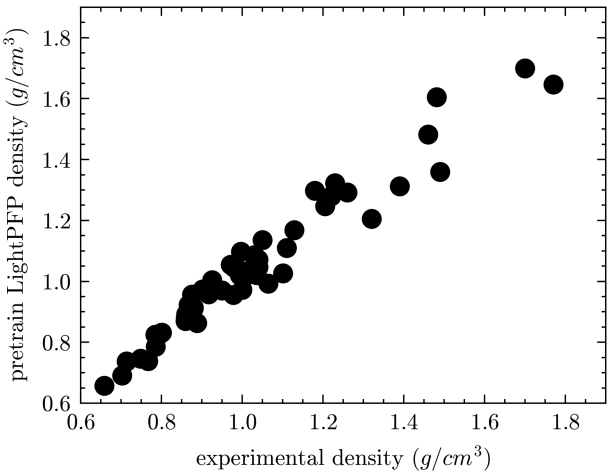

FIG. S5. Comparison between predicted densities from the organic pretrained LightPFP model and experimental densities

Supplementary Note 6. TRAINING STRUCTURE SAMPLING METHODS

A robust training dataset for machine learning interatomic potentials is built by systematically sampling diverse yet physically meaningful configurations around one or more initial structures. Sampling methods can be combined and run independently to cover thermal, mechanical, defect, surface, and chemical degrees of freedom. These strategies balance relevance to the target material with diversity across configuration space, improving both accuracy and robustness of the potential. The illustration of these sampling methods are shown at Figure S6.

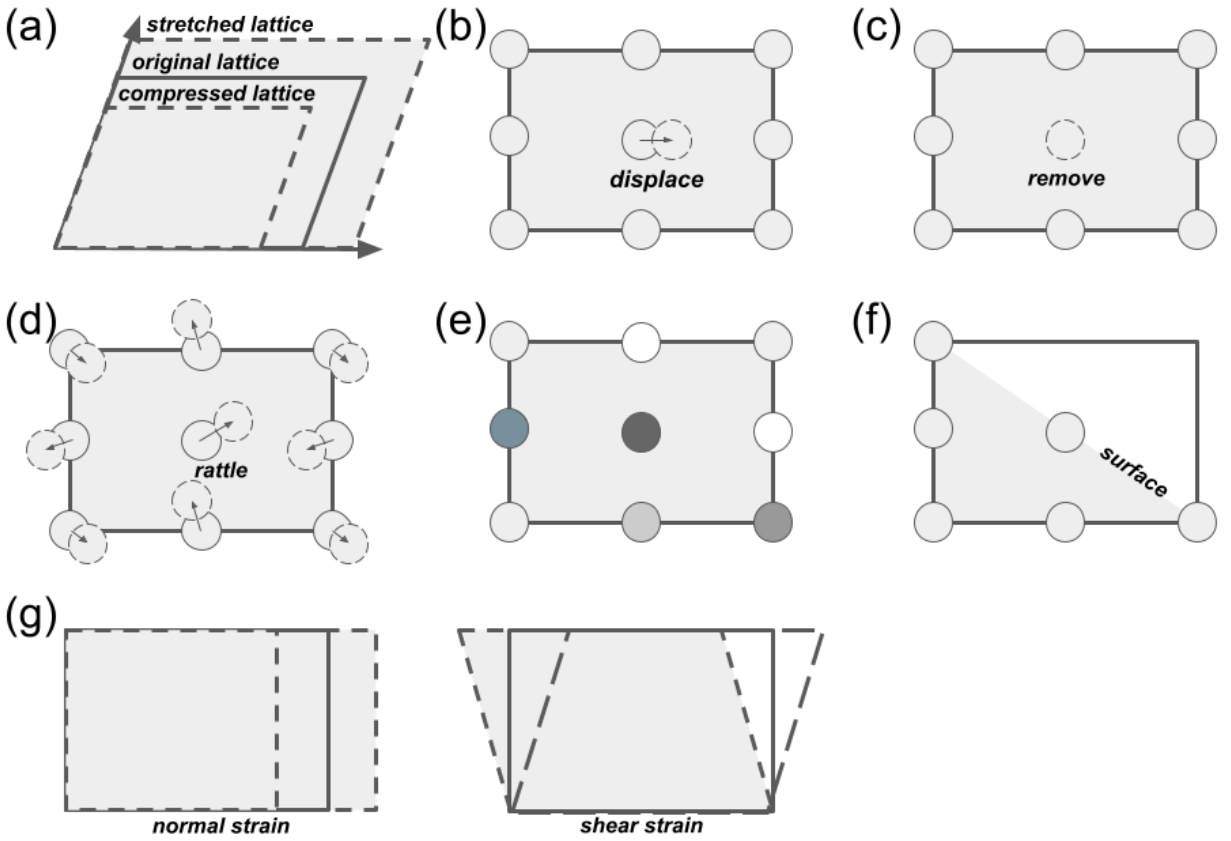

FIG. S6. Illustration of sampling methods used in the LightPFP dataset generation. (a) Uniform compression/stretch sampling, (b) Displacement sampling, (c) Vacancy sampling, (d) Rattle sampling, (e) Substitution sampling, (f) Surface sampling and (g) Deformation sampling. Reproduced from the Matlantis manual (accessed 12 Nov 2025).

A. Molecular dynamics sampling

Molecular dynamics (MD) sampling generates structures by propagating the atomic system under finite-temperature dynamics, optionally at controlled pressure. By choosing ensembles such as NVT (constant volume) or NPT (constant pressure), and by varying temperature, one can explore configurational space from near-equilibrium states to highly disordered regimes. To prevent collecting redundant configurations, snapshots are taken at a fixed stride along the trajectory.

B. Uniform compression/stretch sampling

Uniform compression/stretch sampling produces structures by isotropically scaling the lattice vectors of the input periodic structure, keeping the lattice angles unchanged and preserving fractional atomic coordinates. This method targets the volume–energy relationship and can be augmented by fixed-cell relaxation or MD runs starting from the scaled configurations to enrich the dataset at specific densities.

C. Deformation sampling

Deformation (strain) sampling applies prescribed normal and shear strain components to the unit cell, changing both lattice lengths and angles while maintaining periodicity. By scanning the six independent components of the strain tensor, one obtains structures spanning elastic distortions relevant to mechanical properties. Atomic positions may be further optimized under fixed cell shape to produce relaxed strained configurations, helping the model learn stress–strain behavior and elastic responses.

D. Displacement sampling

Single-atom displacement sampling perturbs one atom from its equilibrium position along a Cartesian direction by a controlled amplitude. Such localized perturbations probe the curvature of the potential energy surface and the force constants around equilibrium, which are essential for learning vibrational responses. Each displaced configuration is generated independently from the same starting structure to map local force landscapes efficiently.

E. Rattle sampling

Rattle sampling introduces random displacements to all atoms simultaneously, drawing each component of the displacement from a specified distribution (e.g., Gaussian). This global perturbation broadens coverage of non-equilibrium configurations and can reveal failure modes of the model under larger distortions. Because it may produce unphysical configurations with extreme forces, filtering based on maximum force thresholds and optional relaxation steps are recommended to maintain data quality. This method is recommended for the molecular systems since it provided useful information of bond breaking.

F. Vacancy sampling

Vacancy sampling creates point-defect structures by randomly removing one or more atoms from the initial configuration. These defective structures can be complemented with fixed-cell relaxations or MD to sample local reconstructions and thermally activated defect configurations. By including vacancy-containing data, the model gains sensitivity to defect energetics and local structural changes associated with missing atoms.

G. Surface sampling

Surface sampling constructs slab models by cleaving the periodic bulk along specified Miller indices and introducing a vacuum layer to isolate the surfaces. Symmetry analysis can be used to avoid duplicate surfaces generated by equivalent indices in high-symmetry crystals. Subsequent fixed-cell relaxations and MD on slab geometries enrich the dataset with surface reconstructions and thermal fluctuations. This approach is intended for periodic crystalline inputs and targets accurate description of surface energetics and structure.

H. Substitution sampling

Element substitution sampling generates chemically disordered structures by stochastically replacing atoms in the initial structure with user-specified species at defined probabilities. This method captures configurational variability in multicomponent systems, such as alloys, by sampling diverse local chemistries. Fixed-cell relaxation and MD can be applied

after substitution to explore thermally accessible configurations, improving robustness and transferability across compositional variations. This method is useful for the solid-solution and high-entropy alloy.

Supplementary Note 7. HYPERPARAMETERS OF LIGHTPFP MODELS

The hyperparameters of the LightPFP models used in the results section is listed here

Material	Cutoff	levmax	C_μ	C_ν	n_q	Neural Network Readout
Ni3Al	6.0	8	1	1	16	None
Li6PS5Cl	6.0	8	1	1	16	[16, 16, 1]
HEA	5.0	8	1	1	16	None
MgO	6.0	8	1	1	16	[16, 16, 1]
SiO2-HF	6.0	8	1	1	16	[16, 16, 1]

To specify the complexity of moment tensor potential, parameter cutoff, levmax, μ, ν, n_q is used. Once such hyperparameters are defined, each admissible basis function must have the level less than levmax. As explained in the main text, the basis function B_α comprises of matrix contractions of moment descriptors $M_{\mu,\nu}$

$$M_{\mu,\nu}(\mathbf{n}_i) = \sum_j f_\mu(\mathbf{r}_{ij}) \underbrace{\mathbf{r}_{ij} \otimes \mathbf{r}_{ij} \otimes \cdots \otimes \mathbf{r}_{ij}}_{\nu \text{ times}}$$

where the level of basis can be calculated as

$$\text{lev} = 2 + C_\mu \times \mu + C_\nu \times \nu \tag{S5}$$

The n_q is the number of radial basis functions in the polynomial function. 27 basis functions are admissible when we set levmax=8, C_μ=1, C_ν=1:

where the level is defined following Shapeev (MTP) convention, "$\cdot$" is a dot product between vectors and " : " is a Frobenius product of two matrices.

Intuitively, one can think that the higher the levmax, the more complex the MTP. The lower C_μ and C_ν, the more complex the MTP. Note that the more complex the MTP, the more memory and the longer the computation time it requires.

Basis index	Moment tensor component	Level
B_0	$M_{0,0}$	2
B_1	$M_{0,0} \times M_{0,0}$	4
B_2	$M_{0,0} \times M_{0,0} \times M_{0,0}$	6
B_3	$M_{0,0} \times M_{0,0} \times M_{0,0} \times M_{0,0}$	8
B_4	$M_{1,0}$	3
B_5	$M_{0,0} \times M_{1,0}$	5
B_6	$M_{0,0} \times M_{0,0} \times M_{1,0}$	7
B_7	$M_{1,0} \times M_{1,0}$	6
B_8	$M_{0,0} \times M_{1,0} \times M_{1,0}$	8
B_9	$M_{2,0}$	4
B_{10}	$M_{0,0} \times M_{2,0}$	6
B_{11}	$M_{0,0} \times M_{0,0} \times M_{2,0}$	8
B_{12}	$M_{1,0} \times M_{2,0}$	7
B_{13}	$M_{2,0} \times M_{2,0}$	8
B_{14}	$M_{3,0}$	5
B_{15}	$M_{0,0} \times M_{3,0}$	7
B_{16}	$M_{1,0} \times M_{3,0}$	8
B_{17}	$M_{4,0}$	6
B_{18}	$M_{0,0} \times M_{4,0}$	8
B_{19}	$M_{5,0}$	7
B_{20}	$M_{6,0}$	8
B_{21}	$M_{0,1} \cdot M_{0,1}$	6
B_{22}	$M_{0,0} \times (M_{0,1} \cdot M_{0,1})$	8
B_{23}	$M_{0,1} \cdot M_{1,1}$	7
B_{24}	$M_{0,1} \cdot M_{2,1}$	8
B_{25}	$M_{1,1} \cdot M_{1,1}$	8
B_{26}	$M_{0,2} \colon M_{0,2}$	8

TABLE S4. Definitions of B_i with corresponding right-hand side expressions and levels.

Supplementary Note 8. 3-STAGES TRAINING METHOD

We propose a three-stage training strategy for the LightPFP model. In Stage I, the optimization focuses on fitting forces; in Stage II, the emphasis shifts to energy and stress; and in Stage III, the loss terms are balanced so that the energy, force, and stress losses are of comparable magnitudes. This is achieved by progressively adjusting the coefficients of the energy (α), force (β) and stress (γ) terms in the loss function.

To assess the effectiveness of this strategy, we trained on the Li_6PS_5Cl dataset and conducted three fixed-weight baselines. Each baseline uses constant loss weights equal to those employed in one stage of the three-stage schedule: Baseline 1 ($\alpha = 10^{-5}, \beta = 10.0, \gamma = 10^{-5}$), Baseline 2 ($\alpha = 1.0, \beta = 0.1, \gamma = 10.0$), and Baseline 3 ($\alpha = 26.2, \beta = 0.034, \gamma = 1383.1$). The evolution of the losses over epochs is shown in Figure S7, and the final losses are summarized in Table S5.

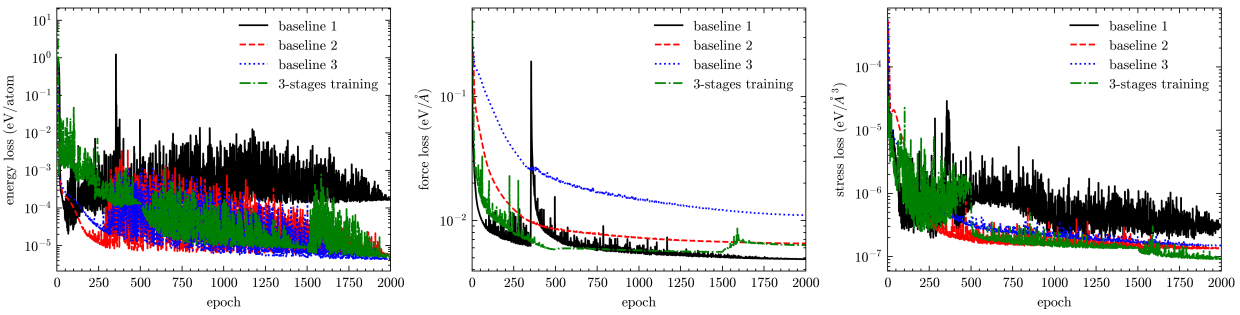

FIG. S7. Convergence curves comparing three fixed-weight baselines and the proposed three-stage training strategy

	baseline 1	baseline 2	baseline 3	3-stages-training
energy	1.67×10^{-4}	5.76×10^{-6}	4.39×10^{-6}	4.96×10^{-6}
forces	0.00486	0.00651	0.0110	0.00631
stress	3.03×10^{-7}	1.35×10^{-7}	1.50×10^{-7}	9.28×10^{-8}

TABLE S5. Comparison of the final energy, force, and stress losses across training strategies.

Baseline 1 fails to fit the energy accurately, yielding the largest energy loss, while Baseline 3 fails to fit forces accurately. Baseline 2 offers a more balanced trade-off; however, its energy, force, and stress losses are all larger than those achieved by the three-stage training. Overall,

the proposed three-stage schedule provides the best balance across the three targets, with near-optimal energy and stress losses and competitive force accuracy.

Supplementary Note 9. ACTIVE LEARNING METHOD

Active learning is a powerful approach for developing accurate and efficient interatomic potentials in molecular dynamics simulations. Here is a brief introduction of the active learning workflow we used for the "Dry etching of SiO_2" example and several other examples in the Supplementary Materials 2 (see Fig. S8):

1. Initial Dataset: A simple initial dataset is necessary for active learning. The initial dataset does not need to be large and robust.

2. Model Training: The initial LightPFP model is trained with this initial dataset.

3. Exploration: The LightPFP model is then used to drive molecular dynamics (MD) simulations, exploring new configurations and areas of the potential energy surface.

4. Quality Check: At certain MD steps, check the accuracy of LightPFP. Calculate the energy, forces, and stress of the MD snapshot using PFP, and compare these with the LightPFP predictions.

5. Data Selection: Include the MD snapshot in the training dataset if the discrepancy between PFP and LightPFP is greater than the minimum threshold and less than the maximum threshold.

6. Model Update: After several MD simulations are finished or cease based on other criteria, update the LightPFP model with the dataset collected in current and previous iterations and the initial dataset.

7. Iteration: Steps 3 to 6 are repeated iteratively. Each iteration improves the potential's accuracy and extends its applicability.

This active learning approach allows for the efficient development of accurate potentials by focusing computational resources on the most informative data points, ultimately resulting in a potential that can reliably reproduce the behavior of the target system across a wide range of conditions.

a. Sampling threshold We use minimum/maximum thresholds to collect high-quality training data. For more efficient training data collection, it also performs early stopping of MD simulations or PFP-based sampling upon detecting outliers. Both PFP and LightPFP are used to calculate the energy, forces, and stress of given MD snapshots. The errors of LightPFP w.r.t PFP are used to determine if certain MD snapshots should be collected. The valid error range is defined by both a minimum (lower bound) and a maximum threshold

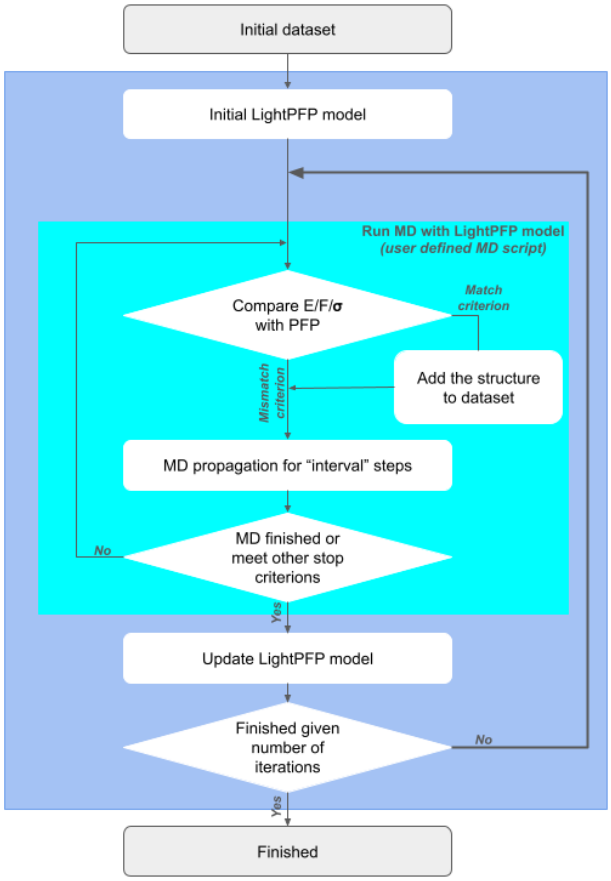

FIG. S8. Illustration of active learning workflow. Reproduced from the Matlantis manual (accessed 12 Nov 2025).

(upper bound). When the error of an MD snapshot is very large, exceeding the upper bound, it indicates an unreasonable structure that is not beneficial for training. Continuing MD from this point can lead to more structures that are not valuable. In such cases, the MD simulation will stop early. In the "Dry etching of SiO_2" example, error checking is performed every 100 steps, and the selection criterion is: energy error between 5.0 and 40.0 times of energy MAE of current using LightPFP model; force error (largest atomic error in the structure) is in between 1.5 and 50 eV/A.

b. MD early stop As mentioned, the MD simulation will stop early when the discrepancy between PFP and LightPFP is very large. This indicates that the MD has reached a configuration where the current LightPFP model is unreliable. While structures with huge errors compared to PFP are not useful for training, the structures leading up to such structures, typically several MD steps before, are critical. Learning from these preceding

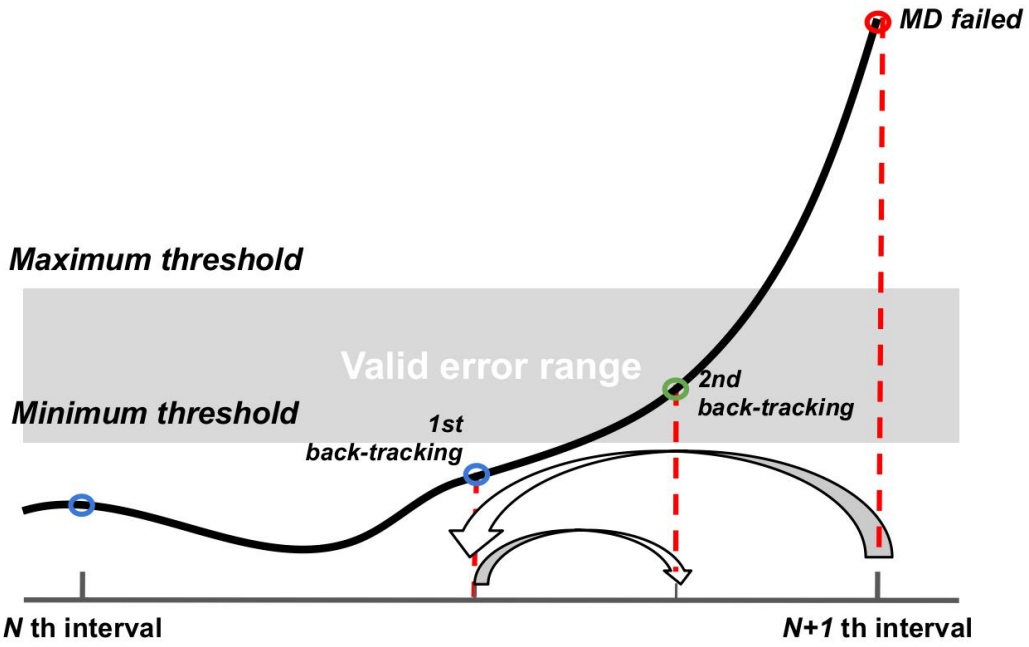

FIG. S9. Back-tracking mechanism of active learning when MD failed. Reproduced from the Matlantis manual (accessed 12 Nov 2025).

structures helps prevent the MD from evolving into unphysical configurations. Our workflow provides two mechanisms for this: (1) Back-tracking Method: When MD stops due to large prediction errors, the algorithm checks previous MD steps using a binary search until a training structure with the error in the specified region is found. MD snapshots are cached to facilitate this process. (2) PFP-based fallback: When MD stops early, the simulation rolls back to the previous checkpoint and continues using PFP instead of the LightPFP model for several more additional samples. In the "Dry etching of SiO_2" example, we collect 5 additional training structures based on PFP when MD failed.

c. Model update The model is updated in each iteration with the latest dataset and all previous datasets. To accelerate active learning, the total number of epochs for model training is adjusted according to the size of whole datasets, and the training time is kept roughly constant. This mechanism is designed to handle the gradually increasing dataset during the active learning iterations. In the "Dry etching of SiO_2" example, we fixed the time cost for each model update to 0.5 hour.

REFERENCES

[1]M. De Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. Van Der Zwaag, J. J. Plata, *et al.*, Scientific data **2**, 1 (2015).

[2]S. Takamoto, C. Shinagawa, D. Motoki, K. Nakago, W. Li, I. Kurata, T. Watanabe, Y. Yayama, H. Iriguchi, Y. Asano, *et al.*, Nature Communications **13**, 2991 (2022).

[3]A. Nichol, J. Achiam, and J. Schulman, arXiv preprint arXiv:1803.02999 (2018).

[4]S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, *et al.*, Nucleic acids research **53**, D1516 (2025).