Keep the Teacher Non-Fine-Tuned: Knowledge Distillation for Stable Molecular Dynamics with Neural Network Potentials

Naoki Matsumura, Yuta Yoshimoto, Yuto Iwasaki, Meguru Yamazaki, Yasufumi Sakai
Fujitsu Limited, Fujitsu Research, Kawasaki, Japan
matsumura-naoki@fujitsu.com

***Abstract*—Neural network potentials (NNPs) accelerate molecular dynamics (MD), but stable NNP-MD requires training data with broad configurational coverage, including rare high-energy states. Conventional knowledge distillation pipelines fine-tune a foundation teacher on density functional theory (DFT) and then use the fine-tuned teacher to generate training data via NNP-MD simulations for training a lightweight student. We show this order can harm MD: fine-tuning steepens the teacher potential energy surface, traps trajectories in low-energy basins, and collapses high-energy coverage. We therefore reverse the workflow. We keep the teacher non-fine-tuned as a generalist model to sample diverse NNP-MD configurations labeled with teacher energies and forces, distill a lightweight student on these soft targets, and then fine-tune only the student on a small DFT-labeled subset. On polyethylene glycol (PEG) and $Li_{10}GeP_2S_{12}$ (LGPS) materials using MatterSim as the generalist teacher, the student achieves near-experimental property accuracy while reducing expensive DFT labeling by 10× (PEG) and 5× (LGPS) relative to representative active learning baselines. It also yields large inference gains (e.g., 82× speedup for a 3,100-atom PEG system) and avoids teacher memory bottlenecks. In PEG, We further show that a fine-tuned-teacher pipeline can match static validation error yet fail in MD due to missing high-energy coverage, showing that configurational coverage is as important as static accuracy for NNP-MD.**

***Index Terms*—Knowledge distillation, Molecular dynamics, Neural network potential.**

I. Introduction

Molecular dynamics (MD) simulations are a core tool for understanding atomistic processes and predicting material properties. Their predictive power depends on accurate interatomic forces, which are traditionally obtained from quantum-mechanical calculations such as density functional theory (DFT). However, DFT becomes computationally prohibitive as system size grows, with costs that scale steeply (commonly cited as $O(N^3)$ with the number of electrons) [1], limiting simulations to small sizes and short timescales.

Neural network potentials (NNPs) address this bottleneck by learning the mapping from atomic coordinates to potential energy, with forces obtained as energy gradients. Once trained, NNPs can deliver near-DFT accuracy at dramatically lower cost, enabling large-scale NNP-driven MD (NNP-MD) [2], [3]. Recently, large foundation NNPs trained on massive, diverse datasets, such as M3GNet [4], CHGNet [5] and MatterSim [6], have further improved robustness across materials and thermodynamic conditions. However, these foundation models are often too slow and memory-intensive for production-scale MD, motivating model compression into lightweight NNPs.

Knowledge distillation (KD) [7] provides a natural route to compress a large teacher model into a fast student model by training the student to match the teacher's energy and force predictions. Recent KD pipelines for NNPs, such as DPA-2 [8] and PFD [9], typically follow a "fine-tune-then-distill" order: (i) fine-tune a pre-trained foundation model on a small, task-specific DFT dataset to obtain an accurate specialist teacher, and then (ii) use that fine-tuned teacher to generate NNP-MD trajectories that serve as training data for the student. This workflow is effective for model compression and static accuracy, but it can lead to a critical failure case when the end goal is robust MD, because the teacher used for NNP-MD determines the sampling distribution through its potential energy surface.

The root cause is that MD data generation is governed by the teacher's potential energy surface (PES). Fine-tuning improves agreement on individual configurations with DFT and corrects known systematic biases where pre-trained foundation models underestimate energies in high-energy regions [10]. However, it can also reshape the PES by increasing effective barriers between configurational states. As a result, teacher-driven MD becomes trapped in low-energy basins, producing a distillation dataset that is strongly biased toward low-energy configurations and lacks coverage of higher-energy structures. This is problematic because high-energy configurations are critical for training NNPs that enable stable and accurate MD simulations [11], [12]. Figure 1 summarizes this mechanism schematically. The fine-tuned teacher (orange) can be closer to the DFT reference PES (black dashed), yet its steeper landscape confines MD trajectories to low-energy valleys, suppressing transitions that would otherwise explore higher-energy configurations. Consequently, students trained on such trajectories can achieve low static validation errors yet fail in MD simulations, as we demonstrate in our PEG case study. This occurs because the training distribution converges to low-energy regions, overlooking crucial configurations essential for dynamic robustness.

To address this issue, we reverse the conventional order. We keep the teacher non-fine-tuned as a generalist foundation model and use it purely for configurational sampling via NNP-MD. Because the generalist teacher tends to exhibit a smoother

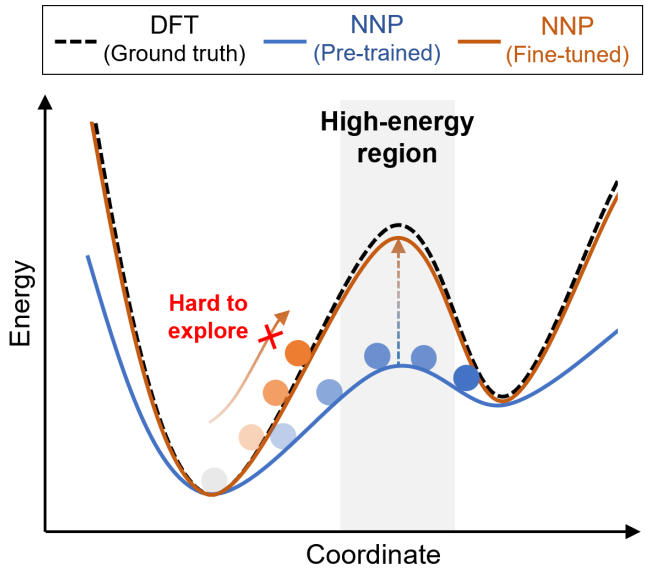

Fig. 1. Schematic potential energy surfaces (PES): DFT reference (black dashed), a pre-trained generalist teacher (blue), and a fine-tuned specialist teacher (orange). Fine-tuning can improve agreement on individual configurations with DFT, especially at higher energies, but it may also increase effective barriers and steepen the PES. As a result, teacher-driven MD becomes confined to low-energy basins, and the generated distillation dataset collapses toward low-energy configurations with reduced high-energy coverage. This motivates using a non-fine-tuned generalist teacher for configurational sampling and deferring DFT alignment to student-only fine-tuning.

PES, it facilitates transitions across configurational space and yields a more diverse set of sampled structures, including higher-energy configurations. We then distill a lightweight student from the teacher-labeled energies and forces and recover quantitative ground-truth accuracy by fine-tuning only the student on a small, strategically selected DFT-labeled subset. This decouples "sampling for coverage" from "alignment to DFT accuracy" in a simple and deployable recipe.

We validate the proposed workflow on two distinct material systems: an organic polymer (polyethylene glycol, PEG) and an inorganic solid-state electrolyte ($Li_{10}GeP_2S_{12}$, LGPS), using MatterSim [6] as the generalist teacher and DeepPot-SE (DP) [13] as a student. Across both systems, we achieve strong data efficiency—reducing expensive DFT labeling by up to 10×—while maintaining high accuracy for target properties. The distilled student also provides substantial inference speedups and overcomes the teacher model's memory limitations, enabling larger-scale simulations that are impractical with the foundation teacher.

Our main contributions are:

- We demonstrate a failure case in PEG where a fine-tune-then-distill pipeline achieves low static validation error but breaks down during MD. We explain this mismatch by a shift in the training distribution: fine-tuning steepens the teacher PES, confines teacher-driven MD to low-energy basins, and therefore under-samples the higher-energy configurations needed for stable trajectories.
- We propose a generalist-teacher distillation workflow that keeps the teacher non-fine-tuned for broad configurational sampling, distills a lightweight student from teacher labels, and fine-tunes only the student with a small, diversity-selected DFT dataset.
- We demonstrate strong data efficiency and deployment benefits on PEG and LGPS, including 5–10× fewer DFT labels and large speedups and memory savings that enable larger-scale NNP-MD.

The remainder of this paper is organized as follows. Section II reviews related work on NNPs, data generation, and KD. Section III describes the proposed workflow. Section IV presents the results on PEG and LGPS, including property prediction, stability analysis, and computational benchmarks. Section V discusses limitations and scope, and Section VI summarizes our findings.

Due to planned commercialization, the dataset and source code used in this study are proprietary and will not be made publicly available. However, key algorithms, model architectures, and hyperparameters are described in detail in the manuscript to support reproducibility.

II. Related Work

A. Neural Network Potentials

NNPs are AI models designed to replace computationally expensive quantum mechanical calculations, predicting energy and atomic forces with DFT-level accuracy. A typical NNP architecture, pioneered by Behler and Parrinello [14] and advanced by a subsequent work [13], consists of two components: a descriptor network that encodes atomic environments into numerical features, and a fitting network that maps these features to energy, with forces calculated as the negative gradient. By learning this mapping from DFT data, NNPs provide a computationally efficient alternative for simulating complex molecular systems [2], [3].

Recently, large pre-trained "foundation" NNPs have emerged by training on massive, diverse datasets across elements and thermodynamic conditions. Models such as M3GNet [4], CHGNet [5], and MatterSim [6] improve robustness and transferability for out-of-the-box deployment and downstream specialization. However, their graph-based architectures often incur substantial inference latency and memory footprint, limiting production-scale MD for large atom counts and long timescales [15], [16]. This motivates compressing foundation teachers into lightweight students without sacrificing MD robustness.

B. Data Generation Strategies for NNPs

NNP performance is primarily constrained by training-data coverage over configurational space and energy scale. *Ab initio* MD (AIMD) provides high-quality trajectories but is expensive and often yields redundant configurations [17], [18]. Active learning (AL) [11], [19]–[21] addresses redundancy by iteratively selecting informative structures for labeling and retraining, enabling systematic improvement with fewer labels than naive AIMD sampling.

A recurring finding in recent NNP-MD studies is that high-energy configurations—although rarely visited—are essential for robust dynamics because MD trajectories can encounter them under thermal fluctuations, during rare events, or when the model extrapolates [11], [12]. Insufficient coverage in these regions can lead to unphysical forces and catastrophic simulation failure even when static test errors appear small. In parallel, systematic biases of universal/foundation NNPs in higher-energy regions have been reported, motivating transfer learning and careful data curation for task-specific deployment [10]. These results collectively suggest that *how* training data are generated directly affects MD stability.

C. Knowledge Distillation for NNPs

KD [7] transfers knowledge from a large teacher to a lightweight student, typically by training the student to match teacher predictions. For molecular and materials modeling, KD has been studied for graph neural networks, including approaches that distill intermediate representations to accelerate inference while retaining accuracy [22], [23].

Recent methods such as DPA-2 [8] and PFD [9] established a practical workflow for compressing large pre-trained atomic models into smaller, faster students for task-specific deployment. A common design choice in these pipelines is the *fine-tune-then-distill* order: first fine-tune the foundation model on a small DFT dataset to obtain a specialist teacher, then run teacher-driven NNP-MD to generate trajectories for student training. This order prioritizes teacher accuracy before distillation and often yields strong static validation metrics.

Fine-tune-then-distill pipelines assume that a more accurate teacher also generates better distillation trajectories. Our work challenges this for NNP-MD, where the teacher used for NNP-MD defines the sampling distribution through its PES: fine-tuning can steepen the PES, trap trajectories in low-energy basins, and reduce high-energy coverage needed for dynamical robustness. We therefore decouple *sampling for coverage* from *DFT alignment* by using a non-fine-tuned generalist teacher for trajectory generation and applying DFT correction only via a brief, student-only fine-tuning step.

III. Methodology

In this section, we detail our KD approach for constructing computationally efficient and highly accurate NNP models for a specific target material by decoupling *sampling for coverage* from *alignment to DFT accuracy*. As illustrated in Figure 2, our approach consists of four main steps: Step 1 generates diverse soft targets using a non-fine-tuned generalist teacher, Step 2 distills knowledge into a lightweight student model, Step 3 selects a small subset of structures for DFT calculations to create hard targets, and Step 4 fine-tunes the student model with these hard targets.

Step 1: Soft Target Generation with a Generalist Teacher

We perform NNP-MD simulations using a non-fine-tuned generalist teacher. The key advantage of using a generalist teacher is its smoother PES, which results from training on vast, diverse datasets without specialization for the target material. This smoother landscape lowers effective barriers between configurational states, enabling MD simulations to escape local minima and sample a broader configurational distribution, including higher-energy configurations that are important for dynamical robustness. The MD simulation conditions (temperature, pressure, duration) are chosen to span the thermodynamic ranges relevant to the target applications. From these simulations, we collect a large set of atomic structures—typically over 10,000 configurations. Each structure is labeled with the teacher's predicted energies and forces, forming a *soft target* dataset $\mathcal{D}_{\text{soft}} = \{(\mathbf{X}_i, E_i^t, \mathbf{F}_i^t)\}_{i=1}^{N_{\text{soft}}}$, where $\mathbf{X}_i$ represents atomic coordinates, E_i^t is the teacher's predicted energy, and $\mathbf{F}_i^t$ are the teacher's predicted forces.

Step 2: Knowledge Distillation to the Student Model

We train a lightweight student NNP to mimic the teacher's energy and force predictions on $\mathcal{D}_{\text{soft}}$. The training objective is:

$$\mathcal{L}_{\text{soft}} = \frac{1}{N_{\text{soft}}} \sum_{i=1}^{N_{\text{soft}}} \left[w_E (E_i^s - E_i^t)^2 + w_F \|\mathbf{F}_i^s - \mathbf{F}_i^t\|^2 \right],$$

where E_i^s and $\mathbf{F}_i^s$ denote the student's predictions, and w_E and w_F are energy and force weights, respectively. We employ mean squared error (MSE) rather than classification-based losses (e.g., KL divergence), as MSE is more suitable for regression tasks in NNP modeling [22]. This distillation process serves two critical purposes: (1) it pre-trains the student's descriptor network to encode rich structural representations spanning diverse atomic environments, including high-energy configurations sampled in Step 1, and (2) the trained student becomes a feature extractor for data selection in Step 3. Importantly, while the student learns the general shape of the PES, its predictions are not yet aligned with DFT ground-truth accuracy—this alignment is deferred to Step 4.

Step 3: Strategic Selection of Hard Targets

While the student model has learned the general shape of the PES from the teacher, its predictions are not yet aligned with DFT ground-truth accuracy. To bridge this gap, we generate a small set of *hard targets* $\mathcal{D}_{\text{hard}} = \{(\mathbf{X}_j, E_j^{\text{DFT}}, \mathbf{F}_j^{\text{DFT}})\}_{j=1}^{N_{\text{hard}}}$ by performing DFT calculations on a strategically selected subset of structures from $\mathcal{D}_{\text{soft}}$. Since DFT calculations are computationally expensive, we aim to minimize N_{hard} while maximizing the informativeness of the selected structures. In practice, we use $N_{\text{hard}} \ll N_{\text{soft}}$ (e.g., 1,000 DFT calculations from 20,000 candidate structures). To achieve this, we employ a structural feature-based screening method [11] that combines two diversity criteria:

Feature diversity: We extract high-dimensional structural features from an intermediate layer of the student model for all structures in $\mathcal{D}_{\text{soft}}$. These features capture the atomic environments learned by the student model. To facilitate visualization and selection, we project them into a two-dimensional space using densMAP [24], a density-preserving dimensionality reduction technique. This projection allows us to identify

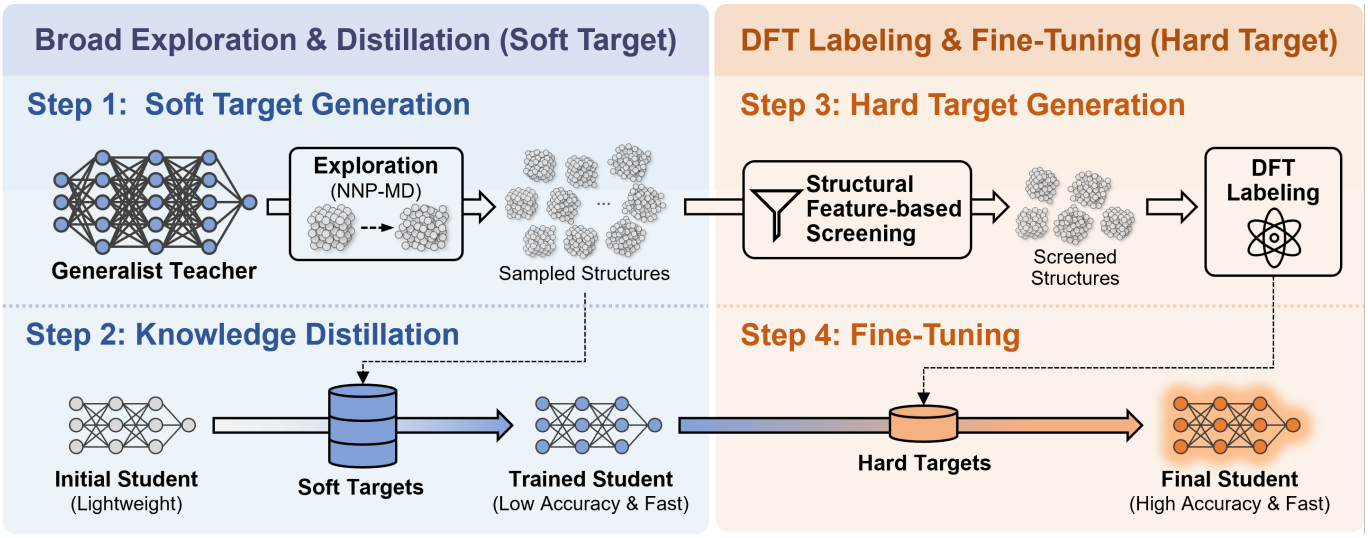

Fig. 2. Overview of our proposed knowledge distillation approach combining broad exploration with targeted refinement. Step 1: A non-fine-tuned generalist teacher performs NNP-MD simulations to generate diverse soft targets spanning wide configurational space. Step 2: A lightweight student NNP is trained to mimic the teacher's predicted energies and forces. Step 3: A small subset of structures is strategically selected for computationally expensive DFT calculations to create hard targets. Step 4: The student model is fine-tuned using only the hard targets to achieve DFT-level accuracy while preserving the broad structural knowledge from distillation.

structures that are widely distributed across the feature space, ensuring coverage of diverse atomic configurations.

Energy diversity: In addition to structural diversity, energy diversity is crucial for generating robust NNP models [25], [26]. We incorporate this by augmenting the two-dimensional feature space with a third dimension: the normalized energy of each structure. This ensures coverage across both configurational space and energy scales.

From this three-dimensional diversity space (2D features + 1D energy), we select N_{hard} structures by maximizing inter-point distances. DFT calculations are performed only on these selected structures to obtain ground-truth labels, forming $\mathcal{D}_{\text{hard}}$.

Step 4: Final Refinement with Hard Targets

We fine-tune the student model using $\mathcal{D}_{\text{hard}}$. The fine-tuning objective is:

$$\mathcal{L}_{\text{hard}} = \frac{1}{N_{\text{hard}}} \sum_{j=1}^{N_{\text{hard}}} \left[w_E (E_j^s - E_j^{\text{DFT}})^2 + w_F \|\mathbf{F}_j^s - \mathbf{F}_j^{\text{DFT}}\|^2 \right],$$

where E_j^s and $\mathbf{F}_j^s$ are the student's predictions, and E_j^{DFT} and $\mathbf{F}_j^{\text{DFT}}$ are ground-truth DFT labels.

This fine-tuning step corrects the student's predictions to match DFT ground-truth accuracy without discarding the broad structural knowledge acquired during distillation. The resulting model achieves an optimal balance: it inherits the generalist teacher's broad configurational coverage from Steps 1–2 and gains quantitative DFT-level accuracy from Steps 3–4. This combination yields a final student NNP that is computationally efficient, robust across diverse atomic environments, and accurate enough for production-level MD simulations.

IV. Results

We evaluate the proposed KD workflow on two material systems, PEG and LGPS. We report (i) accuracy on MD-relevant target properties and (ii) data efficiency in terms of required DFT labels for both systems, and we further provide an explicit MD stability analysis in the PEG case study.

A. Case Study: Polyethylene Glycol (PEG)

1) Experimental Setup: We used MatterSim [6] (MatterSim-v1.0.0-5M) as the generalist teacher. MatterSim was trained on a large dataset spanning high-temperature and high-pressure conditions, which makes it suitable for generating diverse configurations over a wide energy range, including higher-energy structures needed for robust MD. For soft target generation, we performed NVT simulations at four temperatures (300, 400, 500, 600 K) and five densities (±5% around 1.120 g/cm^3) for 150 ps each with a 0.5 fs timestep, collecting 20,000 structures (16,000 training, 4,000 validation) every 100 fs from the final 100 ps. For the student model, we used the DeepPot-SE (DP) architecture [13], implemented within the DeePMD-kit framework [27], [28], with a 6 Å cutoff radius. The descriptor network has three layers (25, 50, 100 nodes) and the fitting network has three layers (240, 240, 240 nodes), both with ResNet-like skip connections. We used the Adam optimizer for 500,000 steps with batch size 4. The learning rate decayed exponentially from 1.0×10^{-3} to 1.0×10^{-8} every 5,000 steps. The loss function combines mean squared errors on energies (weight increasing from 0.02 to 1.0) and forces (weight decreasing from 1000.0 to 1.0). DFT calculations were performed using Quantum ESPRESSO with 100 Ry plane-wave cutoff and Γ-point sampling. We used the BLYP+D3 functional [29], [30] with HGH pseudopotentials [31]. After distillation, we

TABLE I

COMPARISON OF DATA EFFICIENCY AND ACCURACY FOR PEG. OUR APPROACH ACHIEVES STATE-OF-THE-ART ACCURACY USING ONLY 1,000 HARD TARGETS (10× DATA EFFICIENCY). VALUES IN PARENTHESES ARE PERCENTAGE DEVIATIONS FROM EXPERIMENTAL DATA.

Method	**# of hard targets**	**Density (g/cm^3)**	**Diffusion (10^{-6} cm^2/s)**
Experiments [32]	—	1.120	0.297
QRNN [33]	344,654	**1.128 (1%)**	0.390 (31%)
GeNNIP4MD [11]	9,987	1.141 (2%)	0.235 (-21%)
MatterSim (Teacher) [6]	—	0.743 (-34%)	9.676 (3,158%)
This work	**1,000**	1.132 (1%)	**0.334 (12%)**

TABLE II

ABLATION STUDY RESULTS FOR PEG SYSTEM. BOTH SOFT AND HARD TARGETS ARE NECESSARY, AND INTELLIGENT SELECTION OF HARD TARGETS IS CRITICAL FOR CONSISTENT ACCURACY.

Setting	**# of hard targets**	**Density (g/cm^3)**	**Diffusion (10^{-6} cm^2/s)**
Necessity of both soft and hard targets			
Soft targets only	—	0.994 (-11%)	0.775 (161%)
Hard targets only	1,000	0.008 (-99%)	8,373.0 (>10^6%)
Full approach	1,000	**1.132 (1%)**	**0.334 (12%)**
Impact of hard target quantity			
100 hard targets	100	1.237 (10%)	0.003 (-99%)
500 hard targets	500	1.031 (-8%)	0.741 (149%)
1,000 hard targets	1,000	**1.132 (1%)**	**0.334 (12%)**
Importance of intelligent data selection			
Random selection	1,000	1.144 (2%)*	0.216 (-27%)*
Feature-based	1,000	**1.132 (1%)**	**0.334 (12%)**

*Mean of 3 random trials; std dev: ±0.012 g/cm^3, ±0.063 $\times 10^{-6}$ cm^2/s

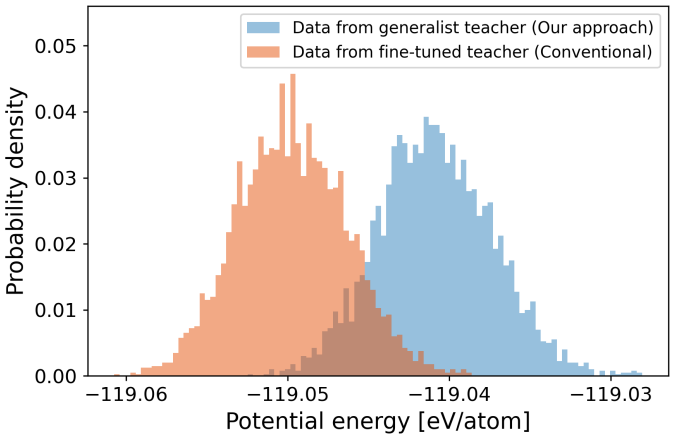

Fig. 3. Energy distributions of training data generated by different teacher approaches, relabeled by DFT for fair comparison. Our generalist teacher (blue) effectively samples high-energy structures (39.8% above -119.040 eV/atom), while the conventional fine-tuned teacher (orange) is largely restricted to low-energy regions (only 0.3% in the same range), confirming that fine-tuning steepens the PES and limits exploration of critical high-energy configurations.

fine-tuned the student model using only 1,000 DFT-labeled data points (hard targets), efficiently selected using the structural feature-based screening method.

2) Accuracy and Data Efficiency: We evaluate the final student model by running large-scale NNP-MD simulations and computing two target properties: density and self-diffusion coefficient. For property evaluation, we performed a 20 ns NNP-MD production run (Δt = 0.5 fs) in the NPT ensemble at 298.15 K and 1 bar using a 3,100-atom PEG system. Table I compares our results with experiments and representative baselines.

MatterSim in a zero-shot setting is not accurate for PEG, showing a 34% density error, which motivates task-specific specialization. After distillation and student-only fine-tuning with 1,000 DFT-labeled structures, our final student achieves near-experimental agreement: 1% error in density and 12% error in self-diffusion. The most significant advantage lies in its remarkable data efficiency. With only 1,000 hard targets, our approach matches or outperforms prior NNP workflows that rely on substantially more DFT labels, such as GeNNIP4MD (9,987) and QRNN (344,654), corresponding to at least a 10× reduction in DFT labeling.

These results support the central design choice of our workflow: broad teacher-driven sampling provides configurational coverage, while student-only DFT fine-tuning restores quantitative accuracy. In other words, the property accuracy is achieved without requiring dense DFT coverage across the full MD distribution.

3) Ablation Study: Table II isolates the contribution of each component. Training only on soft targets yields insufficient quantitative accuracy (-11% density error), while training from scratch on only hard targets leads to catastrophic MD behavior, indicating that distillation pre-training is required for stable deployment. The number of hard targets also matters: 100 and 500 labels are insufficient, whereas 1,000 labels recover near-experimental accuracy. Finally, feature-based selection is consistently better than random selection across three trials, improving both accuracy and stability.

4) Comparison with Conventional Fine-Tuned Teacher Approach: We directly test our hypothesis by comparing against the standard fine-tune-then-distill pipeline used in DPA-2 [8] and PFD [9]: (1) fine-tune the teacher on 1,000 DFT-labeled structures, (2) run NNP-MD with the fine-tuned teacher to generate distillation trajectories, and (3) train a student on these trajectories.

For this comparison, we define MD failure as trajectory breakdown leading to unphysical density collapse or numerical instability within the simulated timescale.

On static metrics, the conventional pipeline appears competitive. The fine-tuned teacher achieves a force MAE of 0.030 eV/Å, and the student trained on its data reaches a validation force MAE of 0.063 eV/Å, close to 0.061 eV/Å for our final student. However, this similarity is misleading: we observe a critical failure case when this conventionally-trained student model is deployed for an actual MD simulation. The system disintegrates, resulting in a near-vacuum density (0.015 g/cm^3) within 0.5 ns. This demonstrates that static validation error does not guarantee dynamical stability.

Figure 3 explains this gap. The fine-tuned teacher gener-

ates a dataset concentrated in low-energy regions, whereas the generalist teacher covers substantially more of the high-energy regime. Quantitatively, only 0.3% of structures from the fine-tuned teacher lie above -119.040 eV/atom, compared to 39.8% for the generalist teacher. This lack of high-energy samples leaves the student under-trained for configurations that challenge stability during MD, leading to catastrophic failure despite a low validation MAE [11], [12].

This failure is consistent with the sampling-collapse mechanism in Fig. 1 and the coverage gap in Fig. 3. It also explains why static validation MAE is insufficient: it does not reflect whether the training distribution includes the rare but destabilizing higher-energy configurations encountered during MD.

B. Case Study: $Li_{10}GeP_2S_{12}$ (LGPS)

1) Experimental Setup: We evaluate our proposed workflow on LGPS, an inorganic solid-state electrolyte. We employed the same generalist teacher (MatterSim-v1.0.0-5M) and student model architecture (DeepPot-SE) as in the PEG case study. For soft target generation, NVT simulations were conducted at five temperatures (300, 600, 900, 1200, 1500 K) for 110 ps each, yielding 10,000 structures (8,000 training, 2,000 validation) sampled every 50 fs from the final 100 ps. DFT calculations were performed with the same computational settings as PEG, but using the PBE functional [34] with PAW method [35]. The student model was fine-tuned with only 1,000 DFT-calculated hard targets.

2) Accuracy and Data Efficiency: We evaluate LGPS using the lithium-ion self-diffusion coefficient, a key property for battery electrolytes that is sensitive to NNP accuracy. For diffusion evaluation, we ran 1 ns NNP-MD production simulations (Δt = 0.5 fs) in the NVT ensemble using a 1,600-atom LGPS system at the temperatures shown in Figure 4.

Figure 4 reports the Arrhenius behavior predicted by our final student across a wide temperature range. The final student model (red circles) agrees well with experimental measurements [36] and AIMD references [37] (black diamonds) over the full temperature range, despite using only 1,000 hard targets for DFT alignment. This yields a substantial improvement in data efficiency over other state-of-the-art NNP generation methods: we achieve comparable or better accuracy with 1,000 hard targets, compared to 5,279 in GeNNIP4MD [11] and 17,268 in the PaiNN-based model [38] (i.e., over 5× and 17× fewer DFT labels, respectively).

Figure 4 also clarifies the role of the two-step workflow. The student model trained only on soft targets (green pentagons) inherits the teacher's bias and overestimates diffusion in the 300–500 K regime, whereas the subsequent student-only fine-tuning with hard targets corrects this bias and aligns predictions with the ground truth. Consistent with this correction, the static force MAE improves from 0.088 eV/Å to 0.054 eV/Å after fine-tuning, providing a more reliable basis for diffusion prediction.

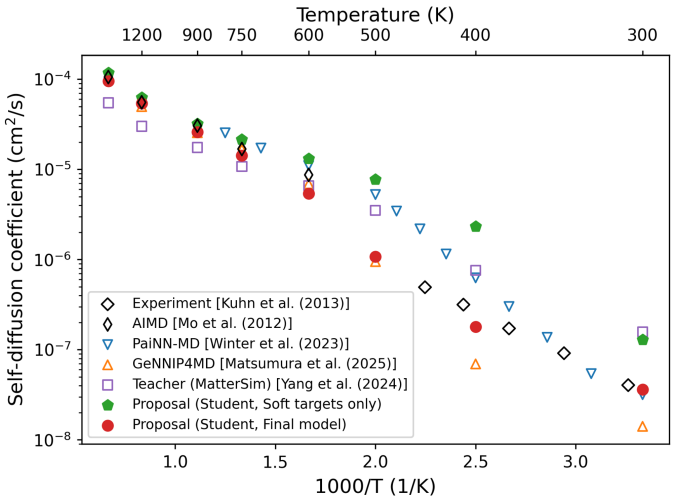

Fig. 4. Arrhenius plot of lithium-ion diffusion coefficient for LGPS. The final student model from the proposed approach achieves excellent agreement with experimental and high-cost AIMD reference data. It demonstrates superior data efficiency compared to other state-of-the-art NNP models (PaiNN-MD, GeNNIP4MD). The student model trained on solely soft targets overestimates them in the 300–500 K range. In contrast, the final student model fine-tuned with only 1,000 hard targets demonstrates excellent agreement with experimental and AIMD values, highlighting the efficiency of our proposed approach.

C. Computational Performance Analysis

We evaluate the computational benefit of KD from two perspectives: (i) inference performance for NNP-MD (runtime and memory constraints), and (ii) end-to-end time to generate a deployable student model.

For inference, we measure the average execution time of 1000-step NNP-MD simulations over five trials on an NVIDIA H100 GPU and compare the per-step time of the teacher (MatterSim) and the distilled student (DP) for both PEG and LGPS systems. Figure 5 summarizes the results. KD yields large inference gains. The student model achieves a speedup of 10× to 106× for PEG and 5× to 46× for LGPS. More importantly for large-scale deployment, the teacher model runs out of memory for PEG systems larger than 4,096 atoms, whereas the student model enables stable simulations up to approximately 20,000 atoms. For the specific property-evaluation systems in this work (3,100-atom PEG and 1,600-atom LGPS), the observed speedups are 82× and 20×, respectively.

We also compare the total wall-clock time required to construct the final student model against a conventional active-learning method (GeNNIP4MD [11]). Table III reports the breakdown into structure sampling, DFT labeling, and training on an NVIDIA V100 GPU. Overall, our workflow reduces the total NNP generation time by 1.9× for PEG and 3.0× for LGPS compared to active learning. The main tradeoff is that structure sampling of soft targets is more expensive because it uses the large GNN-based teacher (e.g., 123 h vs. 38 h for PEG). However, this upfront cost is offset by (i) substantially reduced DFT labeling (10× fewer DFT calculations, saving 93 h for PEG) and (ii) avoiding the iterative retraining cycles

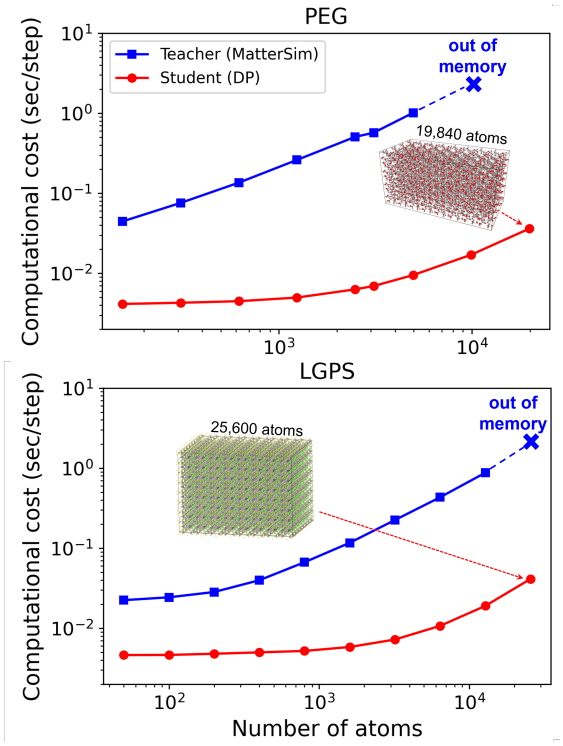

Fig. 5. Computational cost of the teacher model (MatterSim) and the student model (DP) on an NVIDIA H100 GPU. The distilled student model not only achieves a speedup of up to 106× but, more importantly, overcomes the memory limitations of the teacher, enabling simulations of much larger systems.

TABLE III
NNP GENERATION TIME COMPARISON ON AN NVIDIA V100 GPU. VALUES IN PARENTHESES INDICATE SPEEDUP RELATIVE TO GENNIP4MD.

System	Method	Structure sampling	DFT labeling	Training	Total
PEG	GeNNIP4MD	38 h	101 h	121 h	260 h
	This work	123 h	8 h	7 h	**138 h (1.9×)**
LGPS	GeNNIP4MD	4 h	132 h	99 h	235 h
	This work	45 h	25 h	8 h	**78 h (3.0×)**

inherent to active learning (saving 114 h for PEG). In our workflow, the student is trained once on the teacher-labeled dataset and then briefly fine-tuned, which streamlines training and reduces total NNP generation time.

V. LIMITATIONS AND SCOPE

This study evaluates a single generalist teacher (MatterSim) on two target systems (PEG and LGPS). While the results consistently support the central mechanism studied here—that fine-tuning the teacher can reshape the PES, reduce high-energy coverage during teacher-driven MD, and harm MD robustness—we do not claim that the same behavior will hold uniformly for all foundation NNPs or materials. Extending the evaluation to additional teachers (e.g., M3GNet, CHGNet, DPA-2) and broader out-of-distribution settings is an important direction for future work.

The preferred workflow may also depend on the target application. When the primary requirement is extreme quantitative accuracy in well-sampled regions and the MD sampling distribution is not a limiting factor, accuracy-first pipelines (fine-tuning the teacher before distillation) may remain competitive. In contrast, our workflow is designed for settings where reliable MD requires preserving coverage of higher-energy configurations encountered during MD.

Finally, the choice of 1,000 DFT-labeled structures for student fine-tuning reflects a practical tradeoff between labeling cost and accuracy in our experiments. Our ablations indicate that this number is sufficient for PEG and LGPS under the studied conditions, but other materials or thermodynamic regimes may require adjusting the number of DFT labels and/or the selection strategy.

VI. CONCLUSION

In this work, we presented a KD workflow for building lightweight NNPs that are accurate and robust for NNP-driven MD. The key design choice is to decouple *sampling for configurational coverage* from *alignment to DFT accuracy*: we generate diverse training trajectories using a non-fine-tuned generalist teacher, distill a lightweight student from teacher-labeled energies and forces, and then apply DFT alignment via student-only fine-tuning on a small set of strategically selected hard targets.

Across two representative systems (PEG and LGPS), the resulting student achieves near-experimental property accuracy while reducing expensive DFT labeling by up to an order of magnitude. The student also provides substantial deployment gains, including large inference speedups and improved memory scalability compared with the foundation teacher, enabling larger and longer MD simulations.

A central implication, demonstrated explicitly in PEG, is that static validation errors alone are not reliable predictors of MD robustness: models with similar static accuracy can behave very differently in long MD runs if their training data lack sufficient coverage of higher-energy configurations. By explicitly preserving configurational coverage during trajectory generation and applying DFT correction at the student stage, our workflow provides a simple and deployable procedure to obtain fast, accurate, and MD-stable distilled NNPs.

REFERENCES

[1] S. Mohr, L. E. Ratcliff, L. Genovese, D. Caliste, P. Boulanger, S. Goedecker, and T. Deutsch, "Accurate and efficient linear scaling DFT calculations with universal applicability," *Physical Chemistry Chemical Physics*, vol. 17, no. 47, pp. 31 360–31 370, 2015.

[2] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine Learning Force Fields," *Chemical Reviews*, vol. 121, no. 16, pp. 10 142–10 186, 2021.

[3] P. Friederich, F. Häse, J. Proppe, and A. Aspuru-Guzik, "Machine-learned potentials for next-generation matter simulations," *Nature Materials*, vol. 20, no. 6, pp. 750–761, 2021.

[4] C. Chen and S. P. Ong, "A universal graph deep learning interatomic potential for the periodic table," *Nature Computational Science*, vol. 2, no. 11, pp. 718–728, 2022.

[5] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, "CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling," *Nature Machine Intelligence*, vol. 5, no. 9, pp. 1031–1041, 2023.

[6] H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, M. Horton, R. Pinsler, A. Fowler, D. Zügner, T. Xie, J. Smith, L. Sun, Q. Wang, L. Kong, C. Liu, H. Hao, and Z. Lu, "MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures," 2024, arXiv:2405.04967.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," 2015, arXiv:1503.02531.

[8] D. Zhang, X. Liu, X. Zhang, C. Zhang, C. Cai, H. Bi, Y. Du, X. Qin, A. Peng, J. Huang, B. Li, Y. Shan, J. Zeng, Y. Zhang, S. Liu, Y. Li, J. Chang, X. Wang, S. Zhou, J. Liu, X. Luo, Z. Wang, W. Jiang, J. Wu, Y. Yang, J. Yang, M. Yang, F.-Q. Gong, L. Zhang, M. Shi, F.-Z. Dai, D. M. York, S. Liu, T. Zhu, Z. Zhong, J. Lv, J. Cheng, W. Jia, M. Chen, G. Ke, W. E, L. Zhang, and H. Wang, "DPA-2: a large atomic model as a multi-task learner," *npj Computational Materials*, vol. 10, no. 1, p. 293, 2024.

[9] R. Wang, Y. Gao, H. Wu, and Z. Zhong, "PFD: Automatically Generating Machine Learning Force Fields from Universal Models," 2025, arXiv:2502.20809.

[10] B. Deng, Y. Choi, P. Zhong, J. Riebesell, S. Anand, Z. Li, K. Jun, K. A. Persson, and G. Ceder, "Systematic softening in universal machine learning interatomic potentials," *npj Computational Materials*, vol. 11, no. 1, p. 9, 2025.

[11] N. Matsumura, Y. Yoshimoto, T. Yamazaki, T. Amano, T. Noda, N. Ebata, T. Kasano, and Y. Sakai, "Generator of Neural Network Potential for Molecular Dynamics: Constructing Robust and Accurate Potentials with Active Learning for Nanosecond-Scale Simulations," *Journal of Chemical Theory and Computation*, vol. 21, no. 8, pp. 3832–3846, 2025.

[12] Y. Yoshimoto, N. Matsumura, Y. Iwasaki, H. Nakao, and Y. Sakai, "Large-Scale, Long-Time Atomistic Simulations of Proton Transport in Polymer Electrolyte Membranes Using a Neural Network Interatomic Potential," 2025, arXiv:2503.20412.

[13] L. Zhang, J. Han, H. Wang, W. Saidi, R. Car, and W. E, "End-to-end Symmetry Preserving Inter-atomic Potential Energy Model for Finite and Extended Systems," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[14] J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces," *Physical Review Letters*, vol. 98, no. 14, p. 146401, 2007.

[15] D. Wines and K. Choudhary, "CHIPS-FF: Evaluating Universal Machine Learning Force Fields for Material Properties," *ACS Materials Letters*, vol. 7, no. 6, pp. 2105–2114, 2025.

[16] R. Jacobs, D. Morgan, S. Attarian, J. Meng, C. Shen, Z. Wu, C. Y. Xie, J. H. Yang, N. Artrith, B. Blaiszik, G. Ceder, K. Choudhary, G. Csanyi, E. D. Cubuk, B. Deng, R. Drautz, X. Fu, J. Godwin, V. Honavar, O. Isayev, A. Johansson, B. Kozinsky, S. Martiniani, S. P. Ong, I. Poltavsky, K. Schmidt, S. Takamoto, A. P. Thompson, J. Westermayr, and B. M. Wood, "A practical guide to machine learning interatomic potentials – Status and future," *Current Opinion in Solid State and Materials Science*, vol. 35, p. 101214, 2025.

[17] R. Chahal, M. D. Toomey, L. T. Kearney, A. Sedova, J. T. Damron, A. K. Naskar, and S. Roy, "Deep-Learning Interatomic Potential Connects Molecular Structural Ordering to the Macroscale Properties of Polyacrylonitrile," *ACS Applied Materials & Interfaces*, vol. 16, no. 28, pp. 36 878–36 891, 2024.

[18] D. Liu, Y. Wu, M. R. Samatov, A. S. Vasenko, E. V. Chulkov, and O. V. Prezhdo, "Compression Eliminates Charge Traps by Stabilizing Perovskite Grain Boundary Structures: An Ab Initio Analysis with Machine Learning Force Field," *Chemistry of Materials*, vol. 36, no. 6, pp. 2898–2906, 2024.

[19] E. V. Podryabinkin and A. V. Shapeev, "Active learning of linearly parametrized interatomic potentials," *Computational Materials Science*, vol. 140, pp. 171–180, 2017.

[20] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241733, 2018.

[21] Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang, and W. E, "DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models," *Computer Physics Communications*, vol. 253, p. 107206, 2020.

[22] F. E. Kelvinius, D. Georgiev, A. P. Toshev, and J. Gasteiger, "Accelerating molecular graph neural networks via knowledge distillation," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2023.

[23] I. Amin, S. Raja, and A. S. Krishnapriyan, "Towards fast, specialized machine learning force fields: Distilling foundation models via energy hessians," in *The Thirteenth International Conference on Learning Representations*, 2025.

[24] A. Narayan, B. Berger, and H. Cho, "Assessing single-cell transcriptomic variability through density-preserving data visualization," *Nature Biotechnology*, vol. 39, no. 6, pp. 765–774, 2021.

[25] J. Qi, T. W. Ko, B. C. Wood, T. A. Pham, and S. P. Ong, "Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling," *npj Computational Materials*, vol. 10, no. 1, p. 43, 2024.

[26] A. D. Kaplan, R. Liu, J. Qi, T. W. Ko, B. Deng, J. Riebesell, G. Ceder, K. A. Persson, and S. P. Ong, "A Foundational Potential Energy Surface Dataset for Materials," 2025.

[27] H. Wang, L. Zhang, J. Han, and W. E, "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics," *Computer Physics Communications*, vol. 228, pp. 178–184, 2018.

[28] J. Zeng, D. Zhang, D. Lu, P. Mo, Z. Li, Y. Chen, M. Rynik, L. Huang, Z. Li, S. Shi, Y. Wang, H. Ye, P. Tuo, J. Yang, Y. Ding, Y. Li, D. Tisi, Q. Zeng, H. Bao, Y. Xia, J. Huang, K. Muraoka, Y. Wang, J. Chang, F. Yuan, S. L. Bore, C. Cai, Y. Lin, B. Wang, J. Xu, J.-X. Zhu, C. Luo, Y. Zhang, R. E. A. Goodall, W. Liang, A. K. Singh, S. Yao, J. Zhang, R. Wentzcovitch, J. Han, J. Liu, W. Jia, D. M. York, W. E, R. Car, L. Zhang, and H. Wang, "DeePMD-kit v2: A software package for deep potential models," *The Journal of Chemical Physics*, vol. 159, no. 5, p. 054801, 2023.

[29] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials," *Journal of Physics. Condensed Matter: An Institute of Physics Journal*, vol. 21, no. 39, p. 395502, 2009.

[30] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," *The Journal of Chemical Physics*, vol. 132, no. 15, p. 154104, 2010.

[31] C. Hartwigsen, S. Goedecker, and J. Hutter, "Relativistic separable dual-space Gaussian pseudopotentials from H to Rn," *Physical Review B*, vol. 58, no. 7, pp. 3641–3662, 1998.

[32] M. M. Hoffmann, R. H. Horowitz, T. Gutmann, and G. Buntkowsky, "Densities, Viscosities, and Self-Diffusion Coefficients of Ethylene Glycol Oligomers," *Journal of Chemical & Engineering Data*, vol. 66, no. 6, pp. 2480–2500, 2021.

[33] S. Mohanty, J. Stevenson, A. R. Browning, L. Jacobson, K. Leswing, M. D. Halls, and M. A. F. Afzal, "Development of scalable and generalizable machine learned force field for polymers," *Scientific Reports*, vol. 13, no. 1, p. 17251, 2023.

[34] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Physical Review Letters*, vol. 77, no. 18, pp. 3865–3868, 1996.

[35] P. E. Blöchl, "Projector augmented-wave method," *Physical Review B*, vol. 50, no. 24, pp. 17 953–17 979, 1994.

[36] A. Kuhn, V. Duppel, and B. V. Lotsch, "Tetragonal Li10GeP2S12 and Li7GePS8 – exploring the Li ion dynamics in LGPS Li electrolytes," *Energy & Environmental Science*, vol. 6, no. 12, pp. 3548–3552, 2013.

[37] Y. Mo, S. P. Ong, and G. Ceder, "First Principles Study of the $Li_{10}GeP_2S_{12}$ Lithium Super Ionic Conductor Material," *Chemistry of Materials*, vol. 24, no. 1, pp. 15–17, 2012.

[38] G. Winter and R. Gómez-Bombarelli, "Simulations with machine learning potentials identify the ion conduction mechanism mediating non-Arrhenius behavior in LGPS," *Journal of Physics: Energy*, vol. 5, no. 2, p. 024004, 2023.