

Supporting Information for

A Perspective on Training Machine Learning Force Fields for Solid-State Electrolyte Materials

Zihan Yan^{1,2} , Shengjie Tang^{1,2} , and Yizhou Zhu^{2,*} 

¹*School of Materials Science and Engineering, Zhejiang University, Hangzhou, China*

²*Department of Materials Science and Engineering, Westlake University, Hangzhou, China*

*Corresponding author: Yizhou Zhu (zhuyizhou@westlake.edu.cn)

Contents

Supplementary Notes: Additional computational details	S2
Supplementary Note 1. Additional computational details for Figure 1	S2
Supplementary Note 2. Additional computational details for Figure 2	S3
Supplementary Note 3. Additional computational details for Figure 3	S3
Supplementary Note 4. Additional computational details for Figure 4	S4
Supplementary Note 5. Additional computational details for Figure 5	S4
Arrhenius plots calculated by NEP models trained on different data sizes	S5
Dimer dissociation curve tests for NEP models trained on different datasets	S6
Validation of MLFF Transport Predictions against AIMD	S7
Data efficiency of MACE model	S8
Data efficiency of GNEP model	S9
Data efficiency of fine-tuned NEP89 foundation model	S10
Data efficiency of fine-tuned MACE foundation model	S11
Arrhenius plots calculated by NEP models trained with different random seeds	S12
Arrhenius plots of LGPS calculated by MACE models trained with different random seeds	S13
Data quality sensitivity of fine-tuned foundation models	S14
Large-supercell Γ -point validation for LGPS	S15
Structural and vibrational validation of MLFF models	S16
Speed benchmark of NEP models with larger cutoff	S18

Supplementary Notes: Additional computational details

Supplementary Note 1. Additional computational details for Figure 1

Dataset Construction and MLFF Training:

The reference datasets for training were obtained from previous studies: the $\text{Li}_7\text{La}_3\text{Zr}_2\text{O}_{12}$ (LLZO) dataset was sourced from Ref. [1], the Li_3YCl_6 (LYC) dataset from Ref. [2], and the $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ (LGPS) dataset from Ref. [3]. To evaluate data efficiency, we then systematically reduced the dataset using the farthest-point sampling (FPS) algorithm implemented in our GPUMDkit package [4], progressively subsampling by factors of $1/2^n$ ($n = 1, 2, \dots$). Each reduced dataset was used to train a NEP model, with the full 1,978-structure dataset serving as a test set to evaluate how prediction accuracy degrades with decreasing training data.

All machine learning force fields (MLFFs) in this section were trained using the Neuroevolution Potential (NEP) framework as implemented in *Graphics Processing Units Molecular Dynamics* (GPUMD) package [5]. The hyperparameters for the NEP models were configured as follows:

- **Common Settings:** For all materials, the `basis_size` was set to (12, 12), `l_max` to (4, 2, 1), `zbl` was enabled with a value of 2, and the models were evolved for 100,000 generations.
- **LLZO:** A cutoff of (6, 5) Å and a batch size of 200 were used.
- **LYC:** A cutoff of (7.5, 5) Å and a batch size of 500 were used.
- **LGPS:** A cutoff of (6, 5) Å and a batch size of 500 were used.

All other parameters were kept at their default values as specified in the `nep` program. Also, all the training files can be found in the <https://github.com/zhyan0603/SourceFiles>.

Molecular Dynamics Simulations:

Ionic transport properties were calculated via large-scale MD simulations using the GPUMD package. For all systems, the equations of motion were integrated using a time-step of 1 fs (for LLZO) or 2 fs (for LYC and LGPS). Temperature and pressure were controlled by the MTTK (Martyna–Tuckerman–Tobias–Klein) thermostat and barostat under the NPT ensemble.

- **LLZO:** Simulations were performed on a $4 \times 4 \times 4$ supercell (12,288 atoms) at temperatures of 950, 1000, 1100, and 1200 K. Each production run lasted 1 ns to collect Mean Square Displacement (MSD) data.
- **LYC:** An orthorhombic supercell (3,840 atoms) was employed. Production runs were conducted at 450 K (3 ns), 500 K (2 ns), 550 K (2 ns), and 600 K (2 ns).
- **LGPS:** A $6 \times 6 \times 6$ supercell (10,800 atoms) was used. Simulations were carried out at 450 K, 500 K, 550 K, and 600 K for 2 ns.

The self-diffusion coefficients (D) and activation energies (E_a) were extracted from the MSD and Arrhenius plots, respectively, using our GPUMDkit package [4].

Supplementary Note 2. Additional computational details for Figure 2

We tested how k -point density in the reference DFT calculations affects the performance of trained force fields for LLZO, LYC, and LGPS. We selected 50, 150, and 150 structures from their datasets using the FPS algorithm and performed DFT calculations at two levels of accuracy: high-accuracy calculations with $KSPACING = 0.25 \text{ \AA}^{-1}$, and low-accuracy Γ -point-only calculations.

All the DFT calculations were performed using the Vienna *Ab initio* Simulation Package (VASP) [6, 7], based on density functional theory (DFT) [8, 9] within the projector-augmented wave (PAW) method [10]. The exchange-correlation interactions were described by the PBEsol functional for LLZO and LGPS [1, 3], whereas the PBE+optB88b-vdW was employed for LYC following the previous study [2]. The energy cutoff of the plane wave basis was set to 520 eV. The electronic energy convergence criterion was set to 10^{-5} eV. The Γ -centered k -point meshes with a density of $KSPACING = 0.25 \text{ \AA}^{-1}$ in the Brillouin zone were used for electronic self-consistent calculations. For the *ab initio* molecular dynamics simulations, the same settings were adopted except that only the Γ -point was used for Brillouin zone sampling and the energy convergence criterion was relaxed to 10^{-4} eV.

Supplementary Note 3. Additional computational details for Figure 3

Force Field Training:

The NEP and qNEP models were trained using the `nep` program within the GPUMD package. Training hyperparameters were kept consistent with those used for Figure 1, with the qNEP model specifically employing the `charge_mode 2` option in the `nep.in` file to activate the Latent Ewald Summation (LES) method [11] for long-range interactions. Given that the NEP framework utilizes two regularization terms to prevent overfitting, all available data were used for the training set without a separate validation set.

In contrast, the MACE variants (T0-I, T0-E, T1-E, and T1-E-LES) were trained with a dedicated validation set. We deviated from the standard two-stage training strategy (which typically shifts focus from forces to energies) due to observed numerical instabilities; specifically, MACE-T1-E-LES exhibited significant energy fluctuations during the second stage after incorporating LES. Consequently, a one-stage training protocol was adopted with both energy and force weights set to 1000. All trained models, configuration files, and training parameters are available in our GitHub repository for reproducibility.

Molecular Dynamics Simulations:

MD simulations were performed to investigate the Li-ion transport properties. To accommodate the high computational cost of the MACE models, relatively small supercells were utilized for all systems: 192 atoms for LLZO, 120 atoms for LYC, and 200 atoms for LGPS.

- **Lattice Constants:** First, the coefficient of thermal expansion for each material is determined using the NEP model. Then, the lattice parameters are fixed to values corresponding to the target temperature for subsequent simulations.
- **Ensemble and Integration:** All MD runs were conducted in the NVT ensemble using a timestep of 2 fs.
- **Data Collection:** To ensure statistical convergence of the MSD, the simulation duration was set to 1 ns for LLZO and 2 ns for LYC and LGPS. The Arrhenius plots were then constructed based on these trajectories to derive activation energies (E_a).

Supplementary Note 4. Additional computational details for Figure 4

To quantify this efficiency consideration, we benchmarked the computational efficiency of all models on a single NVIDIA V100 32GB GPU at the Open Source Supercomputing Center of S-A-I. NEP and qNEP were executed in GPUMD package, a highly optimized MD package implemented in CUDA C++, while MACE models were run in LAMMPS package [12], except for MACE-T1-E-LES which was evaluated using ASE package [13] due to implementation constraints.

Supplementary Note 5. Additional computational details for Figure 5

To evaluate the spatial locality of the interatomic potentials, we employed a two-zone perturbation scheme. For a selected central atom, the system is divided into two concentric regions: an inner fixed zone of radius $2 \times R_{\text{cut}}$ where all atomic positions remain frozen, and an outer perturbation zone where atoms undergo thermal motion. Unlike conventional manual random displacements, we utilized MD-based perturbations to sample configurations from the Boltzmann distribution. Perturbation temperatures were set to 1000 K for LLZO and 500 K for LYC and LGPS to ensure sufficient sampling of the local potential energy surface.

We performed locality tests using large supercells, specifically an LLZO supercell containing 5184 atoms, an orthorhombic LYC supercell containing 3840 atoms, and an LGPS supercell containing 2400 atoms. Due to the prohibitive cost of DFT at this scale, the MACE-T1-E model was employed as the DFT surrogate. Its message-passing architecture allows it to capture interactions up to approximately $4 \times R_{\text{cut}}$, providing a reliable benchmark for evaluating the $2 \times R_{\text{cut}}$ locality.

The sensitivity of the force on a central atom to environmental changes beyond $2 \times R_{\text{cut}}$ is quantified by the locality error (σ_f), defined as the standard deviation of the atomic force vector $\mathbf{f}$ across N sampled configurations:

$$\sigma_f = \sqrt{\frac{1}{N} \sum_{i=1}^N \|\mathbf{f}_i - \bar{\mathbf{f}}\|^2} \quad (\text{S1})$$

where $\bar{\mathbf{f}} = \frac{1}{N} \sum \mathbf{f}_i$ represents the mean force vector and $\|\cdot\|$ denotes the Euclidean norm.

Arrhenius plots calculated by NEP models trained on different data sizes

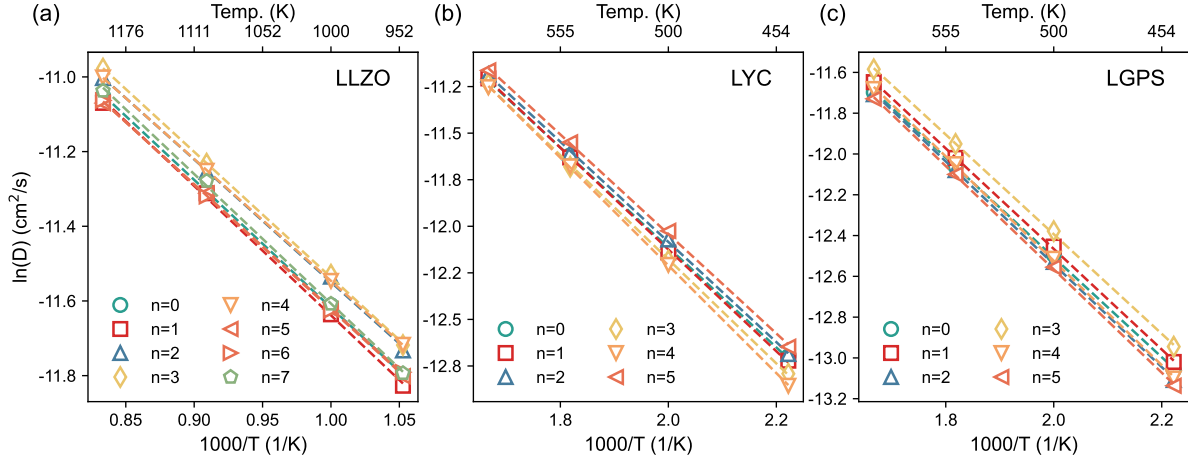


Figure S1: Arrhenius plots of Li^+ ions diffusion calculated by different NEP models trained on different data sizes in (a) LLZO, (b) LYC and (c) LGPS.

To provide a more comprehensive view of how dataset size affects the predicted ion transport properties, we present the full Arrhenius plots for LLZO, LYC, and LGPS in Figure S1. Here, n denotes the factor of dataset reduction, where the training set size is scaled by $1/2^n$. The results show that for all three solid-state electrolyte materials, the Arrhenius plots derived from models trained on significantly reduced datasets (up to $n = 7$) remain remarkably consistent with those from the full-dataset models ($n = 0$).

Specifically, the data points for different n values cluster tightly. This consistency confirms that even when the training data is reduced to approximately 1,000 local Li-ion environments, the NEP model can capture the fundamental diffusivities and activation energies. These findings further substantiate our conclusion that the potential energy surfaces of bulk SSEs are easy to learn, and the bottleneck for accurate transport prediction lies in the quality of the reference data rather than the quantity of configurations.

Dimer dissociation curve tests for NEP models trained on different datasets

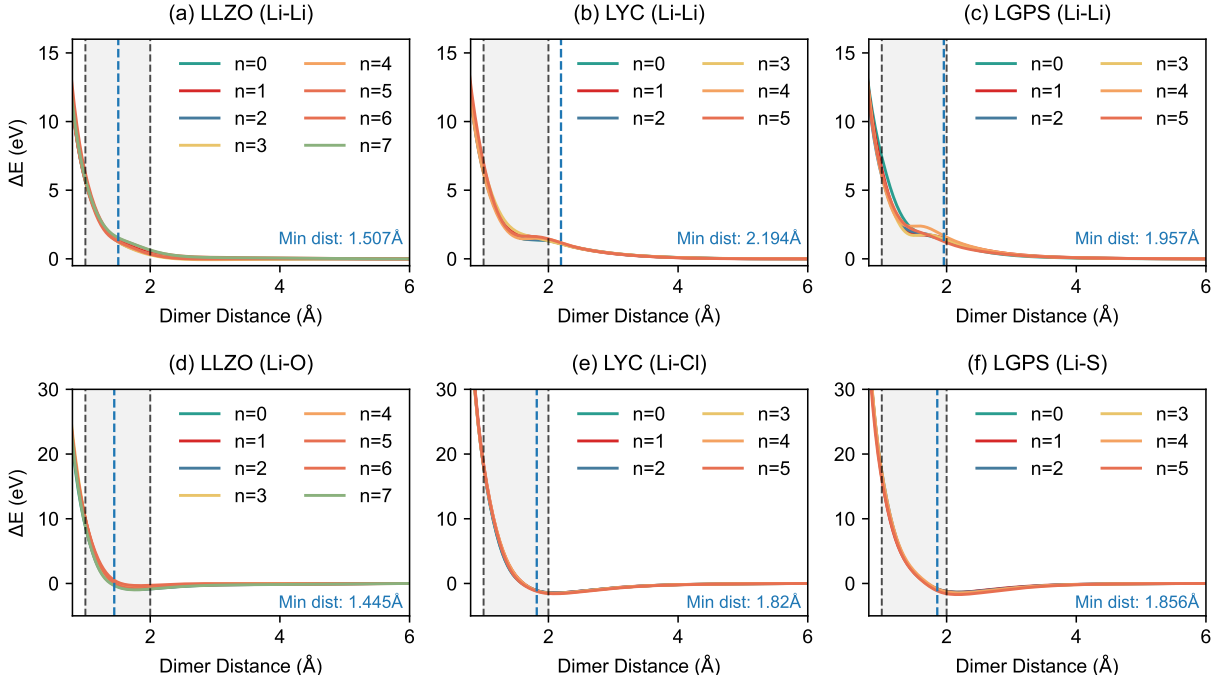


Figure S2: Dimer dissociation curves for (a) Li-Li in LLZO, (b) Li-Li in LYC, (c) Li-Li in LGPS, (d) Li-O in LLZO, (e) Li-Cl in LYC, and (f) Li-S in LGPS, computed from NEP models trained on datasets of different sizes. Here n denotes the dataset reduction level, where the training set contains $1/2^n$ of the full dataset. The shaded region marks the ZBL-NEP transition zone (inner cutoff: 1 Å; outer cutoff: 2 Å), and the blue dashed line indicates the minimum interatomic distance of the corresponding element pair in the full training set.

We performed dimer dissociation curve tests for all three SSE systems, comparing Li-Li and Li-anion pairs (Li-O, Li-Cl, Li-S for LLZO, LYC, and LGPS, respectively) across NEP models trained on progressively reduced datasets (Fig. S2). All NEP models employed here include the ZBL empirical repulsive potential, where the interaction smoothly transitions from the ZBL description to the NEP description between the inner (1 Å) and outer (2 Å) cutoffs. This coupling provides physically reasonable short-range repulsion in regions that are rarely sampled during training and would otherwise place the model in a complete extrapolation regime. As shown in Fig. S2, the dissociation curves remain highly consistent across all reduction levels within the NEP-dominated region. Minor deviations are observed for LGPS at short distances, but these fall within the extrapolation regime and are therefore expected. These results provide additional microscopic validation for the robustness of NEP models under dataset reduction.

Validation of MLFF Transport Predictions against AIMD

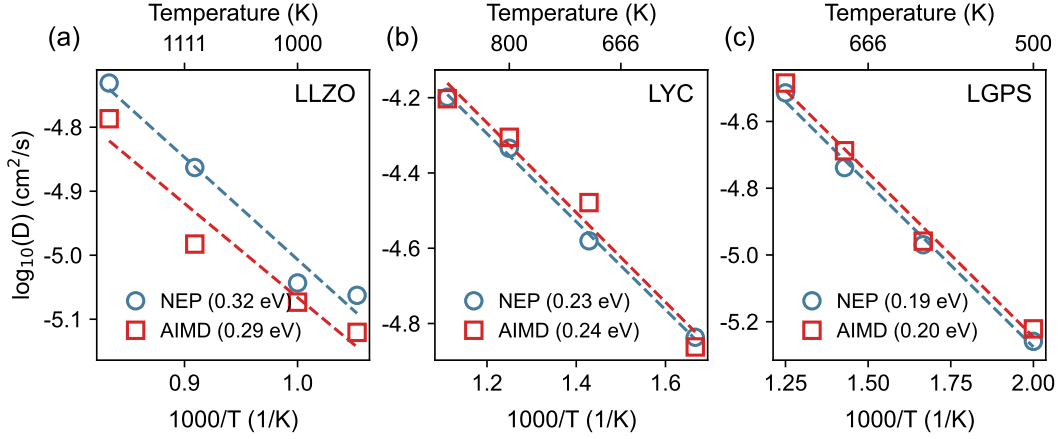


Figure S3: Arrhenius plots of Li-ion diffusivities for (a) LLZO, (b) LYC and (c) LGPS, calculated using the NEP models (blue dots) and AIMD simulations (red squares).

To validate the reliability of the trained force fields, we performed AIMD simulations for all three SSE materials and compared the resulting transport properties with those obtained from NEP-based MD simulations under identical conditions. Due to the finite-size effects present in these systems and the prohibitively high computational cost of large-scale AIMD, both AIMD and NEP simulations were performed using the same small supercells at elevated temperatures to ensure sufficient statistical convergence within accessible simulation timescales. As shown in Fig. S3, the NEP predictions show good agreement with the AIMD reference across all three materials, further confirming the reliability of the trained force fields for describing Li-ion transport in bulk SSEs.

Data efficiency of MACE model

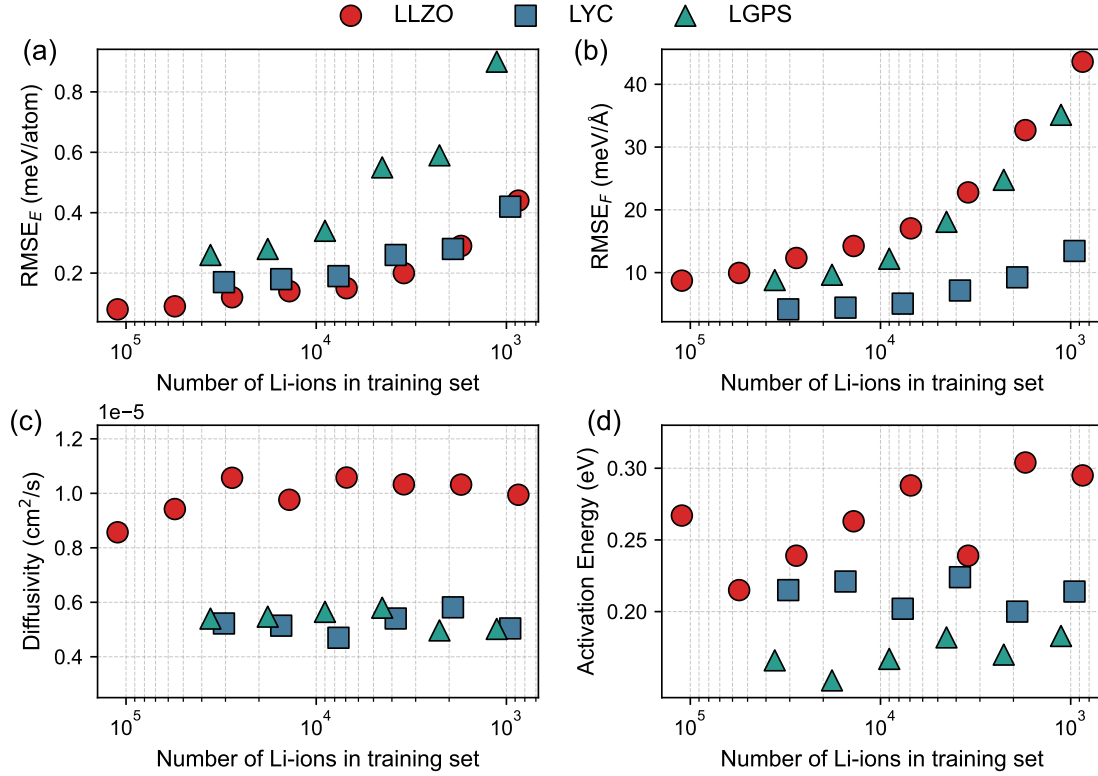


Figure S4: Dataset-size sensitivity tests on MACE-T1-E performance in LLZO, LYC, and LGPS. The validation root mean square errors of (a) energy and (b) force. (c) Li-ion diffusivities (1000 K for LLZO, 500 K for LYC and LGPS). (d) Li-ion activation energies.

To further validate the impact of dataset size across different model architectures, we conducted systematic dataset-size sensitivity tests using the MACE-T1-E model. The results, illustrated in Figure S4, show the evolution of energy and force root-mean-square errors (RMSE) relative to the full dataset as the training dataset is reduced.

Consistent with the findings for the NEP model (Figure 1 in the main text), MACE-T1-E maintains a high degree of accuracy even with reduced data. However, the increase in RMSE during the reduction process is more pronounced for MACE compared to the NEP framework. This difference originates from the underlying optimization strategies: the gradient-descent-based training of MACE, which benefits from larger datasets to effectively explore the complex parameter spaces. In contrast, the separable natural evolution strategy [14] employed by NEP appears to be more robust to small datasets, efficiently extracting core features from sparse configurations.

Despite the steeper rise in error, the absolute RMSE values of MACE-T1-E remain at a remarkably low level throughout the reduction. We have additionally evaluated the transport properties of MACE-T1-E models trained on datasets of varying sizes. As shown in Fig. S4, despite the pronounced sensitivity of force RMSE to dataset reduction, the predicted diffusivities remain largely consistent across all reduction levels, with only slightly larger variations in activation energy observed for LLZO. These results further support the conclusion that bulk SSE materials do not require large datasets to reliably capture ionic transport behavior.

Data efficiency of GNEP model

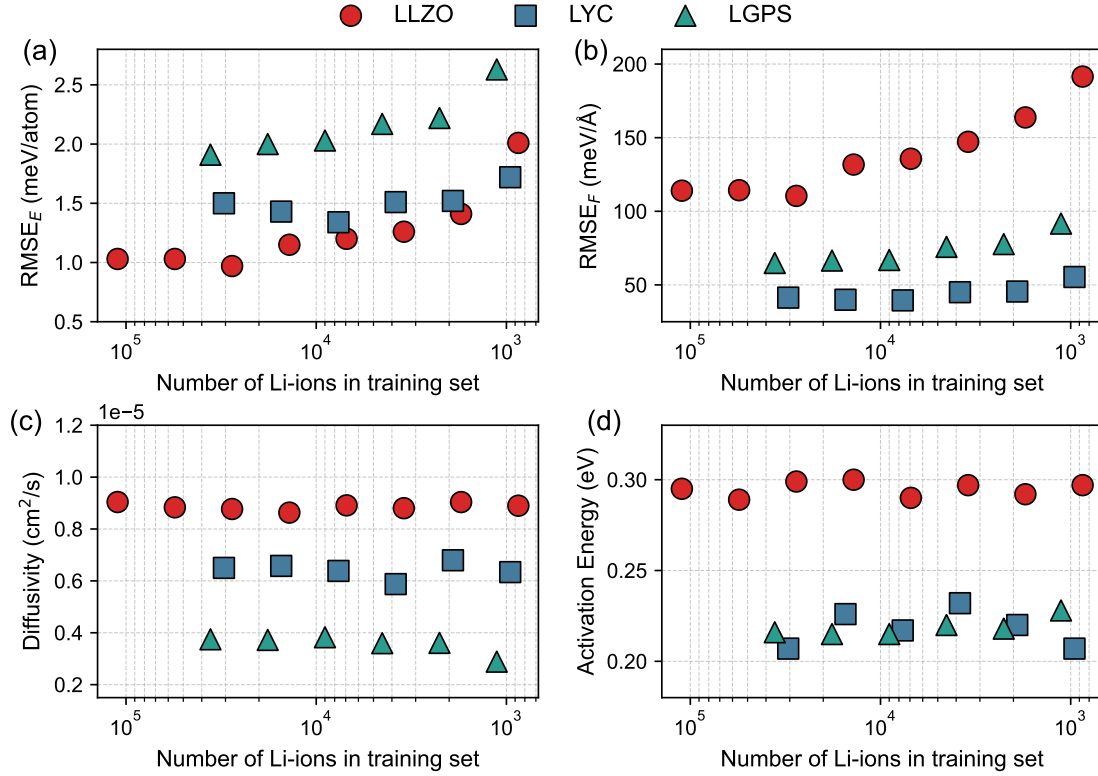


Figure S5: The root mean square errors of (a) energy and (b) force for full datasets calculated by GNEP models trained on different sizes of training datasets.

To further clarify whether the observed data-efficiency gap between NEP and MACE originates from the model architecture or the optimization algorithm, we performed additional benchmarks using the recently developed GNEP (Gradient-optimized Neuroevolution Potential) framework [15]. GNEP maintains the same descriptors and neural network structure as the standard NEP but replaces the natural evolution strategy with a gradient-descent-based optimizer (Adam).

As shown in Figure S5, unlike the standard NEP models, GNEP models exhibit a more pronounced increase in RMSE as the training set size decreases, showing a trend much closer to that of the MACE model. This observation may suggest that the data efficiency of the NEP is attributed to its separable natural evolution strategy [14], which appears more robust at extracting essential features from sparse data without getting trapped or requiring massive batches for gradient updates.

Data efficiency of fine-tuned NEP89 foundation model

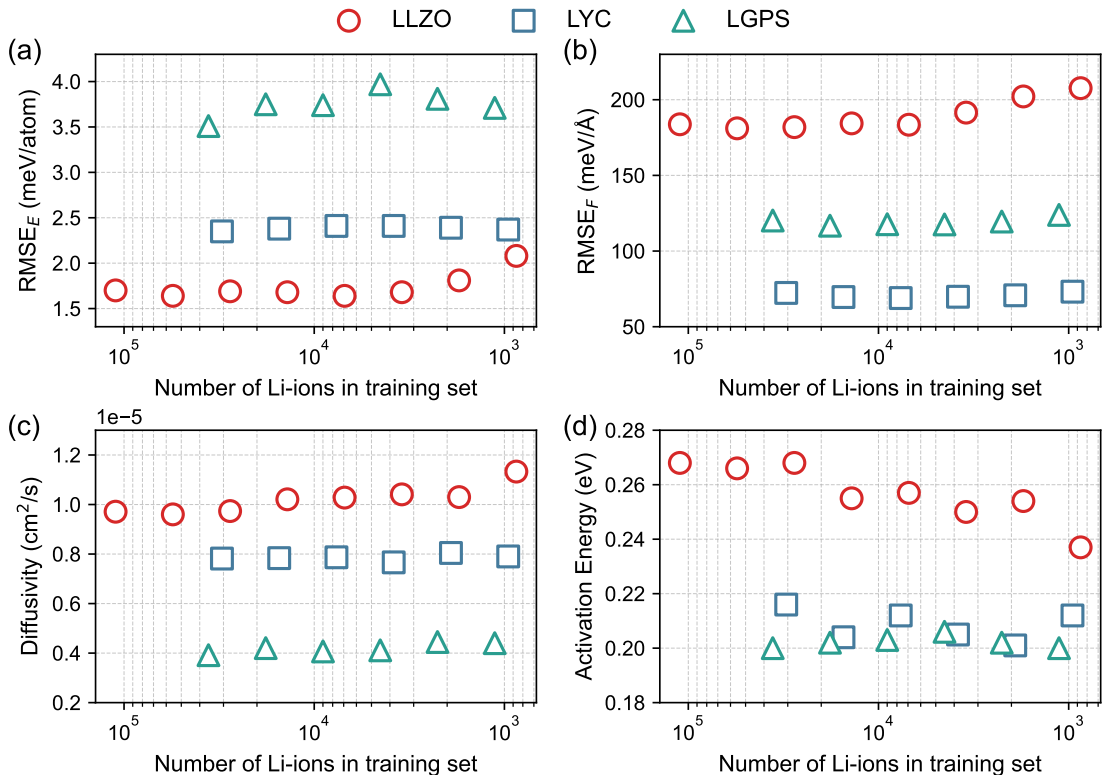


Figure S6: Dataset-size sensitivity tests on **fine-tuned NEP89** performance in $Li_7La_3Zr_2O_{12}$ (LLZO), Li_3YCl_6 (LYC), and $Li_{10}GeP_2S_{12}$ (LGPS). The validation root mean square errors of (a) energy and (b) force. (c) Li-ion diffusivities (1000 K for LLZO, 500 K for LYC and LGPS). (d) Li-ion activation energies.

We further asked whether the dataset-size sensitivity changes when starting from foundation models. To address this question, we performed additional tests using the NEP89 foundation model [16], which covers 89 elements, as a starting point for fine-tuning. Following the same dataset-size sensitivity protocol as in Fig. 1, we fine-tuned NEP89 for 10,000 generations on datasets of varying sizes. The results show that fine-tuned models exhibit a similar insensitivity to dataset size as models trained from scratch (Fig. S6), with diffusivity and activation energy predictions remaining consistent across all reduction levels. We note, however, that fine-tuned NEP89 models yield somewhat higher RMSEs than those trained from scratch, suggesting that the pretrained weights may impose some constraints on model expressivity for these specific systems.

Data efficiency of fine-tuned MACE foundation model

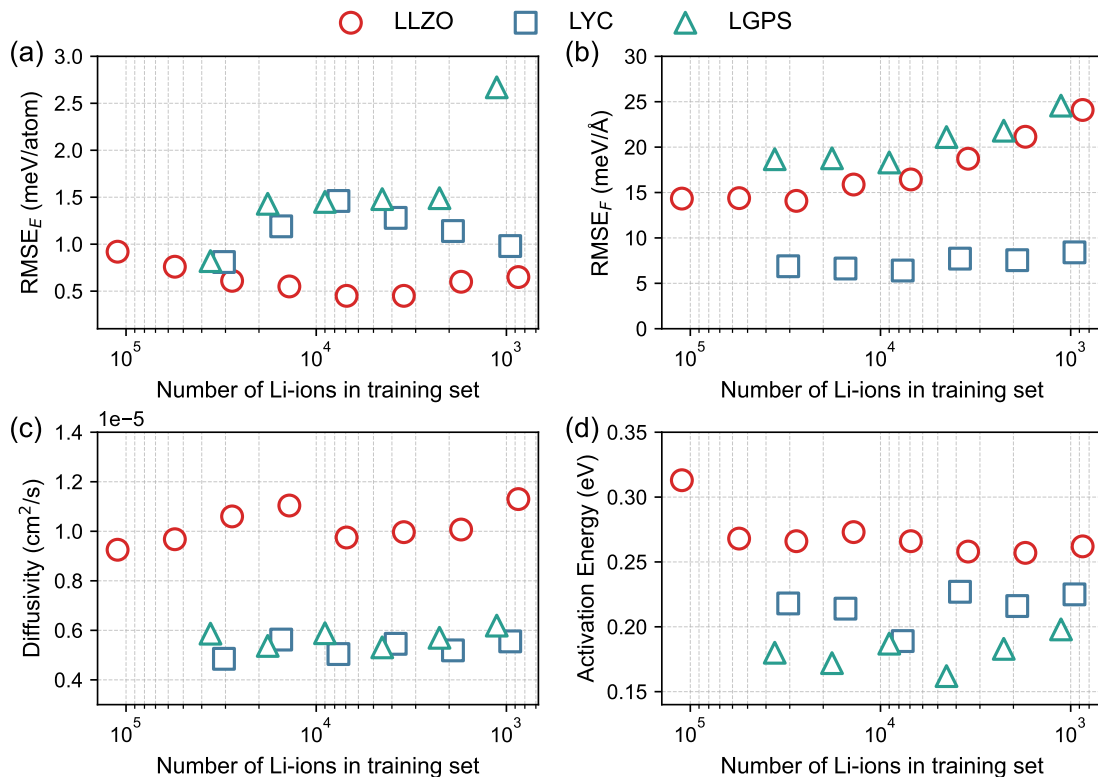


Figure S7: Dataset-size sensitivity tests on **fine-tuned MACE-OMAT-0-MEDIUM** performance in $Li_7La_3Zr_2O_{12}$ (LLZO), Li_3YCl_6 (LYC), and $Li_{10}GeP_2S_{12}$ (LGPS). The validation root mean square errors of (a) energy and (b) force. (c) Li-ion diffusivities (1000 K for LLZO, 500 K for LYC and LGPS). (d) Li-ion activation energies.

We also performed additional tests using the MACE-OMAT-0-MEDIUM foundation model with the LoRA approach (Fig. S7). Interestingly, the force RMSE of the fine-tuned models shows considerably less sensitivity to dataset size compared to models trained from scratch, with only a moderate increase as the training set is reduced. The energy RMSE exhibits somewhat larger fluctuations across reduction levels, which may partly reflect the sensitivity of the energy-force weighting and other hyperparameters in the LoRA fine-tuning setup. Nevertheless, despite these variations in RMSE, the predicted diffusivities and activation energies remain highly consistent across all reduction levels, with reliable transport predictions obtainable with as few as 15 configurations for LLZO. This further reinforces our conclusion that force RMSE is not a reliable metric for evaluating force field quality for ion transport applications, and demonstrates the data efficiency of bulk SSE systems when leveraging pretrained foundation models. These findings are consistent with recent work showing that as few as 10 additional training structures can yield significant accuracy improvements when fine-tuning MLIPs for phonon predictions [17].

Arrhenius plots calculated by NEP models trained with different random seeds

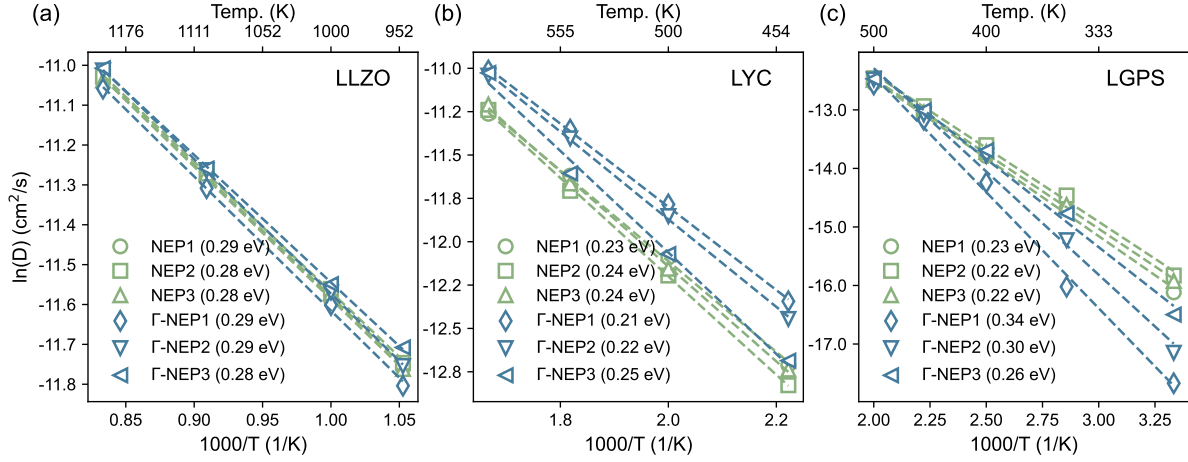


Figure S8: Arrhenius plots of Li-ion diffusion calculated by different NEP models trained with different random seeds in (a) LLZO, (b) LYC and (c) LGPS.

To evaluate the statistical robustness of our NEP models, we trained three independent replicas for both the NEP (trained on high-accuracy DFT data) and the Γ -NEP (trained on Γ -point DFT data) using different random seeds. The resulting Arrhenius plots and calculated activation energies (E_a) for LLZO, LYC, and LGPS are presented in Figure S8.

Our analysis reveals that for all three materials (LLZO, LYC, and LGPS), the three NEP replicas trained on high-accuracy data converge to nearly identical Arrhenius lines with minimal deviation in activation energies (e.g., 0.28 to 0.29 eV for LLZO), indicating that high-quality reference data provides a well-defined potential energy surface for consistent transport predictions. In stark contrast, models trained on low-quality Γ -point data exhibit noticeable discrepancies; while this variation is relatively small for LLZO and LYC, it becomes significantly more pronounced for LGPS, where the E_a values for Γ -NEP replicas fluctuate between 0.26 and 0.34 eV. These results confirm that inadequate reference data quality not only introduces systematic bias but also substantially increases model uncertainty, as the “noise” in low-precision DFT calculations creates an ambiguous training target that makes predicted transport behavior sensitive to random initialization and thus less predictable.

Arrhenius plots of LGPS calculated by MACE models trained with different random seeds

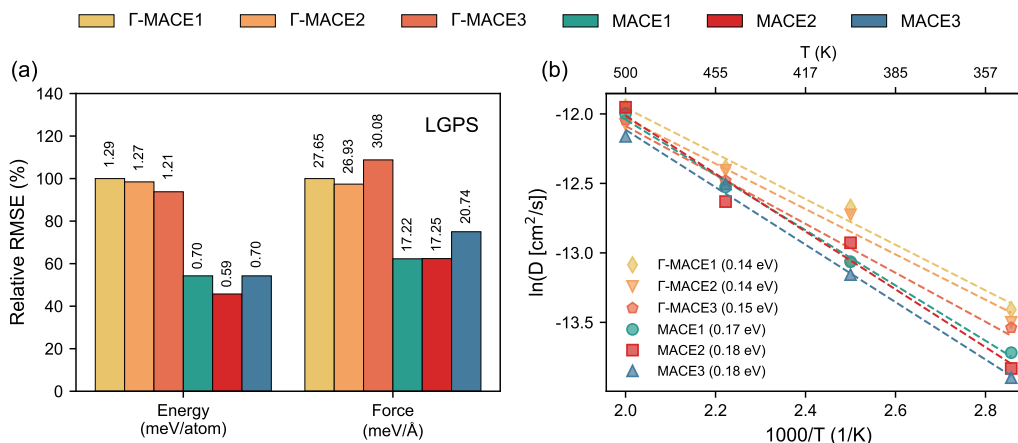


Figure S9: Comparison of (a) relative RMSE of energies and forces and (b) Arrhenius plots of Li-ion diffusivities for LGPS, calculated from three independent MACE and Γ -MACE models trained on high-accuracy and Γ -point DFT reference data, respectively.

We performed additional tests on LGPS using the same dataset as used for the MLFF architecture benchmark, training three independent replicas each of MACE and Γ -MACE models with different random seeds. As shown in Fig. S9, the Γ -MACE models exhibit good reproducibility across replicas, in contrast to the large variance observed for Γ -NEP. However, Γ -MACE slightly underestimates the activation energy (0.14–0.15 eV) compared to MACE trained on high-accuracy data (0.17–0.18 eV). This suggests that the sensitivity to reference data quality is somewhat framework-dependent: while MACE is less susceptible to the instability and variance induced by low-quality data, it is not immune to systematic bias. These results therefore reinforce rather than undermine our recommendation to prioritize data quality, as the detrimental effect of Γ -point reference data on transport predictions is observed across both frameworks, albeit to different degrees.

Data quality sensitivity of fine-tuned foundation models

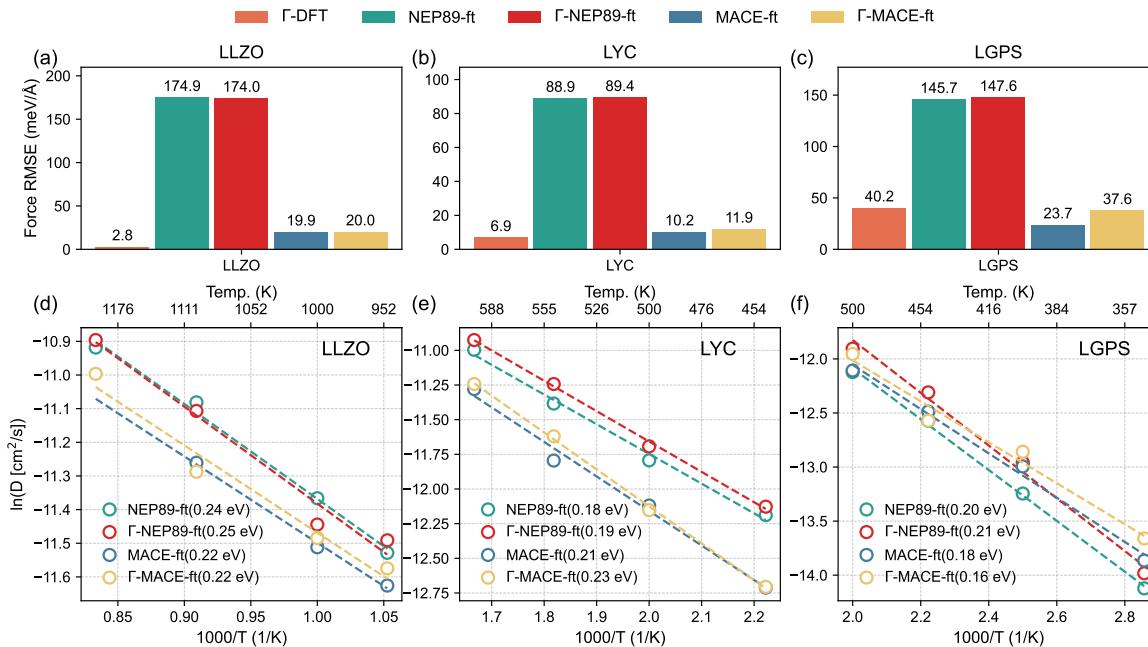


Figure S10: Data quality sensitivity tests for fine-tuned models in LLZO, LYC, and LGPS. (a–c) Force RMSEs of NEP89-ft, Γ -NEP89-ft, MACE-ft, and Γ -MACE-ft models, alongside the Γ -DFT reference error. (d–f) Arrhenius plots of Li-ion diffusivities calculated from the corresponding models.

To examine how reference data quality affects fine-tuning, we trained NEP89 and MACE-OMAT-0-MEDIUM models on both high-accuracy and Γ -point DFT reference data, following the same protocol used in the reference-quality tests. This yields four fine-tuned models: NEP89-ft, Γ -NEP89-ft, MACE-ft, and Γ -MACE-ft. As shown in Fig. S10, the fine-tuned NEP89 models show relatively large force RMSEs regardless of data quality, while the MACE-ft models achieve force RMSEs closer to the Γ -DFT reference error level. Despite these RMSE differences, the predicted diffusivities and activation energies are largely similar between models fine-tuned on high-accuracy and Γ -point data across all three materials. These preliminary results suggest that fine-tuning may be less sensitive to reference data quality than training from scratch, although more systematic tests are required before drawing a general conclusion.

Large-supercell Γ -point validation for LGPS

To test whether the poor LGPS performance observed with small-cell Γ -point reference data can be resolved by using a larger supercell, we performed an additional validation using a $2 \times 2 \times 2$ LGPS supercell containing 400 atoms. We first carried out 2 ns MD simulations over the 300–800 K temperature range and selected 100 representative configurations using the farthest-point sampling algorithm implemented in GPUMDkit [4]. These configurations were then recomputed using Γ -point DFT and used to train three independent NEP models with different random seeds. As shown in Fig. S11, the three independently trained models give highly consistent Arrhenius behavior, indicating that the large-supercell Γ -point dataset leads to reproducible transport predictions for LGPS.

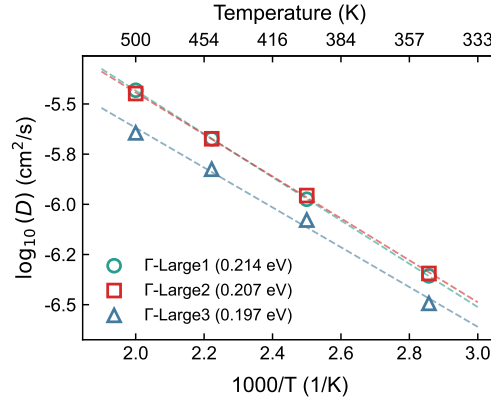


Figure S11: Arrhenius plots of Li-ion diffusivities in LGPS calculated from three independent NEP models trained on Γ -point DFT data generated using a 400-atom $2 \times 2 \times 2$ LGPS supercell.

To further evaluate the accuracy of the large-supercell Γ -point reference data, we randomly selected 20 configurations from the 100-configuration dataset and recomputed them using a higher-accuracy DFT setting with $KSPACING = 0.25 \text{ \AA}^{-1}$. Figure S12 compares the energies, forces, and virials from large-supercell Γ -point DFT against the high-accuracy DFT reference. The corresponding RMSEs are 0.15 meV/atom for energy, 0.75 meV/ $\text{\AA}$ for force, and 0.166 eV for virial. These errors are substantially smaller than the small-cell LGPS Γ -point error and are far below typical MLFF fitting errors, confirming that using a sufficiently large LGPS supercell can greatly suppress Γ -point DFT noise.

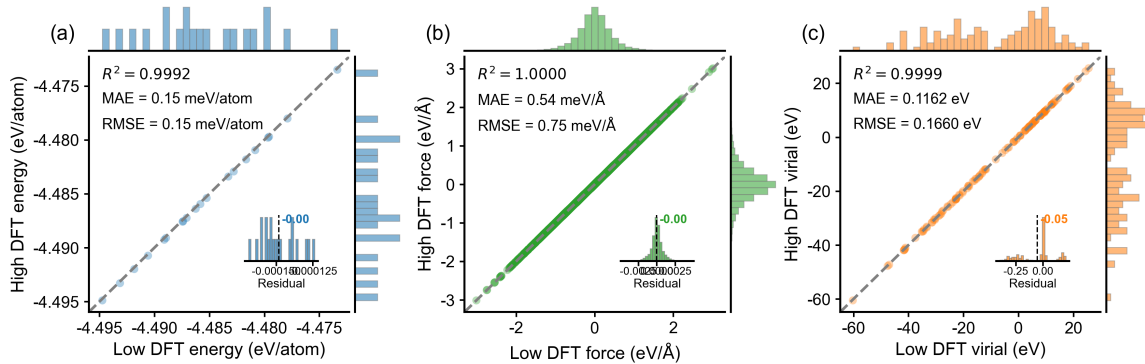


Figure S12: Parity comparison between large-supercell Γ -point DFT and high-accuracy DFT calculations for 20 randomly selected 400-atom LGPS configurations.

Structural and vibrational validation of MLFF models

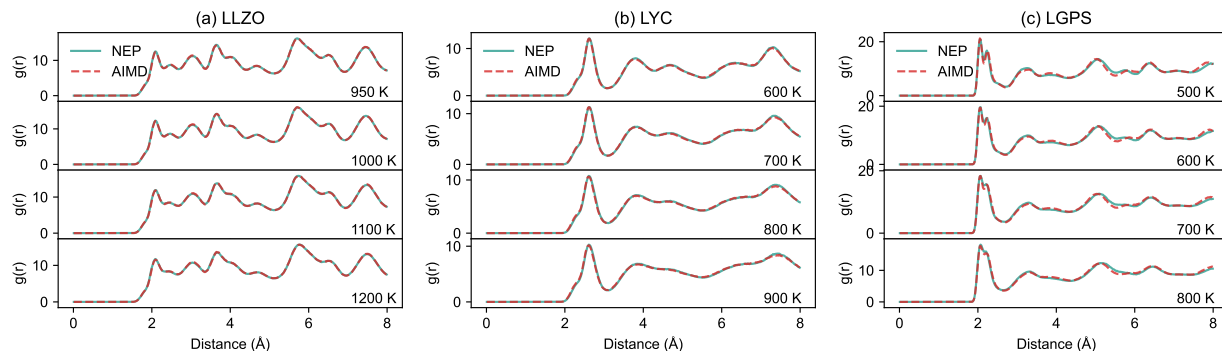


Figure S13: Radial distribution functions of LLZO, LYC, and LGPS calculated from NEP and AIMD simulations at four temperatures.

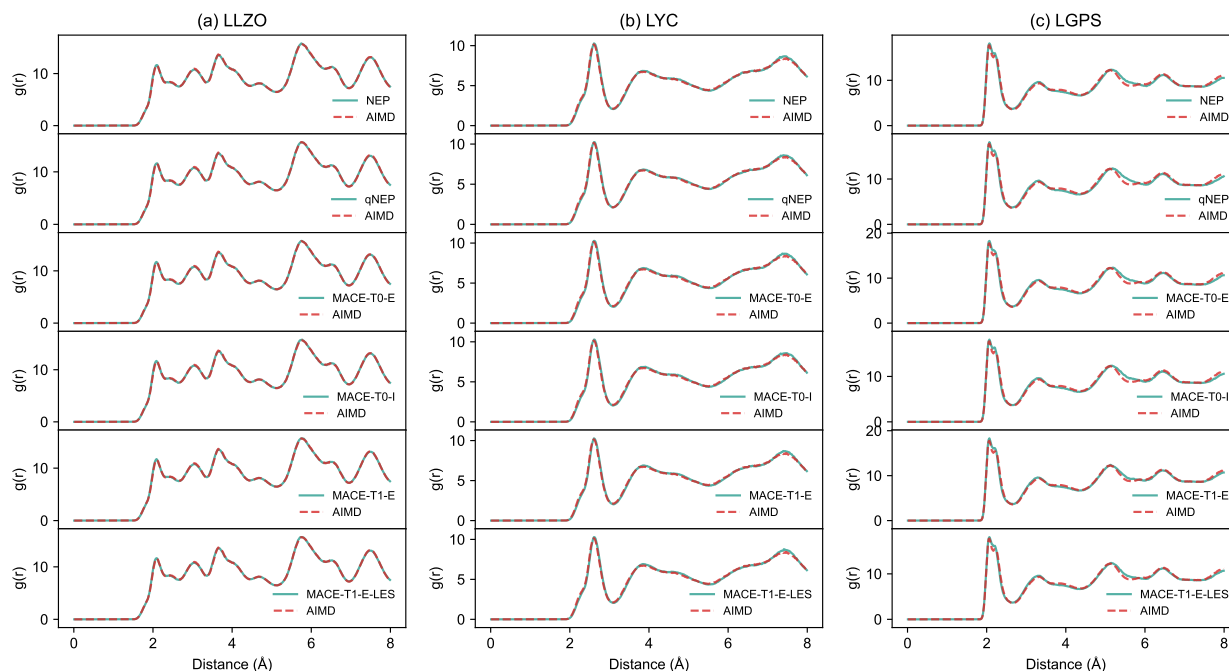


Figure S14: Radial distribution functions of LLZO, LYC, and LGPS calculated from six different MLFF models and AIMD simulations at representative temperatures (1200 K for LLZO, 900 K for LYC, and 800 K for LGPS).

To further validate the physical reliability of the trained force fields beyond diffusivity and activation energy, we calculated radial distribution functions (RDFs) for all three materials. First, the baseline NEP model was compared with AIMD at four temperatures (Fig. S13). The RDFs show excellent agreement across all systems, with only minor deviations observed for LGPS at larger interatomic distances ($\sim 6-8$ Å). We then compared all six MLFF models with AIMD at representative temperatures (Fig. S14). The RDFs remain highly consistent across different architectures, indicating that the models preserve the local structural environments relevant to ion transport.

We also computed the vibrational density of states (VDOS) for the six MLFF models at 1000 K for LLZO and 500 K for LYC and LGPS (Fig. S15). The VDOS profiles are consistent across models, suggesting that the different architectures capture similar vibrational features of the underlying potential energy surfaces.

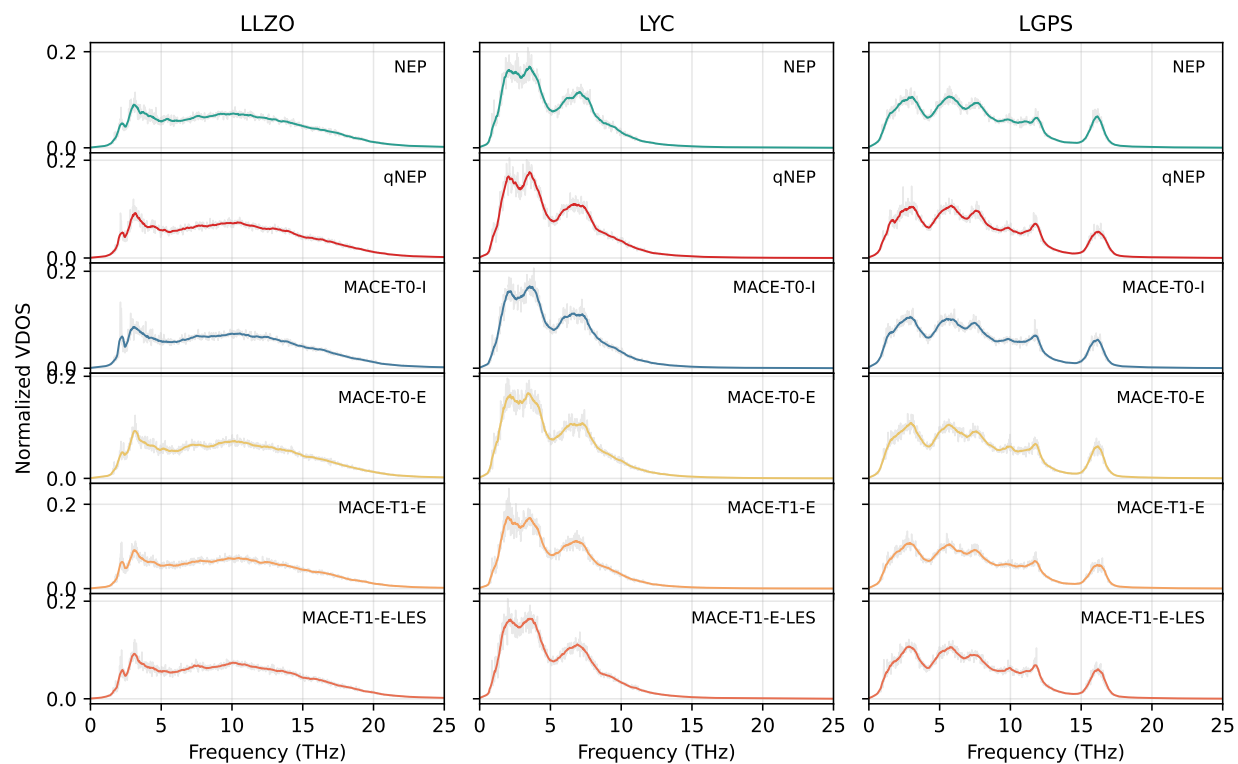


Figure S15: Vibrational density of states (VDOS) of (a) LLZO at 1000 K, (b) LYC at 500 K, and (c) LGPS at 500 K, calculated from different MLFF models.

Speed benchmark of NEP models with larger cutoff

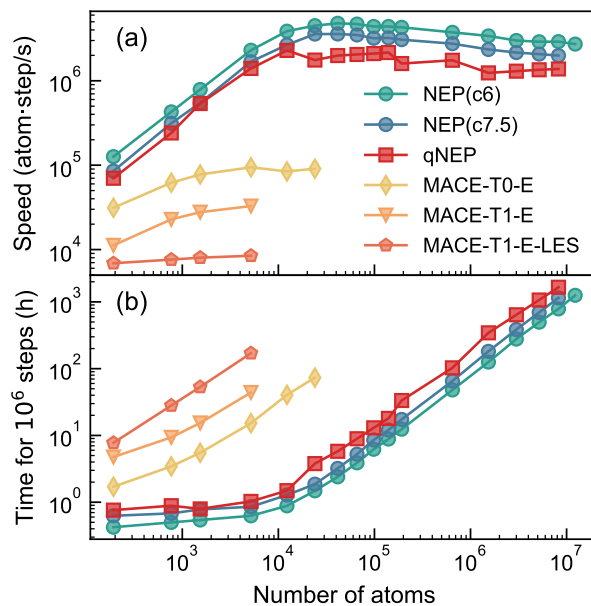


Figure S16: Scaling of (a) simulation speed and (b) total wall-clock time required for 10^6 MD steps as a function of the number of atoms for NEP models with $R_{\text{cut}} = 6$ and 7.5 Å. All tests were performed on the LLZO system using a single NVIDIA V100 GPU (32 GB).

We additionally benchmarked the computational efficiency of the NEP model using a larger cutoff radius of $R_{\text{cut}} = 7.5$ Å for LLZO. As shown in Fig. S16, increasing the cutoff from 6 to 7.5 Å slows the NEP simulation by approximately 1.5 – $1.8\times$, but the model remains substantially faster than message-passing MACE variants across the tested system sizes.

References

- [1] Zihan Yan and Yizhou Zhu. Impact of lithium nonstoichiometry on ionic diffusion in tetragonal garnet-type $\text{Li}_7\text{La}_3\text{Zr}_2\text{O}_{12}$. *Chem. Mater.*, 36(23):11551–11557, 2024. doi:10.1021/acs.chemmater.4c02454.
- [2] Shuo Wang, Yunsheng Liu, and Yifei Mo. Frustration in super-ionic conductors unraveled by the density of atomistic states. *Angew. Chem. Int. Ed.*, 62(15):e202215544, 2023. doi:10.1002/anie.202215544.
- [3] Jianxing Huang, Linfeng Zhang, Han Wang, Jinbao Zhao, Jun Cheng, and Weinan E. Deep potential generation scheme and simulation protocol for the $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ -type superionic conductors. *J. Chem. Phys.*, 154(9):094703, 2021. doi:10.1063/5.0041849.
- [4] Zihan Yan, Denan Li, Xin Wu, Zhoulin Liu, Chen Hua, Boyi Situ, Hao Yang, Shengjie Tang, Benrui Tang, Ziyang Wang, et al. GPUMDkit: A user-friendly toolkit for GPUMD and NEP. *MGE Advances*, 4:e70074, 2026. doi:10.1002/mgea.70074.
- [5] K. Xu, H. Bu, S. Pan, E. Lindgren, Y. Wu, Y. Wang, J. Liu, K. Song, B. Xu, Y. Li, T. Hainer, L. Svensson, J. Wiktor, R. Zhao, H. Huang, C. Qian, S. Zhang, Z. Zeng, B. Zhang, B. Tang, Y. Xiao, Z. Yan, J. Shi, Z. Liang, J. Wang, T. Liang, S. Cao, Y. Wang, P. Ying, N. Xu, C. Chen, Y. Zhang, Z. Chen, X. Wu, W. Jiang, E. Berger, Y. Li, S. Chen, A. J. Gabourie, H. Dong, S. Xiong, N. Wei, Y. Chen, J. Xu, F. Ding, Z. Sun, T. Ala-Nissila, A. Harju, J. Zheng, P. Guan, P. Erhart, J. Sun, W. Ouyang, Y. Su, and Z. Fan. GPUMD 4.0: A high-performance molecular dynamics package for versatile materials simulations with machine-learned potentials. *MGE Advances*, 3:e70028, 2025. doi:doi:10.1002/mgea.70028.
- [6] Georg Kresse and Jürgen Hafner. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B*, 49(20):14251, 1994. doi:10.1103/PhysRevB.49.14251.
- [7] Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54(16):11169, 1996. doi:10.1103/PhysRevB.54.11169.
- [8] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864, 1964. doi:10.1103/PhysRev.136.B864.
- [9] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133, 1965. doi:10.1103/PhysRev.140.A1133.
- [10] Peter E Blöchl. Projector augmented-wave method. *Phys. Rev. B*, 50(24):17953, 1994. doi:10.1103/PhysRevB.50.17953.
- [11] Bingqing Cheng. Latent ewald summation for machine learning of long-range interactions. *npj Comput. Mater.*, 11(1):80, 2025. doi:10.1038/s41524-025-01577-7.
- [12] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.*, 117(1):1–19, 1995. doi:10.1006/jcph.1995.1039.
- [13] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment—a python library for working with atoms. *J. Phys. Condens. Mat.*, 29(27):273002, 2017. doi:10.1088/1361-648X/aa680e.

- [14] Tom Schaul, Tobias Glasmachers, and Jürgen Schmidhuber. High dimensions and heavy tails for natural evolution strategies. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, page 845–852, 2011. doi:[10.1145/2001576.2001692](https://doi.org/10.1145/2001576.2001692).
- [15] Hongfu Huang, Junhao Peng, Kaiqi Li, Jian Zhou, and Zhimei Sun. Efficient gpu-accelerated training of a neuroevolution potential with analytical gradients. *Comput. Phys. Commun.*, 320:109994, 2026. doi:[10.1016/j.cpc.2025.109994](https://doi.org/10.1016/j.cpc.2025.109994).
- [16] Ting Liang, Ke Xu, Eric Lindgren, Zherui Chen, Rui Zhao, Jiahui Liu, Esmée Berger, Benrui Tang, Bohan Zhang, Yanzhou Wang, et al. NEP89: Universal neuroevolution potential for inorganic and organic materials across 89 elements. *arXiv*, 2025. doi:[10.48550/arXiv.2504.21286](https://doi.org/10.48550/arXiv.2504.21286).
- [17] Jonas Grandel, Philipp Benner, and Janine George. Parameter-efficient fine-tuning of machine-learning interatomic potentials for phonon and thermal properties. *arXiv*, 2026. doi:[10.48550/arXiv.2604.0101](https://doi.org/10.48550/arXiv.2604.0101).