Assessing the Accuracy of Machine Learning Interatomic Potentials in Predicting the Elemental Orderings: A Case Study of Li-Al Alloys

Yunsheng Liu[1], Yifei Mo[1,*]

[1] Department of Materials Science and Engineering, University of Maryland, College Park, MD, USA

* Email: yfmo@umd.edu

Abstract. In atomistic modeling, machine learning interatomic potential (MLIP) has emerged as a powerful technique for studying alloy materials. However, given that MLIPs are often trained on a limited set of materials, a concern remains regarding the MLIP's capability to make accurate predictions for a wide variety of phases, compositions, lattice structures, and elemental orderings across alloy systems. This paper presents a detailed analysis of MLIP's performance in the Li-Al alloy system. Even trained only on a very limited number of phases, the MLIPs exhibit good accuracies in predicting a vast array of known and generated intermediate phases and their elemental orderings across the alloy system. We propose and demonstrate several evaluation metrics to assess and quantify the relative stabilities of complex elemental orderings, which is critical for studying the thermodynamics of alloys. Our testing process combined with the evaluation metrics is valuable for quantifying the performance and the transferability of MLIPs and for future improvements of MLIPs.

Keywords: machine learning interatomic potential, lithium aluminum alloys, phase diagram, elemental ordering, Monte Carlo simulations

1. Introduction

As a powerful technique to study the materials phenomena and properties, atomistic modeling and simulations of materials are conducted based on the interactions among atoms, known as potential energy surfaces. While density functional theory (DFT) calculations have been widely used, their high computation cost limits their applications to atomistic models with a small number of atoms (on the level of up to a few hundred) and small system sizes (on the level of ~10^1 Å). The atomistic modeling and simulations of many materials phenomena and properties require the sampling of a large number of atomistic configurations in models with much larger length scale. Thus, alternative techniques to evaluate the potential energy surfaces of atomistic systems with lower computational costs are required. The interatomic potentials, such as modified embedded atom method (MEAM) potentials, are also commonly adopted to calculate and simulate a wide range of materials phenomena and properties in metal and alloy systems using molecular dynamics (MD) simulations.[1–3] Monte Carlo (MC) simulations based on cluster expansion method have been widely employed to evaluate the thermodynamic properties and compute the phase diagrams of alloy systems.[4,5] Machine-learning interatomic potential (MLIP) utilizes ML techniques to reproduce the potential energy surfaces of atomistic systems by training on a variety of configurations and their DFT calculated energies. MLIPs boast multiple advantages, including low computation cost, linear scalability to system sizes, and claimed DFT-level accuracy.[6] MLIPs have been adopted to study many materials phenomena, e.g., fast ionic conduction in ceramic materials,[7–10] phase transitions in amorphous materials,[11] and the phase stabilities and orderings of alloys.[12–16]

A challenge of developing MLIP for alloys is the existence of many phases in different lattice structures over a range of compositions, and each phase can exhibit a wide variety of elemental orderings (or atomistic configurations). To correctly predict the thermodynamics of an alloy system, it is essential for the MLIPs to accurately reproduce the energies and relative stabilities of all these element orderings for stable and unstable phases over a wide range of compositions in different lattice structures. For example, the prediction of the lowest energy configurations for all relevant phases with different structures and compositions is essential for constructing the convex hull and the phase diagrams.[13,14,16–19] Systematic studies based on the cluster expansion method have been conducted by Nguyen et al.[20] to address such challenges for modeling alloys. In addition, MD/MC simulations of many physical phenomena, including anti-site defects, diffusion, element dissolution, solid solution, and disordering, also rely on the accurate predictions of the energies and relative stabilities of different elemental orderings. Therefore, comprehensive examinations are still needed on the MLIPs' capability to accurately predict alloy systems with a large number of complex orderings over many compositions and lattice structures.

In addition, given that training data of the MLIPs are often limited to a few pre-selected phases, compositions, and lattice structures, a common question is whether MLIPs can make accurate predictions on a wide variety of atomistic configurations and elemental orderings from other compositions and lattice structures that are outside of the training data but may occur during the actual MLIPs applications for MD/MC simulations. The capability of interatomic potentials to predict the atomistic configurations outside the training process is known as transferability.[6,11,21–23] Besides choosing a variety of structures and compositions in the training dataset, an alternative approach is using active learning, as demonstrated by Gubaev et al.,[16] to iteratively add new structures into the training dataset and re-train MLIPs, which may greatly mitigate the issue caused by having a limited number of structures in pre-selected training data. Nevertheless, previous studies revealed a number of discrepancies in MLIPs predicting atomic dynamics and materials properties.[6,21,23] Thus, it is critical to test how MLIPs would perform on many possible atomistic configurations that may be encountered in their atomistic simulations of alloys, given that the training dataset cannot cover all of them.

In this study, we perform a systematic test of MLIP performances on complex elemental orderings of different phases over the alloy system. We conduct the MLIP tests on the Li-Al alloy system, which includes a variety of phases with body-centered cubic, face-centered cubic, hexagonal close-packed, and other lattices. We systematically examine the MLIP performance on a large number of elemental orderings in many phases and lattice structures in and out of the training data. We find that MLIPs trained on only a small number of phases may achieve good accuracy and transferability to the orderings in many different phases with low energy errors (section 2.1). In addition, we develop new quantitative metrics for the evaluation of the accuracy of energy rankings of elemental orderings and show that the MLIPs reproduce the energy ranking for other phases not included in the training data (section 2.2). Moreover, we also study the effect of training data, and find that increasing the diversity of training data without significantly increasing its size may lead to worse performance (section 2.3). Finally (section 2.4), the MLIPs are tested in MC and MD simulation for their common application cases. We find discrepancies caused by large supercell sizes and good performance on the energies of

intermediate phases and forces of migrating diffusion atoms. The implications of our results for future improvements of MLIPs are discussed.

2. Results

2.1 Evaluating the MLIP across the alloy phase diagram

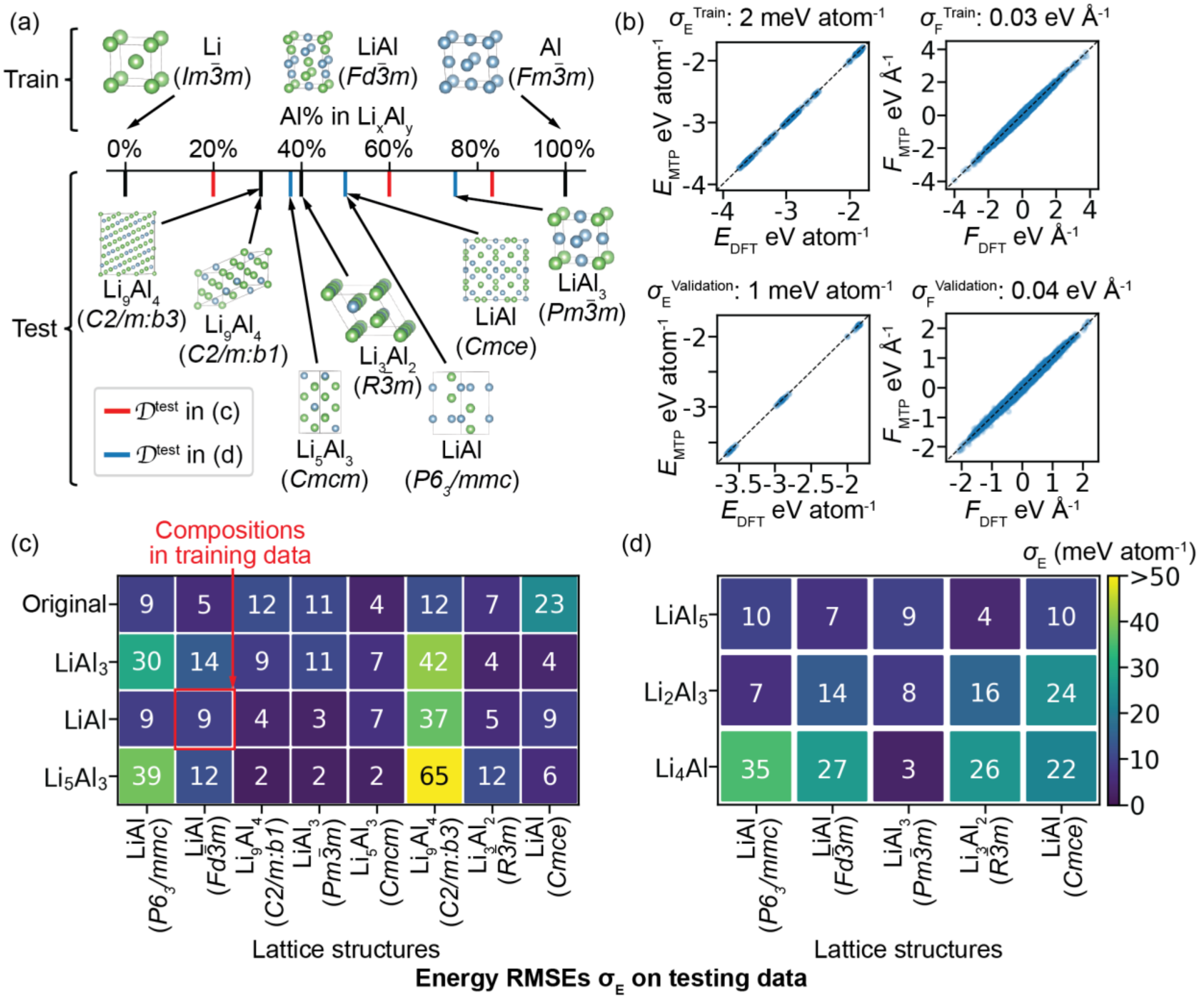

Figure 1. The training and testing of the MLIP in Li-Al alloy system. (a) Among 10 known Li_xAl_y phases, BCC Li ($Im\bar{3}m$), BCC LiAl ($Fd\bar{3}m$), and FCC Al ($Fm\bar{3}m$) are used in the training dataset $\mathcal{D}^{\mathrm{train}}$, and the other seven intermediate phases are used in the testing dataset $\mathcal{D}^{\mathrm{test}}$. (b) Comparison of the energies E and atomic forces F predicted by MTP versus the DFT K4 benchmark on training and validation datasets, $\mathcal{D}^{\mathrm{train}}$ and $\mathcal{D}^{\mathrm{validation}}$. The RMSEs of energies for training and validation data, $\sigma_{\mathrm{E}}^{\mathrm{Train}}$ and $\sigma_{\mathrm{E}}^{\mathrm{Validation}}$, and the RMSEs of forces for training and validation data, $\sigma_{\mathrm{F}}^{\mathrm{Train}}$ and $\sigma_{\mathrm{F}}^{\mathrm{Validation}}$, are given above each plot. (c, d) The RMSEs of predicted energies of up to 30 elemental orderings for phases in given lattice structures (x-axis) generated with different compositions (y-axis).

Our study is performed using the Moment Tensor Potential (MTP) model on the Li-Al alloy system. This binary alloy system includes body-centered cubic (BCC) Li (*Im*$\bar{3}$*m*), face-centered cubic (FCC) Al (*Fm*$\bar{3}$*m*), and eight known intermediate Li_xAl_y phases (**Figure 1**a), such as BCC structures Li_9Al_4 (*C2/m:b1*), Li_9Al_4 (*C2/m:b3*), Li_5Al_3 (*Cmcm*), Li_3Al_2 (*R*$\bar{3}$*m*), and LiAl (*Fd*$\bar{3}$*m*), FCC structure $LiAl_3$ (*Pm*$\bar{3}$*m*), hexagonal close packed (HCP) structure LiAl (*P*6_3*/mmc*), and other structure LiAl (*Cmce*). To train the MTP model for the Li-Al binary alloy system, the training dataset includes the end phases BCC Li (*Im*$\bar{3}$*m*) and FCC Al (*Fm*$\bar{3}$*m*) (Methods), and only one intermediate phase BCC LiAl (*Fd*$\bar{3}$*m*). The training dataset $\mathcal{D}^{\text{train}}$ is generated from these three phases, including a variety of atomic configurations from crystalline bulk structures, strained and distorted bulk structures, liquid structures, defected structures with multiple vacancies or interstitials, interfaces between each pair of the three phases with the {001}, {011} and {111} surfaces, and the snapshots obtained from AIMD simulations (Methods) of bulk crystalline phases, bulk phases with point defects, and interface supercells. The trained MTP model, referred as $\text{MTP}_{\text{Train}}$, accurately predicts atomic forces and energies on $\mathcal{D}^{\text{train}}$ and the validation dataset $\mathcal{D}^{\text{validation}}$ (consisting of 66 snapshots from AIMD simulations, Methods) (Figure 1b). Compared with DFT calculations using a **k**-point mesh of 4×4×4 (DFT K4), the root-mean-squared errors (RMSEs) of energies are as low as 3 meV atom^{-1} for $\mathcal{D}^{\text{train}}$ and 2 meV atom^{-1} for $\mathcal{D}^{\text{validation}}$ and the RMSEs of atomic forces are as low as 0.03 eV Å^{-1} for $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{validation}}$ (Figure 1b).

Here, we test whether the $\text{MTP}_{\text{Train}}$ trained on only three phases (i.e., two end phases and one intermediate phase) can properly predict other intermediate phases, including different orderings in different lattice structures over the composition range in the Li-Al alloy system. The testing dataset, $\mathcal{D}^{\text{test}}$, was constructed based on all intermediate phases, including other known intermediate phases, i.e., four BCCs Li_9Al_4 (*C2/m:b1*), Li_9Al_4 (*C2/m:b3*), Li_5Al_3 (*Cmcm*), and Li_3Al_2 ($R\bar{3}m$), FCC $LiAl_3$ ($Pm\bar{3}m$), HCP LiAl (*$P6_3$/mmc*), and the LiAl (*Cmce*) (Figure 1c), and also the hypothetical phases generated in a wide range of Li-Al compositions in the lattice structures of known intermediate phases (Methods). These hypothetical phases cover many compositions, either the same as known intermediate phases (blue lines, Figure 1a, and lattice structures, Figure 1c) or those deviate significantly from known intermediate phases (red lines, Figure 1a, and lattice structures, Figure 1d). For each phase, up to 30 configurations are generated by swapping the positions of a number of randomly selected atoms in known intermediate phases and by randomly sampling the elemental orderings through the partially substituted sites in the hypothetical phases (Methods) (Supplementary Figure S19). The $\mathcal{D}^{\text{test}}$ allows the test of MLIPs covering a wide range of lattice structures, compositions, and orderings with highly different atomic environments, which are distinct from the training data. It is important to test the MLIPs' capabilities of correctly modeling the relative stabilities of all these phases and orderings, which are crucial for reproducing the most stable phases, the phase transitions, and the phase diagrams.

This $\text{MTP}_{\text{Train}}$ trained on only three phases shows low energy RMSEs of 2 – 14 meV atom^{-1} for most (35 out of 47, Figure 1c and d) of the other intermediate Li-Al phases and the hypothetical generated phases not in the training dataset. The HCP LiAl (*$P6_3$/mmc*) and the LiAl (*Cmce*), which are neither BCC nor FCC covered in training, show energy RMSEs greater than 20 meV atom^{-1} (the 1st row, Figure 1c). The $\text{MTP}_{\text{Train}}$ shows a larger error of energy RMSEs higher than 30 meV atom^{-1} for five hypothetical phases, HCP Li_5Al_3 (*P6₃/mmc*), HCP $LiAl_3$ (*$P6_3/mmc$*), HCP Li_4Al (*$P6_3/mmc$*), BCC Li_5Al_3 (*C2/m:b3*), BCC LiAl (*C2/m:b3*), and BCC $LiAl_3$ (*C2/m:b3*) (Figure 1c and d). The larger errors in HCP structures may be caused by the identical nearest-neighbor atomic environments between HCP and FCC, which may be more challenging to be distinguished by the atomistic descriptor of the MLIP. Similarly, the larger errors on *C2/m:b3* BCC may also be explained by the confusion with another BCC *C2/m:b1* (Figure 1c). Additionally, we test the $\mathrm{MTP_{Train}}$ in other types of structures, such as the structures with a single Li or Al vacancy in the supercells, and the energy RMSEs are as low as 19 meV $\mathrm{atom^{-1}}$ for most intermediate phases (Supplementary Figure S1), which implies the performances of MLIPs can be transferred to the structures and orderings with vacancy defects outside the training data (Methods). However, we also find that the errors of MLIP predictions can be large for the individual defects, such as the defect formation energies (see ***errors of MLIPs on dataset with single vacancies*** in Supporting Information), which may be topic for in future MLIPs studies. In summary, the tests in general show good MLIP performance for predicted energies on many known and hypothetical alloy phases over a wide range of compositions with different lattice structures and orderings. Given that the $\mathrm{MTP_{Train}}$ is trained using only two end phases and one intermediate phase, this implies the good performance and transferability of MLIPs for the energies of the lattice-based model for alloy systems.

2.2 Energy rankings of different elemental orderings

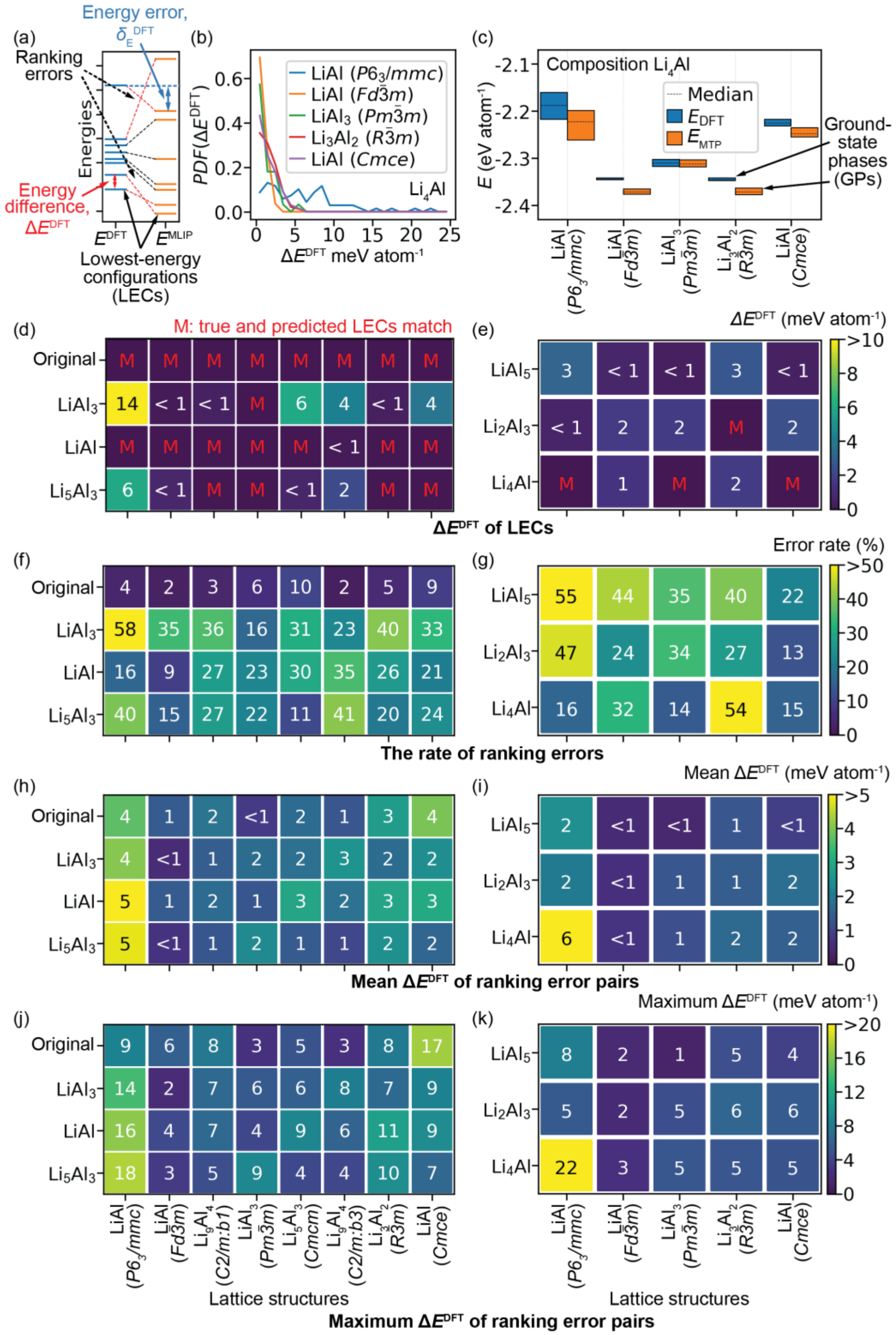

Figure 2. The energy rankings of elemental orderings. (a) The illustration of the energy rankings of different orderings (solid lines) by DFT and MLIP. The dash lines connect the

same ordering configurations, and the crossings of dashed lines indicate ranking errors. δ_E^{DFT} is the energy error between the true DFT and the predicted MLIP energies of the same configuration (blue gap). ΔE^{DFT} is the DFT energy difference of the configuration pair with an energy ranking mismatch (red gap). The lowest-energy lines correspond to the lowest-energy configurations (LECs). (b) The distributions of ΔE^{DFT} for all ranking errors and (c) the box plot of the energy distributions (the maximum, minimum, and median shown in dashed lines) of up to 30 orderings based on five lattice structures in the Li_4Al composition ($R\overline{3}m$ is the true phase, and others are hypothetically generated) from the $\mathcal{D}^{\mathrm{test}}$. (d, e) The differences of DFT energies, ΔE^{DFT}, between the LECs calculated by DFT and $\mathrm{MTP}_{\mathrm{Train}}$, (M indicates the match of LEC predicted by DFT and $\mathrm{MTP}_{\mathrm{Train}}$), (f, g) the rate of ranking errors, (h, i) the mean ΔE^{DFT} and (j, k) the maximum ΔE^{DFT} for all configuration pairs with ranking errors in $\mathcal{D}^{\mathrm{test}}$.

Here we analyze the MLIP's predictions regarding the energies and relative stabilities of different elemental orderings in known and hypothetical phases. The energy ranking of different elemental orderings is relevant for many materials phenomena, such as anti-site defects, diffusion, solid solutions, and phase transitions, and is essential for conducting the random sampling of elemental orderings/configurations in MC simulations for evaluating the thermodynamic properties and the construction of the phase diagrams.

We first test whether the MLIP can correctly predict the lowest-energy configuration (LEC) of each phase, i.e., the most favorable elemental ordering among all orderings that were sampled based on a given lattice structure in $\mathcal{D}^{\mathrm{test}}$ (Methods). For all phases in $\mathcal{D}^{\mathrm{test}}$, the $\mathrm{MTP}_{\mathrm{Train}}$ correctly predicts the LECs of all eight known intermediate phases are by (the 1st row of **Figure 2**d), 50% of 24 hypothetical phases in Figure 2d, and four of 15 hypothetical phases in Figure 2e. Even for the LECs incorrectly predicted by the $\mathrm{MTP}_{\mathrm{Train}}$, the energy differences, ΔE^{DFT}, between the true and the predicted LECs are small, as low as 2 meV atom^{-1} for 13 out of 23 incorrectly predicted LECs, indicating good predictions of the energies. Most errors in the predictions are caused by configurations with similar energies, which were also observed in previous MLIP studies on different defect types or polymorphs of materials.[24–26] In general, the $\mathrm{MTP}_{\mathrm{Train}}$ correctly predicts the LECs of most intermediate phases in the Li-Al alloy system.

We then test whether the MLIP can correctly predict the ground-state phase (GP) of a given composition, which is the most favorable phase at the given composition among all known and hypothetical phases generated based on different lattice structures (Figure 2c, and Supplementary Figure S2 and S3). For example, among all hypothetically generated phases with the $LiAl_4$ composition with the lattice structures of $P6_3/mmc$, $Fd\bar{3}m$, $Pm\bar{3}m$, $R\bar{3}m$, and $Cmce$ (Figure 2c), the $\mathrm{MTP}_{\mathrm{Train}}$ correctly predicts the GP of $LiAl_4$ to be the one with the $R\bar{3}m$ lattice structure. The $\mathrm{MTP}_{\mathrm{Train}}$ also correctly predicted the GPs for LiAl ($Fd\bar{3}m$), $LiAl_3$ ($Pm\bar{3}m$), and $LiAl_5$ ($Pm\bar{3}m$) (Supporting Information). The $\mathrm{MTP}_{\mathrm{Train}}$ incorrectly predicted the GPs of Li_5Al_3 and Li_2Al_3 to be $Fd\bar{3}m$, which should be $Cmcm$ and $Pm\bar{3}m$, respectively, by DFT. Generally, the $\mathrm{MTP}_{\mathrm{Train}}$ shows good performance in predicting LECs and GPs for a wide variety of phases in the alloy system.

To quantify how the MLIP can correctly predict the relative rankings of the energies for many elemental orderings in a given phase, we develop several evaluation metrics based on ranking errors. Ranking error is a pair of atomistic configurations or elemental orderings that exhibit a different energy ranking, i.e., a ranking mismatch, predicted by the MLIP comparing to DFT (indicated as the crossing red dashed lines in Figure 2a). To quantify the error rates of MLIPs on predicting the relative energy ranking, we propose and define the rate of ranking errors as the fraction of mismatched pairs of configurations among all possible pairs, which corresponds to the frequency of how often mismatched pairs occur among all possible pairs of configurations in comparison. This quantity is very similar to the concordance index, which is widely used in biomedical informatics and other fields for quantifying the ranking errors of model predictions.[27]

In addition, we here propose and define the difference of DFT energies, ΔE^{DFT}, between two configurations with a ranking mismatch (red gap, Figure 2a) as another measure of the energy ranking error. This ΔE^{DFT} is different from the commonly used energy error $\delta_{\mathrm{E}}^{\mathrm{DFT}}$ between the MLIP predicted energies and the true DFT for a given atomistic configuration (blue gap, Figure 2a). For the mismatched pairs with large ΔE^{DFT}, the MLIP-based simulations performed in the corresponding materials system are more likely to produce elemental orderings that significantly deviate from the DFT. Thus, the mean ΔE^{DFT} and the maximum ΔE^{DFT} of all mismatched configurations can serve as evaluation metrics to quantify the errors in energy ranking for many elemental orderings predicted by the MLIPs (Methods).

For the $\mathrm{MTP}_{\mathrm{Train}}$, the rate of ranking errors for all known intermediate phases are low (< 10% as shown in the 1st row of Figure 2f). In $\mathcal{D}^{\mathrm{test}}$, a majority (24 out of 39) of hypothetical phases have mismatch probabilities in the range of 11 - 33% (Figure 2f and g), while only three are higher than 50%. As quantified by the mean ΔE^{DFT} and maximum ΔE^{DFT} (Figure 2b), three of the eight intermediate phases have the mean ΔE^{DFT} lower than 2 meV atom^{-1}, so as 20 of the total 39 hypothetical materials. The maximum ΔE^{DFT} for three out of eight known intermediate phases and 21 of 39 hypothetical phases are lower than 6 meV atom^{-1} (Figure 2h - k). Notably, the ΔE^{DFT} distributions of phases with the HCP *P6₃/mmc* lattice structure have larger spreads than most other phases with other lattice structures, indicating a larger ranking error of phases with HCP lattice structures (blue line, Figure 2b and Supplementary Figure S4). The poorer prediction of the HCP lattice structures may be caused by the same nearest-neighbor environments as the FCC configurations, which may be more difficult to differentiate by the atomic descriptors. Overall, the $\mathrm{MTP}_{\mathrm{Train}}$, which is only trained on three phases, shows decent performance in predicting the relative energy rankings of the elemental orderings in most intermediate phases. This implies decent transferability of the MLIPs to many different compositions and lattice structures in the lattice models of alloys. With the aid of these evaluation metrics, the alloy phases with larger errors of energy rankings are identified and are indicated as the potential directions for further improvements.

The error evaluation metrics for the relative energy ranking are important to be considered as independent testing metrics in addition to the widely used average errors in energy and forces. For example, the hypothetical phases of $\mathrm{Li_2Al_3}$ ($P6_3/mmc$) and $\mathrm{LiAl_5}$ (HCP $P6_3/mmc$, BCC $Fd\bar{3}m$, BCC $R\bar{3}m$, and FCC $Pm\bar{3}m$) have small average errors with energy RMSEs at 4 – 10 meV atom^{-1}, but large errors on ranking metrics with the rates of ranking errors at 35 – 55% (Figure 1d, and Figure 2f). In addition, the hypothetical BCC *C2/m:b3* phases have energy RMSEs ranging from 37 to 65 meV atom^{-1} for different compositions, which are significantly larger than the other lattice structures (Figure 1c), while their rates of ranking errors, mean ΔE^{DFT}, and maximum ΔE^{DFT} are small or at comparable levels compared with other phases, ranging from 23% to 41%, 1 to 3 meV atom^{-1}, and 4 to 8 meV atom^{-1}, respectively (Figure 2f, h, and j). Thus, low energy errors may not always indicate good accuracies on energy rankings and vice versa (see ***prediction of ground-state phases for different compositions*** in Supporting Information). It's critical to include the evaluation metrics on ranking errors in the testing procedures of MLIPs in addition to the average errors in energy and forces.

2.3 Effect of training data diversity

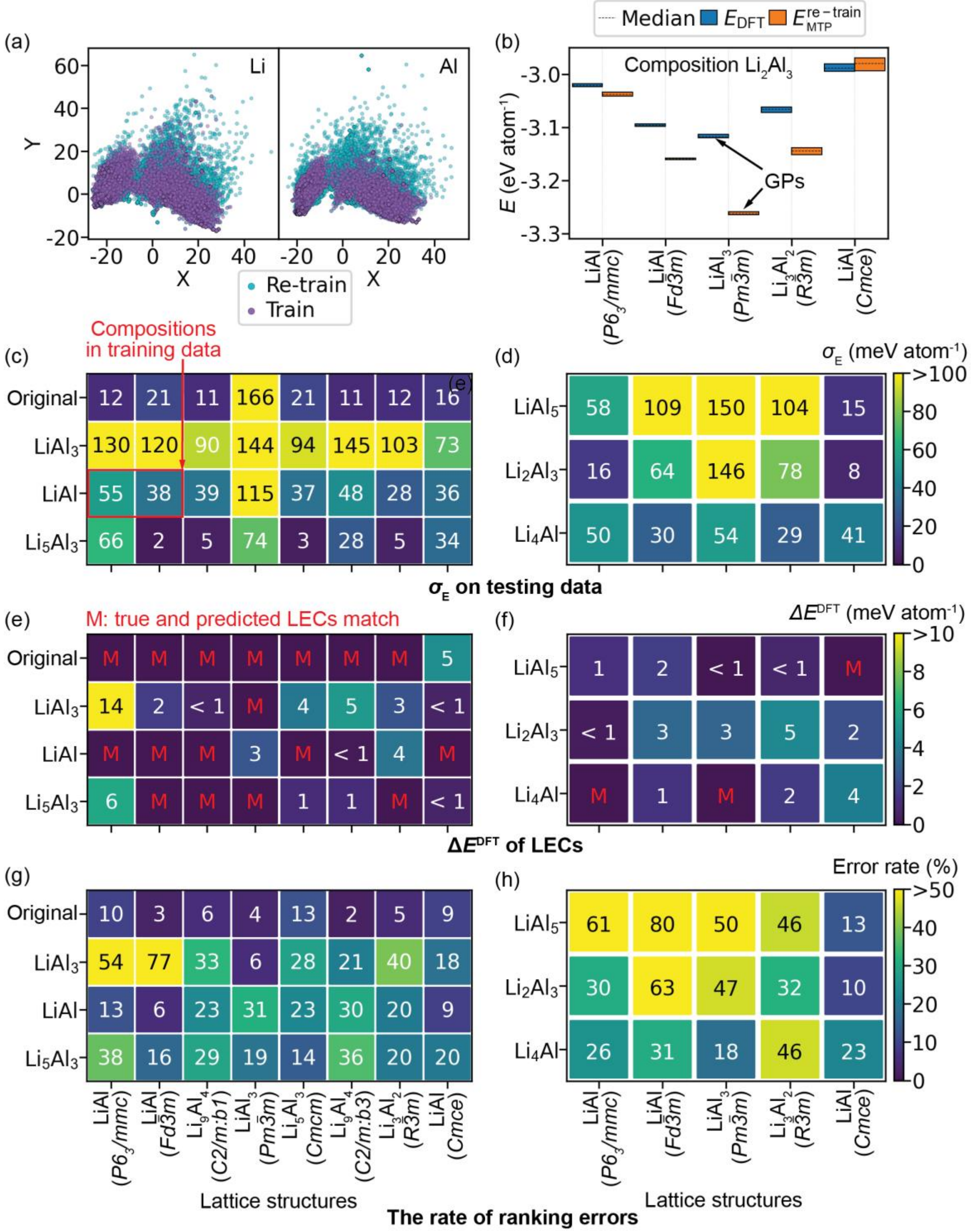

Figure 3. Performances of MTP$_{\text{Re-train}}$ trained with more data diversity. The atomic environments of (a) Li and Al atoms in $\mathcal{D}^{\text{train}}$ (blue) and $\mathcal{D}^{\text{re-train}}$ (orange), plotted by 1st and 2nd principal components from the principal component analysis (PCA) of the Smooth Overlap of Atomic Positions (SOAP) descriptors (Methods). (b) The box plot of the energy

distributions (the maximum, minimum, and median shown in dashed lines) of up to 30 orderings generated in five lattice structures with the Li_2Al_3 composition from $\mathcal{D}^{\mathrm{test}}$ ($Pm\bar{3}m$ as the true phase, and others are hypothetically generated). (c, d) The RMSEs of predicted energies of up to 30 orderings in a lattice structure of known intermediate phases (*x* axis) with different compositions (*y* axis). (e, f) The differences of DFT energies, ΔE^{DFT} between the LECs calculated by DFT and $\mathrm{MTP}_{\mathrm{Re\text{-}train}}$ for $\mathcal{D}^{\mathrm{test}}$ as Figure 2d and e. (g, h) The rate of ranking errors.

In this section, we explore the effect of increasing the diversity in the training data on the performance of MLIP, which is a common practice to improve ML models including MLIPs.[11,21,28,29] The new training dataset, $\mathcal{D}^{\mathrm{re-train}}$, includes one more material, HCP LiAl ($P6_3/mmc$), in addition to the BCC Li ($Im\bar{3}m$), FCC Al ($Fm\bar{3}m$), and BCC LiAl ($Fd\bar{3}m$) in the original training dataset $\mathcal{D}^{\mathrm{train}}$ for $\mathrm{MTP}_{\mathrm{Train}}$ (section 2.1 - 2.2). A total of 144 atomistic configurations with the HCP LiAl ($P6_3/mmc$) were generated in the same manner as other phases, including crystalline bulk with a range of orderings, liquid, point defects, and AIMD snapshots of bulk and defect bulks, and were added into $\mathcal{D}^{\mathrm{re-train}}$ by replacing 24% of configurations in $\mathcal{D}^{\mathrm{train}}$. The $\mathcal{D}^{\mathrm{re-train}}$ has increased data diversity with an additional lattice structure and has a 16% increase in the data size. The $\mathrm{MTP}_{\mathrm{Re\text{-}train}}$ re-trained on $\mathcal{D}^{\mathrm{re-train}}$ is obtained following the same training and validation process (Methods).

Same as the test of $\mathrm{MTP}_{\mathrm{Train}}$ in section 2.2, we perform the test of the $\mathrm{MTP}_{\mathrm{Re\text{-}train}}$ on $\mathcal{D}^{\mathrm{test}}$. For hypothetical phases outside the training data, $\mathrm{MTP}_{\mathrm{Re\text{-}train}}$ performs significantly poorer than $\mathrm{MTP}_{\mathrm{Train}}$, with only six hypothetical phases having energy RMSEs lower than 15 meV atom^{-1}, compared to 28 from $\mathrm{MTP}_{\mathrm{Train}}$ (**Figure 3**c and d). For example, $\mathrm{MTP}_{\mathrm{Re\text{-}train}}$ gives a higher energy RMSE of 103 meV atom^{-1} for the hypothetical $LiAl_3$ in the BCC $R\bar{3}m$ lattice structure, whereas $\mathrm{MTP}_{\mathrm{Train}}$ gives only 4 meV atom^{-1}. The $\mathrm{MTP}_{\mathrm{Re\text{-}train}}$ also gives poorer predictions in energy rankings, especially for hypothetical phases with high rate of ranking errors. There are six materials exhibiting large rate of ranking errors (> 50%), with a rate of ranking error as high as 80% on $LiAl_5$ $Fd\bar{3}m$ (Figure 3g and h). The $MTP_{Re\text{-}train}$ also performs poorer than MTP_{Train} on other metrics of energy rankings. The mean ΔE^{DFT} of ranking errors increases to 4 – 11 meV atom^{-1} from 2 – 4 meV atom^{-1} for known intermediate phases LiAl *P6₃/mmc*, Li_9Al_4 *C2/m:b1*, and LiAl *Cmce*. There are 18 hypothetical phases from $\mathcal{D}^{\text{test}}$ with the mean ΔE^{DFT} of ranking errors lower than 2 meV atom^{-1}, down from 20 of MTP_{Train} (Supporting Information). The number of correctly predicted LECs also decline to 13 (from 16 by MTP_{Train}) (Figure 3e and f).

Three additional re-trained MTPs, $MTP_{Re\text{-}train}^{extra\text{-}1}$, $MTP_{Re\text{-}train}^{extra\text{-}2}$, and $MTP_{Re\text{-}train}^{extra\text{-}3}$, which are selected by different validation procedures (Methods), are examined. All MTP models have low RMSE of energies as low as at 8 meV atom^{-1} on $\mathcal{D}^{\text{train}}$ (24 meV atom^{-1} on $\mathcal{D}^{\text{re-train}}$) but have similarly poor performances in the above tests on $\mathcal{D}^{\text{test}}$ (Supplementary Figure S14, S16, and S18 in Supporting Information). Therefore, the poorer performance of $MTP_{Re\text{-}train}$ is caused by the new training dataset and not by the selection of a particular model (Supplementary Figure S14, S16, and S18 in Supporting Information). Given in this case the MTPs trained with more diverse data does without a much larger training data size may not show improved performance, further studies are needed to understand how to select the training data for further MLIP improvement.

2.4 Errors in the applications of MLIPs

In the previous sections, the MLIPs were tested on a diverse range of orderings/configurations, most of which were hypothetically generated for testing and were not fully covered in the training dataset. While the generation of these orderings/configurations for testing aim to mimic those in atomistic simulations, this section further tests MLIPs in three application cases commonly conducted for alloy materials research.

2.4.1 Errors on elemental orderings from Monte Carlo simulations

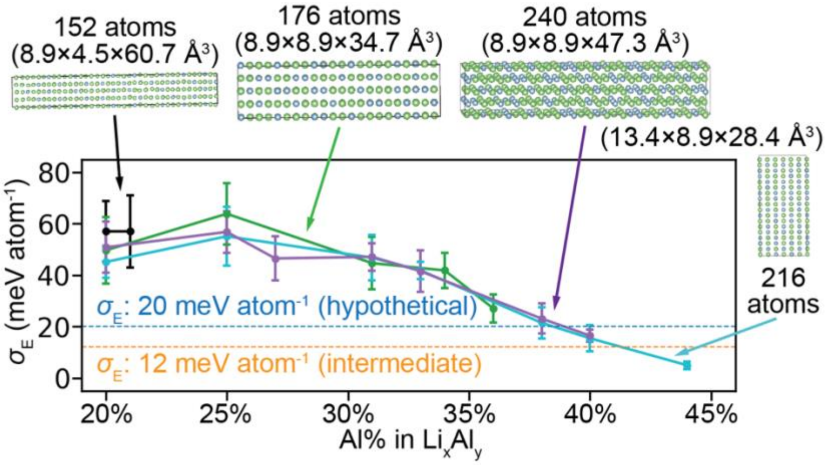

Figure 4. The energy RMSEs of the supercells with different sizes over a range of Li_xAl_y compositions and different elemental orderings from Monte Carlo Simulations. Dashed lines are the energy RMSEs of all known intermediate phases (orange) and hypothetical phases from $\mathcal{D}^{\text{test}}$ (blue).

Given that a key advantage of MLIP is to conduct atomistic simulations in larger length scales not accessible by DFT, we here test the MTP_{Train} in MC simulations using large supercells. The MC simulations are conducted in four large supercells with dimensions 8.9×4.5×60.7 Å^3, 8.9×8.9×34.7 Å^3, 8.9×8.9×47.3 Å^3, and 13.4×8.9×28.4 Å^3 consisting of 152, 176, 216, and 240 atoms, respectively (Methods) (**Figure 4**). For each supercell, we test multiple compositions ranging from Li_4Al to Li_5Al_4 based on the same lattice structure, with the initial structures as an interface slab between BCC metal Li and BCC LiAl ($Fd\overline{3}m$) with different numbers of atomic layers for each phase. Monte Carlo simulations were performed for this interface until an equilibrium was achieved (Methods). The structures from these MC simulations are referred to as the dataset $\mathcal{D}^{\text{slabs}}$ (Methods). While the $\text{MTP}_{\text{Train}}$ gives the energy RMSEs as low as 12 meV atom^{-1} for all known intermediate phases and 20 meV atom^{-1} for all phases in $\mathcal{D}^{\text{test}}$ as shown in section 2.1 – 2.2, the $\text{MTP}_{\text{Train}}$, as tested on these large supercells, exhibits significantly higher energy RMSEs of > 30 meV atom^{-1} at < 35% Al percentage (Figure 4) and as high as 50 – 60 meV atom^{-1} at around 25% Al (i.e., Li_3Al) (Figure 4).

To further identify the causes of the discrepancies, a separate test was performed on large supercells with elemental orderings generated in the same manner as the testing dataset for the BCC LiAl ($Fd\overline{3}m$) lattice at 25% Al percentage. The results show that the energy RMSEs of these supercells are in the range of 24 to 26 meV atom^{-1} (Supplementary Figure S25, see ***errors on large supercells*** in Supporting Information) much lower than those (50 to 60 meV atom^{-1}) encountered in the MC simulations in Figure 4. The PCA of SOAP descriptors on the atomic environments from the MC simulations are confirmed to be different (Supplementary Figure S19, Supporting Information). These results suggest that the errors may be caused by the different elemental orderings or configurations encountered during the MC simulations with large cells rather than merely having large cell sizes.

Given the low errors of $\text{MTP}_{\text{Train}}$ as tested on the typical supercell sizes in $\mathcal{D}^{\text{test}}$ (section 2.1 - 2.2), these discrepancies indicate a potential issue in applying MLIPs in larger simulation cells, as the errors may emerge for those orderings/configurations encountered in the large supercells from actual atomistic simulations that were not encountered in the testing process.

2.4.2 Convex hull and phase transition

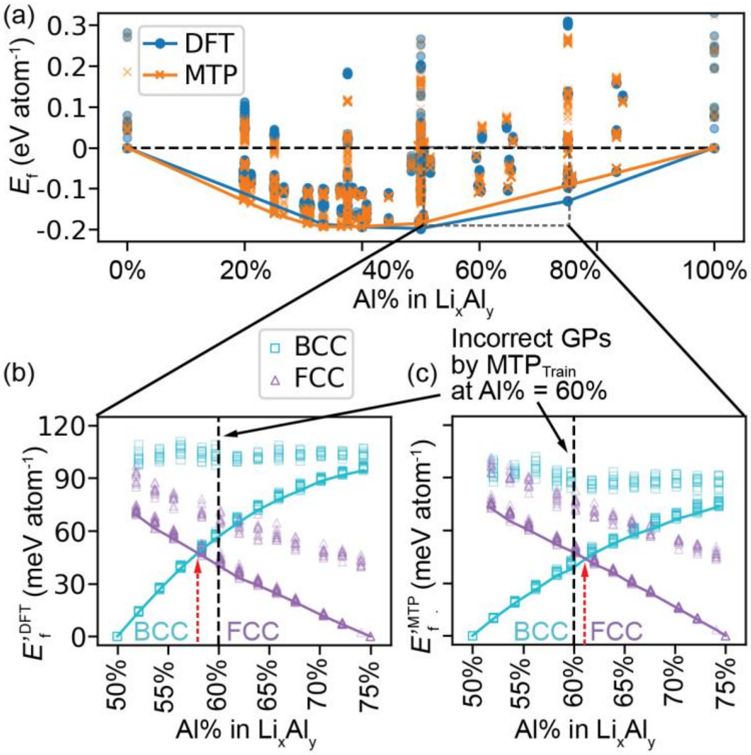

Figure 5. The convex hull of Li-Al alloys. (a) The formation energies E_f and the convex hulls of Li-Al alloy system by DFT (blue, circles) and MTP_{Train} (orange, crosses). (b, c) The phase transition between BCC LiAl (*Fd*$\bar{3}$*m*) (cyan squares) and FCC $LiAl_3$ (*Pm*$\bar{3}$*m*) (purple triangles) predicted by (b) DFT and (c) MTP_{Train}. Red dash arrow indicates the transition point based on E_f. The black dashed line at 60% Al indicates Li_2Al_3, which has incorrectly predicted GP by MTP_{Train} (Figure 2c) (Supporting Information).

Here, we compare the energy convex hulls of the Li-Al alloy system constructed using MTP_{Train} and DFT energies (**Figure 5a**). The convex hull is constructed using the calculated formation energies E_f (Methods) of all configurations from $\mathcal{D}^{\mathrm{train}}$ and $\mathcal{D}^{\mathrm{test}}$ (Figure 1c and d), the relaxed structures of all eight intermediate phases from the ICSD (Figure 1a), the structures with large supercells in $\mathcal{D}^{\mathrm{slabs}}$ in section 2.4.1 (Figure 4a), and the structures with single vacancies in $\mathcal{D}^{\mathrm{Vacancy}}$, which includes complex elemental orderings of known intermediate phases with either a Li or an Al vacancy (Supporting Information). Compared with DFT, the $\mathrm{MTP}_{\mathrm{Train}}$ largely reproduces the overall shape of the convex hull, with some discrepancies in equilibrium phases and corresponding energies (Figure 5). The phase with minimum E_f in the convex hull is predicted to be LiAl and $\mathrm{Li_5Al_3}$ by DFT and $\mathrm{MTP}_{\mathrm{Train}}$, respectively. While there are only four stable intermediated phases predicted by DFT, the $\mathrm{MTP}_{\mathrm{Train}}$ predicted four additional stable intermediate phases in the 19% – 40% Al composition range.

We further examine the BCC-FCC phase transition by comparing the energies of all configurations generated based on the BCC ($Fd\bar{3}m$) and FCC ($Pm\bar{3}m$) lattice structures in the $\mathcal{D}^{\mathrm{bcc/fcc}}$ dataset (Methods) over the 50% to 75% Al composition range (Figure 5b and c). The BCC-FCC transition is predicted to be at 57% Al by the $\mathrm{MTP}_{\mathrm{Train}}$ in good agreement with that at 62% Al by DFT. The deviations in the energies of these phases over the composition range may cause the minor discrepancy. These small energy errors also cause the incorrect prediction of the GP of $\mathrm{Li_2Al_3}$ by $\mathrm{MTP}_{\mathrm{Train}}$ as shown in section 2.2 (Figure 2c) and can be directly observed in Figure 5b and c (Supplementary Figure S2, Supporting Information). This indicates another case where minor energy errors can lead to different physical outcomes of MLIP predictions.

2.4.3 Errors of forces on migrating atoms in MD simulations

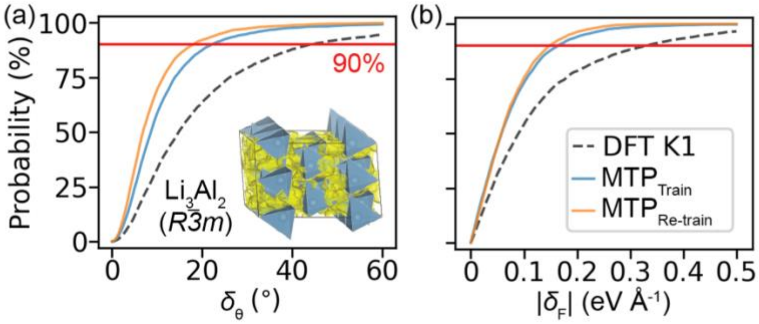

Figure 6. Errors of forces predicted by MLIPs on migrating atoms in MD simulations. The cumulative distribution functions (CDFs) of the force errors, including (a) the directions, δ_θ, and (b) the magnitudes, $|\delta_F|$, of atomic forces predicted by DFT K1 (dashed black lines), MTP_{Train} (blue), and $MTP_{Re\text{-}train}$ (orange) on migrating atoms in snapshots from AIMD simulations of Li_3Al_2 ($R\bar{3}m$). Red lines are visual guidance for 90%.

Here we test the MLIP in the study of atom diffusion in MD simulations, another key application of MLIPs. Previous study reveals that atomic forces on migrating atoms are a key error source of MLIPs.[25] Here, we quantify the errors of atomic forces on migrating atoms from 200 snapshots extracted from AIMD simulations of LiAl ($Fd\bar{3}m$), Li_3Al_2 ($R\bar{3}m$), Li_5Al_3 (*Cmcm*), and Li_9Al_4 (*C2/m:b1*) with a single Al vacancy in the supercell model (Methods) (Supporting Information, **Figure 6**a, b, and Supplementary Figure S20). MTP_{Train} and $MTP_{Re\text{-}train}$ achieve good accuracy in predicting atomic forces on migrating atoms for all four materials, even though the three intermediate materials Li_3Al_2, Li_5Al_3, and Li_9Al_4 are not in the training data. As shown in the CDF curves of force errors δ_θ and $|\delta_F|$, on migrating atoms obtained from Li_3Al_2 AIMD diffusion snapshots (Figure 6), over 90% of migrating atoms predicted by MTP_{Train} has less than 22 degrees of errors in force direction δ_θ and less than 0.16 eV Å^{-1} of errors in force magnitude $|\delta_F|$. The forces errors of the MTPs are compared to the DFT calculations with a single Γ-centered **k**-point (DFT K1) as in Ref. [25], which is commonly used in AIMD simulations for lower computation costs. The DFT K1 shows larger errors of 44 degrees in force direction δ_θ and of 0.33 eV Å^{-1} in force magnitude $|\delta_\mathrm{F}|$, to meet the same 90% cut-off (Figure 6a, b, and Supplementary Figure S20). The force performance score $P(\mathcal{D}_{\mathrm{Li_3Al_2}})$, which is a metric proposed by Ref.[25], is 0.69 and 0.73 for the $\mathrm{MTP_{Train}}$ and $\mathrm{MTP_{Re\text{-}train}}$, respectively, compared to 0.50 from DFT K1, in the dataset $\mathcal{D}_{\mathrm{Li_3Al_2}}$ of forces on migrating atoms for AIMD snapshots in material Li_3Al_2 ($R\overline{3}m$). Even for Li_3Al_2 ($R\overline{3}m$) and Li_5Al_3 (*Cmcm*) that are not in the training datasets $\mathcal{D}^{\mathrm{train}}$ or $\mathcal{D}^{\mathrm{re-train}}$, the $\mathrm{MTP_{Train}}$ and $\mathrm{MTP_{Re\text{-}tain}}$ outperform DFT K1 on force errors. It should be noted that DFT calculations for metals and alloys often require a denser **k**-point mesh, and K1 is a relatively coarse **k**-point sampling for the metal supercell models used. Overall, the MTP models show good performances in predicting forces on migrating atoms, which are not covered in the training data. Since force predictions on migrating atoms are critical for diffusional properties, this good performance of MTP on the migrating atoms not included in the training data is very encouraging. Nevertheless, careful testing is always needed before conducting the atomistic simulations of the MLIPs.

3. Discussion and Conclusions

In this study, we conducted a systematic test of the MLIPs using Li-Al alloy as a model system to answer the question of whether the MLIPs can make accurate predictions for many different configurations that may be encountered in atomistic modeling of alloys. In the model alloy system, our results show that the MLIPs trained using only a few phases can largely predict other intermediate phases (both known and hypothetically generated) with a wide variety of compositions, elemental orderings, and lattice structures. The MLIPs trained on only a few phases with carefully selected configurations capture the key physics of interatomic bonds in lattice-based models to accurately predict other alloy intermediate phases with deviated compositions and non-equilibrium elemental orderings. These results along with many previous studies[12,13,33,14–16,18,19,30–32] show great promises and potentials of applying MLIPs to studying complex alloy systems.

Our process of testing MLIPs can be further developed into a general test, including 1) generating the testing data based on hypothetical phases that mimic the atomistic simulations, and 2) quantifying the error evaluation metrics for the relative stabilities of elemental orderings. First, the MLIPs testing should be performed on the testing data that are generated to mimic the key physical (or error-prone) scenarios in atomistic simulations. For example, in this study, different hypothetical phases with many elemental orderings were generated based on distinct lattice structures over a wide range of compositions, mimicking the random sampling during MC simulations with canonical (fixed composition) or grand canonical ensemble (varying compositions) commonly employed for alloy modeling. Such tests allow a systematic examination of the MLIP performance on a diverse range of atomic configurations in many phases with a wide range of compositions, lattice structures, and complex elemental orderings across the alloy phase diagram.

To evaluate the MLIP's performance on these complex elemental orderings that are relevant to many materials properties/phenomena, multiple evaluation metrics are developed and employed in the testing process. In addition to the commonly used averaged errors, e.g., RMSEs of energies, we propose and demonstrate the rate of ranking errors based on the occurrence frequency of mismatched energies between MLIPs and DFT, the mean and the maximum of the energy differences on mismatched pairs of MLIPs and DFT energies to quantify the predicted energy ranking of different elemental orderings in different phases. The accuracy of MLIPs is also assessed by the correct predictions of GPs and the LECs for a variety of phases. As revealed in our study, even if the MLIPs may have small average errors predicting the energies of different phases, relative stabilities of different elemental orderings and configurations, such as energy ranking, LECs and GPs, may still be incorrectly predicted. Given the importance of correctly predicting the relative stabilities of various configurations and phases for studying many materials properties and phenomena, these evaluation metrics based on energy rankings and relative stabilities should be considered as key components, in addition to average errors, of the MLIP performances in the future testing process.

Our MLIP testing process can also be considered as a testing and quantification of the 'transferability' of MLIPs in predicting other phases, elemental orderings, compositions, and lattice structures, outside of the training data. The testing on the known phases that are intentionally leave out of training is in principle similar to the leave-one-out cross-validation or bootstrapping strategies in general machine-learning. Employing this strategy in the materials system space allows the error quantifications on the materials outside of the training data, and also mimics the situation of encountering unknown materials in the atomistic simulations using MLIPs. This strategy may be further developed and generalized for testing MLIPs and different training processes.[16]

As shown in our study, the complex atomistic configurations and distinct elemental orderings encountered in large supercell models in actual atomistic simulations may still show notable discrepancies. Developing systematic testing processes that mimic the real applications of MLIPs in larger models is needed to further test and improve the MLIPs in those application scenarios.[34] The difficulty in addressing this issue is that these errors may only be encountered in large supercells that would be computationally expensive to calculate by DFT, thus impeding the direct test and correction of these errors. One future research direction is to generate relevant configurations that can represent those in larger length-scale simulations but in adequately small supercells to be verified by DFT calculations.

Our results also provide insights about the complexity regarding the training strategies of MLIPs. While good transferability of MLIPs to other intermediate alloy phases can be achieved by using a small number of phases in the training data, it is unexpected that the MLIPs trained with more data diversity (i.e., more phases) have worse performances on most metrics for most phases, even for those included in the training dataset. As shown in this studied case, increasing the data diversity alone without substantially increasing the amount of training data may not always lead to improved MLIPs. This observed phenomenon and previous studies[21] suggest that the effects of data size and data diversity on MLIP performance are complex. More research is needed to fully understand strategies for balancing the training data size and diversity for achieving improved performance of MLIPs. The testing process and aforementioned metrics in quantifying the performance and transferability of MLIPs can be used to assess different training strategies for improving MLIPs.

In summary, our study provides a comprehensive case study using a binary alloy system on the performance, testing, transferability, and application errors of MLIPs. These insights provide some guidance for future improvement of MLIPs. Our testing process and evaluation metrics for quantifying the performance and transferability of MLIPs should be considered in future studies of MLIPs.

Methods

First-principles computation. Vienna ab initio simulation package[35] (VASP) with Perdew-Burke-Ernzerhof[36] (PBE) functionals by generalized-gradient approximation (GGA) was used to perform all density functional theory calculations. The projector augmented-wave approach was adopted in DFT calculations to relax crystal structures and get energies and forces. All true values of energies and forces for training, validating, and testing MTPs were calculated using 4×4×4 **k**-point mesh (K4). The Gaussian smearing method was adopted to calculate partial occupancies for each orbital (ISMEAR = 0) in combination with smearing width set to 0.05 eV (SIGMA = 0.05). Furthermore, all DFT calculations were spin-polarized with an energy cutoff set to 520 eV, electronic relaxation cutoff set to 10^{-7} eV, and other parameters set to be compatible with the Materials Project.[37,38]

Ab initio molecular dynamics simulation. Ab initio molecular dynamics (AIMD) simulations were performed as the same scheme described in Ref. [39]. The same GGA for PBE functionals as described in *First-Principles Computation* section were used. The AIMD simulations were non-spin-polarized, with the setting of electronic energy convergence cutoff to 10^{-4} eV, a time step of 2 fs, and a Γ-centered 1×1×1 **k**-point. All simulations adopted the NVT ensemble with Nose-Hoover thermostat (SMASS = 0). The initial structures were heated up from 100 K to the target temperatures with a constant heating rate using velocity scaling during a period of 2 ps. All supercell models for AIMD simulations have the lattice parameters at around 10 Å or larger. The temperatures of 600

K, 2000 K, and 400 K were used to generate the near-equilibrium configurations of intermediate phases, liquid phases, and interfaces, respectively.

The AIMD simulations for the vacancy diffusion in LiAl (*Fd$\overline{3}$m*), Li_9Al_4 (*C2/m:b1*), Li_5Al_3 (*Cmcm*), and Li_3Al_2 (*R$\overline{3}$m*) intermediate phases in section 2.4.3 and in Supporting Information (Supplementary Figure S20) were carried out following the established process in Ref. [39]. To compute accurate distributions of force errors on migrating atoms and the corresponding force performance metrics, such as normalized area under the curve (NAC), an adequate number of events with atom migration need to be obtained from AIMD simulations. The diffusion simulations of all four materials were performed at the temperatures of 400, 450, 500, 550, 600, 650, and 700 K for 400 to 1000 ps. Migrating atoms in each snapshot of all AIMD simulations were identified following the process in Ref. [40] and the details were clarified in the section ***identifying migrating atoms and evaluating force errors***. 200 snapshots with the highest number of migrating atoms were collected from AIMD simulations at all temperatures for each material. The migrating atom in these 200 snapshots were then used to calculate the force errors of MTP models. The supercells used in AIMD simulations of four materials were the same with the supercells of original intermediate phases in section 2.1 of main text, removing an Al atom to create a single vacancy.

Datasets for training, validation, and testing. A total of eight datasets were generated for training, validating, and testing MTPs, such as two training datasets $\mathcal{D}^{\mathrm{train}}$ and $\mathcal{D}^{\mathrm{re-train}}$, two validation datasets $\mathcal{D}^{\mathrm{validation}}$ and $\mathcal{D}^{\mathrm{irregular}}$, and four testing datasets $\mathcal{D}^{\mathrm{test}}$, $\mathcal{D}^{\mathrm{slabs}}$, $\mathcal{D}^{\mathrm{Vacancy}}$, and $\mathcal{D}^{\mathrm{bcc/fcc}}$. The energies and the forces of all configurations in these datasets were converged using K4 by single-step self-consistent DFT calculations (except for $\mathcal{D}^{\text{irregular}}$, or indicated otherwise). The numbers of atoms for the supercells of each intermediate phase in $\mathcal{D}^{\text{test}}$ are, 54 for Li (*Im*$\overline{3}m$), 32 for LiAl (*P6*$_3$*/mmc*), 144 for LiAl (*Fd*$\overline{3}m$), 104 for Li_9Al_4 (*C2/m:b1*), 108 for $LiAl_3$ (*Pm*$\overline{3}m$), 48 for Li_5Al_3 (*Cmcm*), 234 for Li_9Al_4 (*C2/m:b3*), 60 for Li_3Al_2 (*R*$\overline{3}m$), 96 for LiAl (*Cmce*), and 108 for Al (*Fm*$\overline{3}m$).

The $\mathcal{D}^{\text{train}}$ dataset contains 356 configurations based on BCC Li metal (*Im*$\overline{3}m$, ICSD 77370), LiAl (*Fd*$\overline{3}m$, ICSD 240122), and FCC Al metal (*Fm*$\overline{3}m$, ICSD 606000). All supercells in $\mathcal{D}^{\text{train}}$ have the lattice parameters around or larger than 10 Å. For each of these three phases, a variety of configurations were generated as follows:

(1) The ground-state configurations fully relaxed by DFT K4.

(2) Five near-equilibrium configurations of the ground-state generated by NVT AIMD simulations at 600 K. Each of the near-equilibrium configurations was generated by a separate AIMD simulation of 2 ps.

(3) Two liquids configurations generated by NVT AIMD simulations at 2000 K, following the same process as 2).

(4) Strained supercells with a) -15% or +15% strain on each of the three lattice parameters, and b) -15% or +15% strain on a pair of distinct lattice parameters with four combinations (+15%, +15%), (-15%, -15%), (+15%, -15%), and (-15%, +15%) strains.

(5) Distorted supercells generated by distorting each of the three lattice angles by -13.5° or 13.5° for cubic lattices (-15% or +15% of corresponding lattice angles if lattice angles are not 90°).

(6) Supercells with vacancies. We removed 5%, 10%, or 15% of atoms in the supercell of Li and Al. For LiAl ($Fd\overline{3}m$), the defected supercells with either Li or Al vacancies were generated, but not both Li and Al vacancies at the same time. All supercells with vacancies were then fully relaxed by DFT K4.

(7) Five near-equilibrium configurations for each supercell with vacancies. All configurations were generated by NVT AIMD simulations at 600 K, following the same process in 2).

(8) Supercells with interstitials. We added one, two, or four atoms of either Li or Al in the supercells of Li, Al, and LiAl. All bulk supercells were then fully relaxed by DFT K4.

(9) Five near-equilibrium configurations for each of the supercells with interstitials. All configurations were generated by NVT AIMD simulations at 600 K, following the same process in 2).

(10) Additionally, 20 interfaces with every pair of two phases in the dataset. All 20 interfaces were generated by directly joining the conventional unit cell of each material by the algorithm by Zur and McGill [41] implemented in Pymatgen[37] Python package. The maximum matching area of the algorithm was set to 400 $Å^2$ and the maximum angle tolerance was set to 0.01. Slabs of both materials on either side of the interface had four layers of unit cells. We generated interfaces using the Miller indices of {100}, {110}, and {111}. The gap distance and the vacuum space at the interface core were both set to 2.5 Å. Due to the high computational cost of evaluating energies and forces for these large interface supercells, we excluded the supercells that had more than 250 atoms or had lattice parameters larger than 25 Å. For each interface supercell, four configurations were generated as follows,

i. A supercell of the interface configuration without relaxing lattice or atom positions.

ii. A supercell of the interface configuration relaxed by DFT K1. The energy and atomic forces of the relaxed supercell were then calculated by DFT K4 self-consistent run with fixed lattice and atom positions.

iii. Two near-equilibrium configurations for the relaxed supercell generated by NVT AIMD simulations at 400 K.

The $\mathcal{D}^{\mathrm{re-train}}$ dataset contains 414 configurations based on four phases: BCC Li metal (*Im*$\overline{3}$*m*, ICSD 77370), LiAl (*Fd*$\overline{3}$*m*, ICSD 240122), LiAl (*P6$_3$/mmc*, ICSD 262069), and FCC Al metal (*Fm*$\overline{3}$*m*, ICSD 606000). All configurations were generated following the same process as $\mathcal{D}^{\mathrm{train}}$, except for near-equilibrium configurations generated in 2), 7), and 9). In $\mathcal{D}^{\mathrm{re-train}}$, near-equilibrium configurations were generated by displacing each atom to a random direction and the displacement of each atom was randomly selected using a uniform distribution between 0 and 0.5 Å, and only three near-equilibrium configurations were generated for 2), 7), and 9) in $\mathcal{D}^{\mathrm{re-train}}$. A total of 42 interface supercells were generated in the $\mathcal{D}^{\mathrm{re-train}}$ following the same process of $\mathcal{D}^{\mathrm{train}}$. Three configurations of each interface supercell were generated, including a supercell configuration without relaxation, a supercell configuration relaxed by DFT K1, and a near-equilibrium configuration generated by the random displacement.

The $\mathcal{D}^{\mathrm{validation}}$ dataset contains 66 configurations generated from NVT AIMD simulations of relaxed bulk, vacancy, and interstitial supercells at 600 K, following the same process used in 2) for $\mathcal{D}^{\mathrm{train}}$. The bulk, vacancy, and interstitial supercells were the same as 1), 6), and 8) in $\mathcal{D}^{\text{train}}$ (a total of 33 supercells), and two near-equilibrium configurations of each supercell were taken following the same as 2), 7), and 9) in $\mathcal{D}^{\text{train}}$.

The $\mathcal{D}^{\text{test}}$ dataset contains 1391 configurations based on eight intermediate phases from the ICSD and 39 hypothetical phases (Figure 1c and d). Hypothetical phases were generated using the lattice structures of eight known intermediate phases and changing the site occupancies to given compositions, consisting of 24 in three compositions $LiAl_3$, LiAl, and Li_5Al_3 in eight known intermediate phases (Figure 1c) and 15 phases with three handpicked compositions, $LiAl_5$, Li_2Al_3, and Li_4Al (Figure 1d). The hypothetical phases were generated by adjusting their compositions from the original intermediate phases to the target compositions in the relaxed supercells. The compositions were adjusted by correspondingly changing the occupancies of Li and Al sites in the relaxed supercells of the original intermediate phases, and the partial occupancies may be adjusted to include vacancies to guarantee the numbers of Li and Al atoms are integers (such as Li_5Al_3 with lattice structures $Pm\bar{3}m$, *C2/m:b3*, and $R\bar{3}m$, and Li_4Al with lattice structures $P6_3/mmc$, $Pm\bar{3}m$, and *Cmce*). Then, the structure with modified site partial occupancies was ordered following the scheme in Ref.[42] The configurations of different elemental orderings for all intermediate phases were generated by swapping random numbers of Li and Al atoms in the supercell as follows. First, we randomly picked *n* Li and *n* Al atoms (*n* is a random number between 1% and 10% of the total number of atoms in the supercell) and swapped their positions. We generated up to 30 configurations for each intermediate phase. All configurations were symmetrically distinct to each other, as checked by the scheme used previously in Ref.[39,42].

The $\mathcal{D}^{\text{slabs}}$ dataset contains 420 configurations of interface supercells with a range of compositions with different orderings generated by MC methods. Interface supercells in $\mathcal{D}^{\text{slabs}}$ were between BCC metal Li and BCC LiAl ($Fd\overline{3}m$) based on the relaxed conventional unit cell, following the same process of generating interfaces in $\mathcal{D}^{\text{train}}$. The maximum matching area of the Zur and McGill algorithm was set to 100 Å^2 and the maximum angle tolerance was set to 0.01. The Miller indices of both the substrate and the film were set to {100}. four types of supercells were used, 8.9×4.5×60.7 Å^3, 8.9×8.9×34.7 Å^3, 8.9×8.9×47.3 Å^3, and 13.4×8.9×28.4 Å^3, corresponding to 2×2×5.5, 2×2×7.5, 2×1×9.5, and 3×2×4.5 of the conventional LiAl ($Fd\overline{3}m$) unit cell (Figure 4). The compositions of supercells were handpicked and rounded to the closest Al concentration at 20%, 21%, 25%, 33%, 34%, 36%, 38%, 40%, and 45%. The compositions were changed by adjusting the number of layers of unit cells for Li and LiAl. A total of 21 different compositions in four supercells were generated by MC simulations using MTP$_{\text{Train}}$ (Figure 4). The MC simulation was performed at 300 K (equivalent to k_BT = 0.026 eV) to find the lowest energy configurations by swapping atom positions of a random Li and a random Al for each attempted random move and was terminated until the relative standard deviations of the total energies of the last 1000 attempts were lower than 10^{-5}. During the MC simulations, the supercells were relaxed to relieve the strain caused by the difference of bond lengths between BCC Li and BCC LiAl, with energy convergence set to 10^{-5}. Since the bonds in BCC LiAl are 8% less than Li-Li bonds in BCC Li, the volume change between initial interfaces and the final mixed configurations cannot be neglected and the additional relaxations are therefore necessary. The MC simulations were conducted using LAMMPS by our house-customized scripts. The final 20 configurations that were accepted in each MC simulation were included in $\mathcal{D}^{\mathrm{slabs}}$, and their energies were calculated by DFT K4 without relaxing lattices or atom positions.

The $\mathcal{D}^{\mathrm{Vacancy}}$ dataset contains 350 configurations based on eight intermediate phases with a single vacancy of either Li or Al (see ***errors of MLIPs on dataset with single vacancies*** Supporting Information). In the relaxed supercells of intermediate phases, a random Li or Al atom was chosen to swap with another random atom (either Li or Al) and was then removed. By using this method, we generated up to 30 configurations that were symmetrically distinct (including an anti-site in some cases) for each phase containing one vacancy.

The $\mathcal{D}^{\mathrm{bcc/fcc}}$ dataset, used to examine the BCC-FCC phase transition in section 2.4.2, contains 701 configurations based on the structures with BCC and FCC lattices with 12 compositions with Al % of 52%, 54%, 56%, 57%, 60%, 62%, 64%, 66%, 68%, 70%, 72%, and 74% for each lattice. The structure lattice of metal Li ($Im\bar{3}m$) and LiAl ($Fd\bar{3}m$) for BCC, and metal Al ($Fm\bar{3}m$) and $LiAl_3$ ($Pm\bar{3}m$) for FCC, were adopted to generate configurations with different elemental orderings. A total of 48 hypothetical phases were generated. The compositions of these phases were adjusted following the same scheme as used for $\mathcal{D}^{\mathrm{test}}$. For each phase, following the same process of ordering of hypothetical phases in $\mathcal{D}^{\mathrm{test}}$, up to 10 symmetrically distinct configurations were generated for metal Li ($Im\bar{3}m$) and Al ($Fm\bar{3}m$), and up to 20 configurations were generated for LiAl ($Fd\bar{3}m$) and $LiAl_3$ ($Pm\bar{3}m$).

The $\mathcal{D}^{\mathrm{irregular}}$ dataset was a validation dataset only used to select the optimized MTPs as described in ***Optimizing MTPs and fine-tuning the hyperparameters***. This dataset contains 141 configurations, which were intentionally generated as far-from-equilibrium configurations that give either exceptionally low energies predicted (< -10 eV atom^{-1}, compared with -1.9 eV atom^{-1} for bulk Li and -3.7 eV atom^{-1} for bulk Al) or exceptionally high predicted forces (> 15 eV Å^{-1} on any atom) by $\mathrm{MTP}_{\mathrm{Train}}^{\mathrm{extra\text{-}2}}$ (see ***test and compare multiple MTPs*** in Supporting Information). Since the configurations in this dataset were far from equilibrium, predicted energies were expected to be high (> -4 eV atom^{-1} for typical $\mathrm{Li}_x\mathrm{Al}_y$), but instead the exceptionally low energies were predicted by the MLIPs contradicting to the physical nature of far-from-equilibrium configurations and indicating errors of the MLIPs. This dataset was used in the validation step to exclude those MTP models that have exceptionally low predicted energies, which can be viewed as errors and failures in capturing the atomic interactions. These 141 configurations were selected from the snapshots produced by MTP-MD simulations performed using grand canonical ensemble (allowing the change of supercell size and the number of Li atoms) at 600 K. The supercell of LiAl ($Fd\bar{3}m$) has a size of 25.3 × 25.3 × 24.0 Å^3 and contains 576 Al atoms with varying Li atoms (289 to 578). The supercell of the interface between bulk Li and Al has a size of 27.9 × 19.7 × 22.6 Å^3 and 384 Al atoms with varying Li atoms (157 to 260). A random amount of Li atoms was removed (between 1 to 287 Li atoms removed for supercells with 576 Al atoms, and between 124 to 227 Li atoms removed for supercells with 384 Al atoms) or inserted (< 2 Li atoms inserted for supercells with 576 Al atoms) from the supercells in the dataset. During the MD simulations using NPT ensemble, an attempt to either remove or insert (50% probability for either action) a Li atom is tried every 0.1 ps. Then, by fixing the number of Al atoms, the acceptance of each removal/insertion attempt follows the Metropolis algorithm and the probability is determined by the final energies predicted by MTPs with a given constant chemical potential of Li, using the scheme developed in Ref. [43,44]. Many of these generated configurations were intentionally generated to be less reasonable, far-from-equilibrium configurations, as some atoms get as close as within 1 Å, which were given by the erroneous prediction (very low energies) of $\mathrm{MTP}_{\mathrm{Train}}^{\mathrm{extra\text{-}2}}$. These configurations produced by $\mathrm{MTP}_{\mathrm{Train}}^{\mathrm{extra\text{-}2}}$ (see ***test and compare multiple MTPs***, Supporting Information) during MTP-MD simulations using grand canonical ensemble were re-used in the validation process (see ***test and compare multiple MTPs***, Supporting Information) to select MtPs having the minimum amount of erroneously predicted energies (< -10 eV atom^{-1}). For some configurations with atoms very close to each other, large forces were predicted by the $\mathrm{MTP}_{\mathrm{Train}}^{\mathrm{extra\text{-}2}}$ even on atoms that were far away (> 5 Å) from the erroneous pairs. Given The $\mathcal{D}^{\mathrm{irregular}}$ dataset was a validation dataset only used to select the optimized MTPs as described in the section ***Optimizing MTPs and fine-tuning the hyperparameters***, no DFT values are needed.

Training MTPs. The MTP models were trained using the scheme implemented in MAML[45] Python package interfaced with MTP[46,47] as in previous studies (Zuo et al.[48]). A grid search approach was employed to identify the combination of hyperparameters, including the cutoff radius, the choice of the radial basis function sets, and the maximum number of iterations, as in Zuo et al.[48] Two to ten values were tested for each hyperparameter, giving a total of 2488 MTP models trained from $\mathcal{D}^{\mathrm{train}}$ and 1537 MTP models trained from $\mathcal{D}^{\mathrm{re-train}}$.

Optimizing MTPs and fine-tuning the hyperparameters. The processes of evaluating the validation scores, fine-tuning the hyperparameters, and selecting the optimal MTPs with the best scores were as follows. To calculate the validation scores, 10 criteria, consisting of the RMSE energies and forces and the *NAC* of force errors (see ***Identifying migrating atoms and evaluating force errors)***, were used:

(1) The RMSE of energies for training data $\mathcal{D}^{\mathrm{train}}$/$\mathcal{D}^{\mathrm{re-train}}$ (with respect to the training process of $\mathrm{MTP}_{\mathrm{Train}}$ or $\mathrm{MTP}_{\mathrm{Re\text{-}train}}$), $\sigma_{\mathrm{E}}^{\mathrm{train}}$;

(2) The RMSE of energies for validation data $\mathcal{D}^{\mathrm{validation}}$, $\sigma_{\mathrm{E}}^{\mathrm{validation}}$;

(3) The RMSE of forces for $\mathcal{D}^{\mathrm{train}}$/$\mathcal{D}^{\mathrm{re-train}}$, $\sigma_{\mathrm{F}}^{\mathrm{train}}$;

(4) The RMSE of forces for $\mathcal{D}^{\mathrm{validation}}$, $\sigma_{\mathrm{F}}^{\mathrm{validation}}$;

(5) The normalized area of the cumulative distribution function curve (*NAC*) of errors on force magnitudes for $\mathcal{D}^{\mathrm{train}}$/$\mathcal{D}^{\mathrm{re-train}}$, Δ*NAC*($|\delta_{\mathrm{F}}|$, $\mathcal{D}^{\mathrm{train}}$) = 1 - *NAC*($|\delta_{\mathrm{F}}|$, $\mathcal{D}^{\mathrm{train}}$);

(6) The *NAC* of errors on force directions for $\mathcal{D}^{\mathrm{train}}$/$\mathcal{D}^{\mathrm{re-train}}$, Δ*NAC*($\delta_{\theta}$, $\mathcal{D}^{\mathrm{train}}$) = 1 - *NAC*($\delta_{\theta}$, $\mathcal{D}^{\mathrm{train}}$);

(7) The *NAC* of errors on force magnitudes for $\mathcal{D}^{\mathrm{validation}}$, Δ*NAC*($|\delta_{\mathrm{F}}|$, $\mathcal{D}^{\mathrm{validation}}$) = 1 - *NAC*($|\delta_{\mathrm{F}}|$, $\mathcal{D}^{\mathrm{validation}}$);

(8) The *NAC* of errors on force directions for $\mathcal{D}^{\mathrm{valid}}$, Δ*NAC*($\delta_{\theta}$, $\mathcal{D}^{\mathrm{validation}}$) = 1 - *NAC*($\delta_{\theta}$, $\mathcal{D}^{\mathrm{validation}}$);

(9) The number of large forces predicted for all atoms in $\mathcal{D}^{\mathrm{irregular}}$, N_{F};

(10) The number of configurations that the energies predicted were < -10 eV atom^{-1} in $\mathcal{D}^{\mathrm{irregular}}$, N_{E};

The *NAC*s were the normalized area of the cumulative distribution function (CDF) curves as proposed in Ref [25] and were calculated following the same steps as in Ref

[25]. The CDF curves of force errors were generated to show the distributions of errors over a specific error range, and the *NAC* was then calculated as dividing the area under the CDF curve by the total area (between 0 to 100% on *y*-axis and the specified error range on *x*-axis). Here, the *NAC* of errors on force magnitudes, *NAC*($|\delta_{\mathrm{F}}|$, $\mathcal{D}$), was calculated by using the CDF curves over an error range of 0 to 1.5 eV Å^{-1}, and the *NAC* of errors on force directions, *NAC*(δ_{θ}, $\mathcal{D}$), was calculated by using the CDF curves over an error range of 0 to 180 degrees.

The value of each criterion, *c*, was then standardized by subtracting its minimum value min(*c*) and dividing by its standard deviation std(*c*),

$$v = [c - \min(c)]/\mathrm{std}(c). \quad \text{Eq. (1)}$$

The total validation score was calculated as

$$V = \sum_{v}(w_v v)^2, \quad \text{Eq. (2)}$$

where the weights w_v of N_F and N_E were set to 1.5 and w_v of all the other criteria were set to 1. The MTPs with the lowest total validation score *V* were selected as MTP$_{\mathrm{Train}}$ and MTP$_{\mathrm{Re\text{-}train}}$.

Principal component analysis of atomic environments. We adopted the Smooth Overlap of Atomic Positions (SOAP) descriptors to quantify the atomic environments in all datasets. The SOAP descriptors were calculated using the QUIP[49] package and the parameters set as follows (or default values if not specified),

(1) The band limit of spherical harmonics basis function (*l_max*) was set to 12.

(2) The number of radial basis function (*n_max*) was set to 6.

(3) The covariance function type was set to ‘dot_product’ and its zeta was set to 4.

(4) The cutoff radius was set to 6 Å.

(5) The number of sparse points was set to 4000.

We visualized the quantified atomic environments by performing the principal component analysis (PCA) function implemented in *sci-kit learn* package. We first standardized all elements of the SOAP descriptors by subtracting the means and dividing by the standard deviations. Then, we conducted PCA and plotted the 1st and the 2nd principal component (Figure 3 and Supplementary Figure S19).

Identifying migrating atoms and evaluating force errors. The process of identifying migrating atoms to evaluate the force errors of MTPs (section 2.4.3) on migrating atoms follows the same approach in Ref [25]. Since most of the diffusion events are by Li atoms in Li-Al materials systems, Li migrating atoms were selected following a similar approach as identifying migrating atoms in Ref [25]. If the distances of the atom between their 1st and 2nd nearest static sites have a difference below 0.9 Å (approximately 33% of distances between two static sites), the atom is identified as a migrating atom. The errors of force magnitudes and force directions between the MTP predicted atomic forces and the true values given by DFT K4 were then calculated for these migrating atoms, and the error distribution were plotted as the CDF curves in section 2.4.3 (Figure 6). In addition, forces using Γ-centered 1×1×1 **k**-point (K1) were also calculated on migrating atoms as benchmarks for the comparison with MTPs. The *NAC*s of force errors were evaluated based on these CDF curves and were then quantified and used for selecting the MTP models (see ***Optimizing MTPs and fine-tuning the hyperparameters***). The force performance metrics $P(\mathcal{D})$ is then calculated as the product of the *NAC* of force errors on magnitudes, $|\delta_{\mathrm{F}}|$, and the *NAC* of force errors on directions, δ_{θ}, $P(\mathcal{D})$ = $NAC(|\delta_F|, \mathcal{D}) \times NAC(\delta_{\theta}, \mathcal{D})$ as proposed by Ref [25].

Convex hull and formation energies. The formation energies, E_f, of intermediate phases in section 2.4.2 were calculated as the energy differences between the target intermediate phase Li_xAl_y and the reference phases. Specifically, the E_f in Figure 5a is,

$$E_{\mathrm{f}}(\mathrm{Li_xAl_y}) = E(\mathrm{Li_xAl_y}) - \frac{x}{x+y}E(\mathrm{Li}) - \frac{y}{x+y}E(\mathrm{Al}), \qquad \text{Eq. (3)}$$

where $E(Li_xAl_y)$, $E(Li)$, and $E(Al)$ are the energies per atom of bulk-phase Li_xAl_y, Li ($Im\bar{3}m$), and Al ($Fm\bar{3}m$).

In Figure 5b, the formation energy of interphase intermediate phase Li_xAl_y is evaluated using LiAl and $LiAl_3$ as references,

$$E'_{\mathrm{f}}(\mathrm{Li_xAl_y}) = E(\mathrm{Li_xAl_y}) - \frac{3x-y}{x+y}E(\mathrm{LiAl}) - 2\frac{y-x}{x+y}E(\mathrm{LiAl_3}), \qquad \text{Eq. (4)}$$

where $E(LiAl)$ and $E(LiAl_3)$ are the energies per atom of bulk-phase LiAl ($Fd\bar{3}m$) and $LiAl_3$ ($Pm\bar{3}m$).

Error evaluation metrics: rate of ranking error, mean ΔE^{DFT}, and maximum ΔE^{DFT}. For a list of configurations (elemental orderings), their energies, E^{DFT} and E^{MLIP}, were calculated by DFT and by MLIPs, and were then compared. To calculate the evaluation metrics of energy ranking, all pairs of configurations were enumerated. For a pair of A and B configurations was determined to have a ranking error (or mismatch) if $(E_A^{\mathrm{DFT}} - E_B^{\mathrm{DFT}})(E_A^{\mathrm{MLIP}} - E_B^{\mathrm{MLIP}}) \leq 0$. The rate of ranking error was calculated as the fraction of mismatched configuration pairs among all pairs. The difference of DFT energies for this pair was evaluated as $\Delta E^{\mathrm{DFT}} = E_{\mathrm{A}}^{\mathrm{DFT}} - E_{\mathrm{B}}^{\mathrm{DFT}}$. The mean and maximum ΔE^{DFT} are the mean and maximum for all mismatched pairs of configurations.

Acknowledgement. The authors acknowledge the funding support from National Science Foundation Award #2004837 and the computational facilities from the University of Maryland supercomputing resources.

Author contribution. Y.M. supervised the project. Both authors designed the computation and analyses, and Y.L. performed them. Y.L. and Y. M. wrote the manuscript.

Competing interests. The authors declare no competing interests.

Code availability. The computation codes and programs to support the finding of this study is available from the corresponding author on reasonable request.

Data availability. The structural (POSCAR files), energies, and forces data to support the finding of this study, including $\mathcal{D}^{\mathrm{train}}$, $\mathcal{D}^{\mathrm{re-train}}$, $\mathcal{D}^{\mathrm{validation}}$, $\mathcal{D}^{\mathrm{irregular}}$, $\mathcal{D}^{\mathrm{test}}$, $\mathcal{D}^{\mathrm{slabs}}$, $\mathcal{D}^{\mathrm{Vacancy}}$, and $\mathcal{D}^{\mathrm{bcc/fcc}}$ are available from:

https://github.com/mogroupumd/Li-Al_MLIP_datasets

References

[1] G. Esteban-Manzanares, A. Ma, I. Papadimitriou, E. Martínez, J. LLorca, Basal dislocation/precipitate interactions in Mg–Al alloys: an atomistic investigation, Model. Simul. Mater. Sci. Eng. 27 (2019) 075003. https://doi.org/10.1088/1361-651X/ab2de0.

[2] Z. Huang, V. Turlo, X. Wang, F. Chen, Q. Shen, L. Zhang, I.J. Beyerlein, T.J. Rupert, Dislocation-induced Y segregation at basal-prismatic interfaces in Mg, Comput. Mater. Sci. 188 (2021) 110241. https://doi.org/10.1016/j.commatsci.2020.110241.

[3] B.-J. Lee, B.D. Wirth, J.-H. Shim, J. Kwon, S.C. Kwon, J.-H. Hong, Modified embedded-atom method interatomic potential for the Fe−Cu alloy system and cascade simulations on pure Fe and Fe-Cu alloys, Phys. Rev. B. 71 (2005) 184205. https://doi.org/10.1103/PhysRevB.71.184205.

[4] W. Chen, G. Xu, I. Martin-Bragado, Y. Cui, Non-empirical phase equilibria in the Cr–Mo system: A combination of first-principles calculations, cluster expansion and Monte Carlo simulations, Solid State Sci. 41 (2015) 19–24. https://doi.org/10.1016/j.solidstatesciences.2015.01.012.

[5] A.F. Kohan, P.D. Tepesch, G. Ceder, C. Wolverton, Computation of alloy phase diagrams at low temperatures, Comput. Mater. Sci. 9 (1998) 389–396. https://doi.org/10.1016/S0927-0256(97)00168-7.

[6] P. Rowe, V.L. Deringer, P. Gasparotto, G. Csányi, A. Michaelides, An accurate and transferable machine learning potential for carbon, J. Chem. Phys. 153 (2020) 034702. https://doi.org/10.1063/5.0005084.

[7] S. Wang, Y. Liu, Y. Mo, Frustration in Super-Ionic Conductors Unraveled by the

Density of Atomistic States, Angew. Chemie Int. Ed. 135 (2023) e202215544. https://doi.org/10.1002/anie.202215544.

[8] E.A. Wu, S. Banerjee, H. Tang, P.M. Richardson, J.-M. Doux, J. Qi, Z. Zhu, A. Grenier, Y. Li, E. Zhao, G. Deysher, E. Sebti, H. Nguyen, R. Stephens, G. Verbist, K.W. Chapman, R.J. Clément, A. Banerjee, Y.S. Meng, S.P. Ong, A stable cathode-solid electrolyte composite for high-voltage, long-cycle-life solid-state sodium-ion batteries, Nat. Commun. 12 (2021) 1256. https://doi.org/10.1038/s41467-021-21488-7.

[9] J. Qi, S. Banerjee, Y. Zuo, C. Chen, Z. Zhu, M.L. Holekevi Chandrappa, X. Li, S.P. Ong, Bridging the gap between simulated and experimental ionic conductivities in lithium superionic conductors, Mater. Today Phys. 21 (2021) 100463. https://doi.org/10.1016/j.mtphys.2021.100463.

[10] A. Marcolongo, T. Binninger, F. Zipoli, T. Laino, Simulating Diffusion Properties of Solid-State Electrolytes via a Neural Network Potential: Performance and Training Scheme, ChemSystemsChem. 2 (2020) e1900031. https://doi.org/10.1002/syst.201900031.

[11] V.L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D.A. Drabold, S.R. Elliott, Origins of structural and electronic transitions in disordered silicon, Nature. 589 (2021) 59–64. https://doi.org/10.1038/s41586-020-03072-z.

[12] L. Tang, Z.J. Yang, T.Q. Wen, K.M. Ho, M.J. Kramer, C.Z. Wang, Development of interatomic potential for Al–Tb alloys using a deep neural network learning method, Phys. Chem. Chem. Phys. 22 (2020) 18467–18479.

https://doi.org/10.1039/D0CP01689F.

[13] L. Tang, Z.J. Yang, T.Q. Wen, K.M. Ho, M.J. Kramer, C.Z. Wang, Short- and medium-range orders in Al90Tb10 glass and their relation to the structures of competing crystalline phases, Acta Mater. 204 (2021) 116513. https://doi.org/10.1016/j.actamat.2020.116513.

[14] S. Kharabadze, A. Thorn, E.A. Koulakova, A.N. Kolmogorov, Prediction of stable Li-Sn compounds: boosting ab initio searches with neural network potentials, Npj Comput. Mater. 8 (2022) 136. https://doi.org/10.1038/s41524-022-00825-4.

[15] S. Yin, Y. Zuo, A. Abu-Odeh, H. Zheng, X.-G. Li, J. Ding, S.P. Ong, M. Asta, R.O. Ritchie, Atomistic simulations of dislocation mobility in refractory high-entropy alloys and the effect of chemical short-range order, Nat. Commun. 12 (2021) 4873. https://doi.org/10.1038/s41467-021-25134-0.

[16] K. Gubaev, E. V. Podryabinkin, G.L.W. Hart, A. V. Shapeev, Accelerating high-throughput searches for new alloys with active learning of interatomic potentials, Comput. Mater. Sci. 156 (2019) 148–156. https://doi.org/10.1016/j.commatsci.2018.09.031.

[17] T. Zubatiuk, O. Isayev, Development of Multimodal Machine Learning Potentials: Toward a Physics-Aware Artificial Intelligence, Acc. Chem. Res. 54 (2021) 1575–1585. https://doi.org/10.1021/acs.accounts.0c00868.

[18] C.W. Rosenbrock, K. Gubaev, A. V. Shapeev, L.B. Pártay, N. Bernstein, G. Csányi, G.L.W. Hart, Machine-learned interatomic potentials for alloys and alloy phase diagrams, Npj Comput. Mater. 7 (2021) 24. https://doi.org/10.1038/s41524-020-00477-2.

[19] G.L.W. Hart, T. Mueller, C. Toher, S. Curtarolo, Machine learning for alloys, Nat. Rev. Mater. 6 (2021) 730–755. https://doi.org/10.1038/s41578-021-00340-w.

[20] A.H. Nguyen, C.W. Rosenbrock, C.S. Reese, G.L.W. Hart, Robustness of the cluster expansion: Assessing the roles of relaxation and numerical error, Phys. Rev. B. 96 (2017) 014107. https://doi.org/10.1103/PhysRevB.96.014107.

[21] D. Montes de Oca Zapiain, M.A. Wood, N. Lubbers, C.Z. Pereyra, A.P. Thompson, D. Perez, Training data selection for accuracy and transferability of interatomic potentials, Npj Comput. Mater. 8 (2022) 189. https://doi.org/10.1038/s41524-022-00872-x.

[22] R. Batra, S. Sankaranarayanan, Machine learning for multi-fidelity scale bridging and dynamical simulations of materials, J. Phys. Mater. 3 (2020) 031002. https://doi.org/10.1088/2515-7639/ab8c2d.

[23] L. Zhang, D.-Y. Lin, H. Wang, R. Car, W. E, Active learning of uniformly accurate interatomic potentials for materials simulation, Phys. Rev. Mater. 3 (2019) 023804. https://doi.org/10.1103/PhysRevMaterials.3.023804.

[24] A. Koneru, H. Chan, S. Manna, T.D. Loeffler, D. Dhabal, A.A. Bertolazzo, V. Molinero, S.K.R.S. Sankaranarayanan, Multi-reward reinforcement learning based development of inter-atomic potential models for silica, Npj Comput. Mater. 9 (2023) 125. https://doi.org/10.1038/s41524-023-01074-9.

[25] Y. Liu, X. He, Y. Mo, Discrepancies and error evaluation metrics for machine learning interatomic potentials, Npj Comput. Mater. 9 (2023) 174. https://doi.org/10.1038/s41524-023-01123-3.

[26] Y. Luo, J.A. Meziere, G.D. Samolyuk, G.L.W. Hart, M.R. Daymond, L.K. Béland, A

Set of Moment Tensor Potentials for Zirconium with Increasing Complexity, J. Chem. Theory Comput. 19 (2023) 6848–6856. https://doi.org/10.1021/acs.jctc.3c00488.

[27] E. Longato, M. Vettoretti, B. Di Camillo, A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models, J. Biomed. Inform. 108 (2020) 103496. https://doi.org/10.1016/j.jbi.2020.103496.

[28] V.L. Deringer, A.P. Bartók, N. Bernstein, D.M. Wilkins, M. Ceriotti, G. Csányi, Gaussian Process Regression for Materials and Molecules, Chem. Rev. 121 (2021) 10073–10141. https://doi.org/10.1021/acs.chemrev.1c00022.

[29] N. Bernstein, G. Csányi, V.L. Deringer, De novo exploration and self-guided learning of potential-energy surfaces, Npj Comput. Mater. 5 (2019) 1–11. https://doi.org/10.1038/s41524-019-0236-6.

[30] A. Seko, Machine learning potentials for multicomponent systems: The Ti-Al binary system, Phys. Rev. B. 102 (2020) 174104. https://doi.org/10.1103/PhysRevB.102.174104.

[31] P.A. Santos-Florez, S.-C. Dai, Y. Yao, H. Yanxon, L. Li, Y.-J. Wang, Q. Zhu, X.-X. Yu, Short-range order and its impacts on the BCC MoNbTaW multi-principal element alloy by the machine-learning potential, Acta Mater. 255 (2023) 119041. https://doi.org/10.1016/j.actamat.2023.119041.

[32] R.E. Ryltsev, N.M. Chtchelkatchev, Deep machine learning potentials for multicomponent metallic melts: Development, predictability and compositional transferability, J. Mol. Liq. 349 (2022) 118181. https://doi.org/10.1016/j.molliq.2021.118181.

[33] J. Qi, Z.H. Aitken, Q. Pei, A.M.Z. Tan, Y. Zuo, M.H. Jhon, S.S. Quek, T. Wen, Z. Wu, S.P. Ong, Machine Learning Moment Tensor Potential for Modelling Dislocation and Fracture in L1$_0$-TiAl and D0$_{19}$-Ti$_3$Al Alloys, Http://Arxiv.Org/Abs/2305.11825. (2023) 1–24. http://arxiv.org/abs/2305.11825.

[34] S. Divilov, H. Eckert, D. Hicks, C. Oses, C. Toher, R. Friedrich, M. Esters, M.J. Mehl, A.C. Zettel, Y. Lederer, E. Zurek, J. Maria, D.W. Brenner, X. Campilongo, S. Filipović, W.G. Fahrenholtz, C.J. Ryan, C.M. DeSalle, R.J. Crealese, D.E. Wolfe, A. Calzolari, S. Curtarolo, Disordered enthalpy–entropy descriptor for high-entropy ceramics discovery, Nature. 625 (2024) 66–73. https://doi.org/10.1038/s41586-023-06786-y.

[35] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B - Condens. Matter Mater. Phys. 54 (1996) 11169–11186. https://doi.org/10.1103/PhysRevB.54.11169.

[36] J.P. Perdew, M. Ernzerhof, K. Burke, Rationale for mixing exact exchange with density functional approximations, J. Chem. Phys. 105 (1996) 9982–9985. https://doi.org/10.1063/1.472933.

[37] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. 68 (2013) 314–319. https://doi.org/10.1016/j.commatsci.2012.10.028.

[38] A. Jain, G. Hautier, C.J. Moore, S. Ping Ong, C.C. Fischer, T. Mueller, K.A. Persson, G. Ceder, A high-throughput infrastructure for density functional theory calculations, Comput. Mater. Sci. 50 (2011) 2295–2310.

https://doi.org/10.1016/j.commatsci.2011.02.023.

[39] Y. Liu, S. Wang, A.M. Nolan, C. Ling, Y. Mo, Tailoring the Cation Lattice for Chloride Lithium-Ion Conductors, Adv. Energy Mater. 10 (2020) 2002356. https://doi.org/10.1002/aenm.202002356.

[40] Y. Liu, X. He, Y. Mo, Discrepancies and the Error Evaluation Metrics for Machine Learning Interatomic Potentials, (2023). https://doi.org/https://doi.org/10.48550/arXiv.2306.11639.

[41] A. Zur, T.C. McGill, Lattice match: An application to heteroepitaxy, J. Appl. Phys. 55 (1984) 378–386. https://doi.org/10.1063/1.333084.

[42] X. He, Q. Bai, Y. Liu, A.M. Nolan, C. Ling, Y. Mo, Crystal Structural Framework of Lithium Super-Ionic Conductors, Adv. Energy Mater. 9 (2019) 1–12. https://doi.org/10.1002/aenm.201902078.

[43] T.P. Senftle, M.J. Janik, A.C.T. van Duin, A ReaxFF Investigation of Hydride Formation in Palladium Nanoclusters via Monte Carlo and Molecular Dynamics Simulations, J. Phys. Chem. C. 118 (2014) 4967–4981. https://doi.org/10.1021/jp411015a.

[44] T.P. Senftle, R.J. Meyer, M.J. Janik, A.C.T. van Duin, Development of a ReaxFF potential for Pd/O and application to palladium oxide formation, J. Chem. Phys. 139 (2013) 044109. https://doi.org/10.1063/1.4815820.

[45] C. Chen, Y. Zuo, W. Ye, Q. Ji, S.P. Ong, Maml - materials machine learning package, GitHub Repos. (2020). https://github.com/materialsvirtuallab/maml.

[46] A. V Shapeev, Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials, Multiscale Model. Simul. 14 (2016) 1153–1173.

https://doi.org/10.1137/15M1054183.

[47] E. V. Podryabinkin, A. V. Shapeev, Active learning of linearly parametrized interatomic potentials, Comput. Mater. Sci. 140 (2017) 171–180. https://doi.org/10.1016/j.commatsci.2017.08.031.

[48] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A.P. Thompson, M.A. Wood, S.P. Ong, Performance and Cost Assessment of Machine Learning Interatomic Potentials, J. Phys. Chem. A. 124 (2020) 731–745. https://doi.org/10.1021/acs.jpca.9b08723.

[49] A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons, Phys. Rev. Lett. 104 (2010) 136403. https://doi.org/10.1103/PhysRevLett.104.136403.