npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01500-6

Systematic softening in universal machine learning interatomic potentials

Bowen Deng [1,2], Yunyeong Choi [1,2], Peichen Zhong [1,2], Janosh Riebesell [3], Shashwat Anand [2], Zhuohan Li [2], KyuJung Jun [1,2], Kristin A. Persson [1,2] & Gerbrand Ceder [1,2]

Machine learning interatomic potentials (MLIPs) have introduced a new paradigm for atomic simulations. Recent advancements have led to universal MLIPs (uMLIPs) that are pre-trained on diverse datasets, providing opportunities for universal force fields and foundational machine learning models. However, their performance in extrapolating to out-of-distribution complex atomic environments remains unclear. In this study, we highlight a consistent potential energy surface (PES) softening effect in three uMLIPs: M3GNet, CHGNet, and MACE-MP-0, which is characterized by energy and force underprediction in atomic-modeling benchmarks including surfaces, defects, solid-solution energetics, ion migration barriers, phonon vibration modes, and general high-energy states. The PES softening behavior originates primarily from the systematically underpredicted PES curvature, which derives from the biased sampling of near-equilibrium atomic arrangements in uMLIP pre-training datasets. Our findings suggest that a considerable fraction of uMLIP errors are highly systematic, and can therefore be efficiently corrected. We argue for the importance of a comprehensive materials dataset with improved PES sampling for next-generation foundational MLIPs.

Artificial intelligence (AI) is increasingly shifting the paradigm of scientific discovery to accelerate research and solve real-world scientific challenges[1]. While ab-initio quantum mechanical simulation methods, such as density functional theory (DFT), offer the theoretical foundation to investigate material and chemical science problems at the atomic scale, their computational demands limit their applicability in both spatial and temporal scales. Recent advancements in machine learning interatomic potentials (MLIPs)[2,3] have enabled the opportunity to scale up quantum mechanical methods to million atoms simulations such as water, copper[4], and biomolecules[5].

Alongside improvements in atomic environment descriptors and graph neural networks that enhance the expressivity of MLIP models[3,6], universal machine learning interatomic potentials (uMLIPs) have demonstrated another avenue by taking advantage of pre-training on large and comprehensive material datasets[7–13]. These uMLIPs enable out-of-box atomic modeling covering the entire periodic table as well as providing robust machine-learning foundations for fine-tuning downstream tasks. While uMLIPs hold considerable promise, a critical challenge lies in their ability to reliably generalize to complex and diverse chemical environments, particularly those that deviate significantly from the pre-training data distribution. Several recent benchmark efforts have tested the uMLIPs' ability to identify stable materials[14], surface energies[15], lattice relaxations and vibrational properties[16], etc. A systematic understanding of the ability of uMLIPs to extrapolate to common atomic-modeling tasks, especially those with atomic environments that are out of distribution (OOD), remains an open question with implications for their real-world applicability in material discovery and design.

In this work, we systematically investigate the extrapolative capabilities of three foundational uMLIPs – M3GNet[7], CHGNet[8], and MACE-MP-0[10] (hereafter referred to as MACE) – across a diverse suite of material modeling tasks, including surface energies, defect energies, solid-solution energetics, phonon vibrational modes, and ion migration barriers. Across all benchmark tests for all uMLIP models, our analysis shows consistent underpredictions of energies and forces. To quantify and explain thse underpredictions, we investigate the behavior of uMLIPs in high-energy transition states and reveal a systematic potential energy surface (PES) softening behavior in the uMLIPs as illustrated in Fig. 1. We attribute the PES softening issue to the combination of the biased sampling of near-ground-state configurations in the uMLIP pre-training datasets[17], which primarily comprise DFT ionic relaxation trajectories near PES local energy minima. The uMLIPs trained predominantly on small energy and force labels suffer from distribution shifts and experience increased but systematic prediction errors in high-energy PES regions which are important for the

[1]Department of Materials Science and Engineering, University of California, Berkeley, CA, USA. [2]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [3]Cavendish Laboratory, University of Cambridge, Cambridge, UK. e-mail: gceder@berkeley.edu

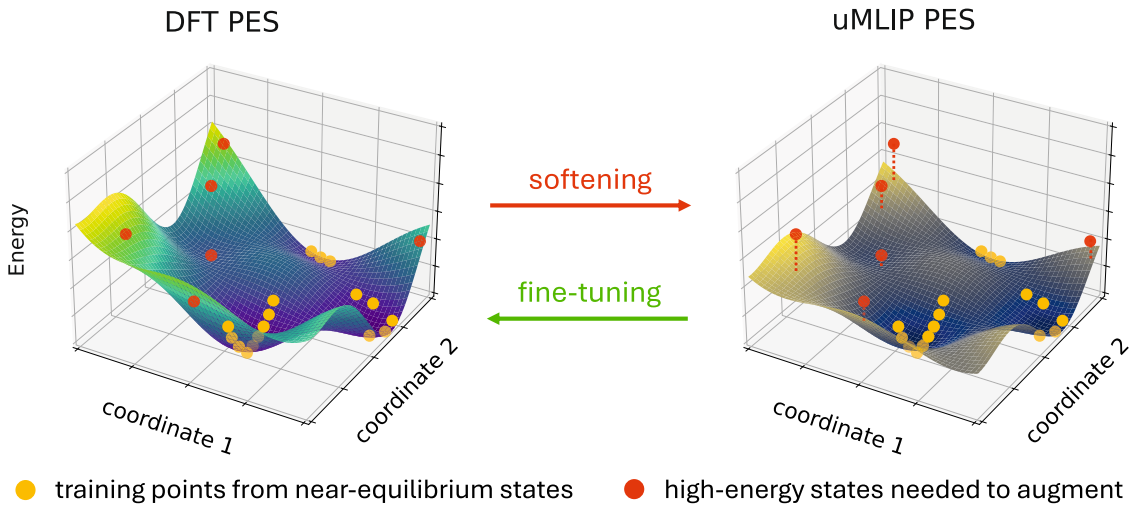

Fig. 1 | Potential energy surface softening in uMLIPs. Left: schematic representation of the potential energy surface (PES) described in density functional theory (DFT), with two arbitrary coordinate axes. Right: PES described by universal machine learning interatomic potentials (uMLIPs), which well describes the PES regions sampled by near-equilibrium states in the pre-training dataset (orange), but experience larger errors in high-energy regions (red) with underprediction of energies and forces. The softening behavior is largely systematic in a given chemical space and, can therefore be efficiently fixed locally with a small amount of data augmentation. Using a linear correction, we demonstrate the data efficiency of uMLIP fine-tuning.

kinetics of rare events, such as ion migrations, and for the energy of defects with undercoordinated atoms, such as vacancies and surfaces.

We demonstrate that this systematic PES softening can be effectively mitigated by fine-tuning with a minimal amount of data points. We find that a simple linear correction derived from a single DFT reference label is sufficient to remove much of the PES softening issue in a specific chemical system of interest, significantly enhancing the performance and robustness of uMLIPs for a given application. We rationalize this observation by arguing that a considerable amount of prediction errors in pretrained uMLIPs are highly systematic, and therefore can be efficiently corrected by modifying a limited fraction of the model parameters with only a small amount of data augmentation. Our work provides a theoretical foundation for the widely observed data-efficient performance boosts achieved by fine-tuning uMLIPs and highlights the advantage of atomic modeling with large and comprehensive foundational AI models.

Results

Machine learning interatomic potentials framework

MLIPs approximate the total energy of a system as a sum of atomic contributions, each dependent on the positions and chemical identities of the atoms in their local environment:

$$E = \sum_{i}^{n} \phi(\{\vec{r}_j\}_i, \{C_j\}_i), \quad \vec{f}_i = -\frac{\partial E}{\partial \vec{r}_i} \tag{1}$$

ϕ is a learnable function that maps the set of position vectors $\{\vec{r}_j\}_i$ and chemical species $\{C_j\}_i$ of the neighboring atoms j to the energy contribution of atom i. The force $\vec{f}_i$ acting on each atom is calculated as the derivative of the total energy with respect to its position. In the training process, the parameters of the MLIP model are optimized to minimize the discrepancy between the predicted energies and forces and the corresponding reference values from the DFT labels.

The design of the atomic environment descriptor function ϕ is crucial to developing accurate and efficient MLIPs. To capture the essential physics and chemistry of the system, ϕ should be informative and satisfy proper translational and rotational symmetries. This is typically achieved through the use of graph representations[18], high-order interactions[6,7], the preservation of SE(3)/E(3)-equivariance using tensor products based on spherical harmonics[3,10], Fourier basis[19], or Cartesian-coordinates-based atomic density expansion[20]. Additionally, the incorporation of chemical information, such as charge[21] or atomic magnetic moment[8], has been shown to enhance the predictive power of MLIPs.

In addition, recent efforts have been made to pre-train MLIPs on large open-sourced materials datasets such as the Materials Project[17], which primarily consists of DFT ionic relaxation trajectories of various compounds and elements across the periodic table. While initial benchmarks have shown the promising applicability of universal MLIPs in predicting bulk materials energetics[14,16], their performance and limitations in OOD atomic configurations require more benchmarking as the energy of these configurations is often directly relevant for practical materials behavior. The following sections present a systematic assessment of the uMLIPs' ability to extrapolate to low-symmetry OOD atomic configurations that are crucial for atomic-modeling tasks.

Surface energies

Surface energies play an important role in determining the stability and morphology of materials, especially at the nano-scale where the surface-to-volume ratio is significant. Accurate prediction of surface energies is crucial for various applications such as catalysis[22], corrosion[23], adhesion[24], nucleation[25], and thin film growth[26]. In this section, we assess uMLIP's performance in predicting surface energies, which are calculated as

$$\gamma_{\text{surface}} = \frac{E_{\text{slab}} - E_{\text{bulk}}}{2A_{\text{slab}}}, \tag{2}$$

where $E_{\text{slab}}/E_{\text{bulk}}$ are the relaxed energies of the slab/bulk structures that can be obtained independently using DFT or MLIP methods in a large supercell approach. A_{slab} denotes the surface area of the slab.

he energies of 147 surfaces with multiple Miller indices of 29 elements and binary compounds are evaluated, including Si, Cu, Al_2O_3, LiF, ZnS, etc. The DFT and uMLIP calculation details are listed in the Methods section and Supplementary Table 1 lists the full set of elements and compounds with their corresponding prediction errors. Figure 2a shows the uMLIP surface energies versus the DFT surface energies for the three uMLIPs tested, where MAE stands for the model's mean absolute error. MACE exhibits relatively better performance compared to CHGNet and M3GNet, achieving a MAE of 0.032 eV/Å^2. All three uMLIPs consistently underestimate the surface energies compared to DFT, except for a few predictions made by MACE and

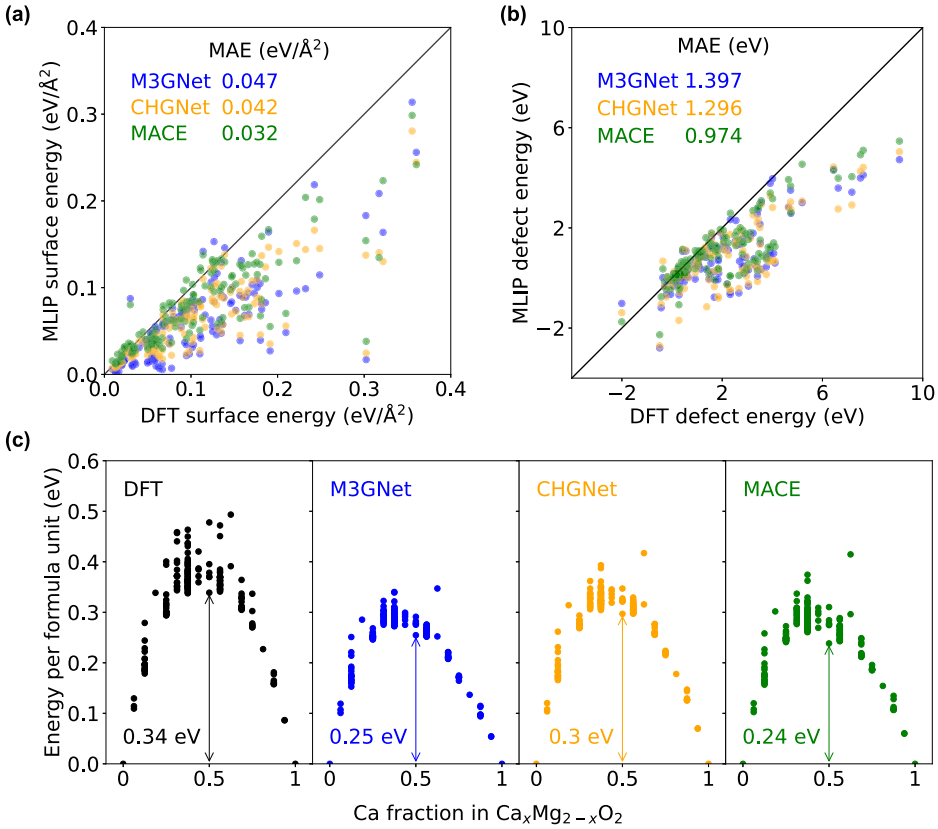

Fig. 2 | uMLIP performance on surfaces, defects, and solid solutions. a Comparison of DFT surface energies and MLIP surface energies, evaluated on 147 surfaces from 29 chemical systems. **b** Comparison of DFT defect energies and MLIP defect energies, evaluated on 134 point defects from 32 chemical systems. **c** Formation energies in $Ca_xMg_{2-x}O_2$ solid solution from DFT and uMLIPs. Each point corresponds to the energy of a specific Ca-Mg cation arrangement at a given Ca fraction. The distributions of energies are collectively underestimated, which would lead to an under-prediction of the miscibility gap temperature in uMLIPs compared to DFT.

M3GNet. The trend in our result is consistent with the recent evaluation of Focassio et al.[15] on the surface energies of elemental crystals.

Defect energies

We also analyze the accuracy of uMLIPs in calculating point defect energies, which is crucial for understanding a material's vacancy formation[27], dopabilities[28], mechanical properties[29], and ionic mobilities[30]. Specifically, we perform benchmarks for point defects including vacancies, interstitials, and anti-site defects. In metallic systems, the point defect energy can be calculated from the energy of a defect structure referenced to the corresponding perfect structure and the external chemical potential of the species added or removed

$$E_i^{\text{point defect}} = E_i^{\text{defect}} - E^{\text{bulk}} - \Sigma \mu_i \Delta N_i, \quad (3)$$

where μ_i is the chemical potential of the species i forming the defect and ΔN_i is the number of atoms of i added (+ 1) or removed (− 1) at the defect. To avoid additional errors in the defect energy introduced by the equilibrium chemical potentials determined from the phase diagram, we used the energy of the pure elemental phases μ_i for this benchmark section. This choice does not affect the benchmark, but only shifts the value of the point defect energy.

Figure 2b presents a comparison between uMLIP and DFT defect energies for 129 point defects across 32 chemical systems, including AlNi, $CaSn_3$, Cu_3Au, $NaPb_3$, $NaAg_4$, etc. Calculation details are listed in the Methods section and the complete list of materials is provided in Supplementary Table 2. Interestingly, the uMLIP calculated defect energies are mostly underestimated, similar to the trend observed in the surface energies in Fig. 2a.

Solid-solution energetics

Thermodynamic modeling of solubility in solid-state systems such as metallic alloys[31] and high-entropy ceramics[32] requires accurate energetics to capture the dependence of the energy on substitutional arrangements[33,34]. This dependence, relative to k_BT, determines the temperature scale at which mixing or order/disorder transitions occur[35]. In this section, we use the mixing of Ca^{2+} and Mg^{2+} in the $Ca_xMg_{2-x}O_2$ rocksalt as an example to examine the ability of uMLIPs to predict the behavior of the solid solution. The end members of the system, MgO and CaO are both rocksalts and the phase diagram has been previously studied both experimentally[36] and computationally[37].

We explore different possible Ca-Mg cation arrangements in the rocksalt at various CaO-MgO ratios and evaluate the corresponding energies (see Methods). These 0K formation energies are shown in Fig. 2c, where each point corresponds to the energy of a specific Ca-Mg cation arrangement at a given Ca fraction. The predicted formation energies from all uMLIPs are positive, consistent with the low T immiscibility of CaO and MgO[37]. We observe a systematic underprediction of the mixing energies and the energy difference between the uMLIPs and DFT at a specific Ca fraction. Among the uMLIPs, CHGNet's predictions closely approximate those of DFT, followed by those of M3GNet and MACE. We note that an underprediction of the formation energy would lead to an underestimation of the solubilization temperature in phase diagram calculations and an overestimation of the solubility limits at a given temperature[35].

Ion migration barriers

The migration barrier for an ion to move through a crystal structure forms the basis for evaluating the diffusion constants in a material and as such is critical to understand its functional or processing behavior. An accurate description of ion mobility is directly relevant in various applications, such as lithium-ion conductors for battery technologies[38], and proton conductors for fuel cells[39], etc. Because the migration barrier is determined by the extrapolation of the energy along the path between two stable sites, it is by definition also a poorly sampled configuration when uMLIPs are only fitted to local equilibrium configurations.

We employ uMLIPs and DFT to conduct a comprehensive assessment of 470 Mg-ion migration pathways in 110 distinct structures including

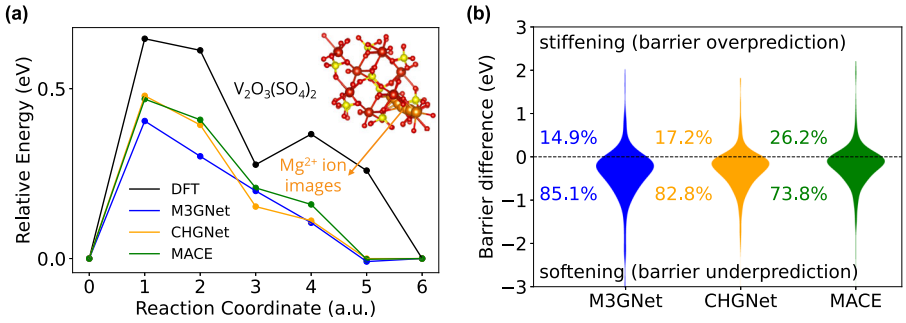

Fig. 3 | Underpredicted ion migration barriers in DFT and uMLIPs. a An example of a Mg-ion migration path in $V_2O_3(SO_4)_2$ (mp-28207) with 5 intermediate images. The energies of initial and final images are referenced to 0, and the kinetically resolved activation (KRA) migration barrier is defined as the highest energy along the path. **b** The distribution of 477 energy barrier differences between uMLIPs and DFT, showing uMLIPs' tendency to underestimate the ion migration barriers.

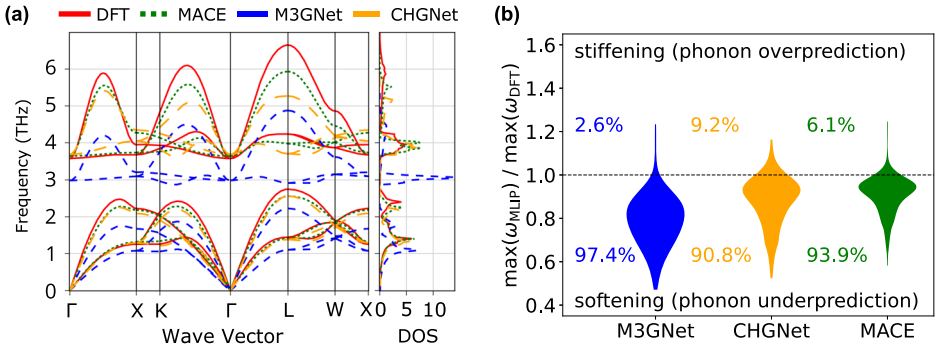

Fig. 4 | Softened phonon vibration modes in uMLIPs. a the phonon dispersion relation and density of states (DOS) of CsF(mp-1784) calculated with DFT and uMLIPs. Systematic underpredictions of phonon vibration frequencies are observed with all uMLIPs. **b** Distribution of ratios between uMLIP maximum frequency to DFT maximum frequency for 229 different compounds.

oxides, halides, and sulfides[40]. For all ion migration paths, we generate an initial guess of the minimum energy pathway based on the DFT charge density[41] and subsequently evaluate it with the approximate nudged elastic band (ApproxNEB) method[42](see method section). ApproxNEB is different from regular NEB in that it does not perform a relaxation of the pathway but solely evaluates the energy along the predefined trajectory[42]. Figure 3a presents the energy landscape of one Mg ion migration path in $V_2O_3(SO_4)_2$ (Materials Project ID mp-28207), where the energies of the initial and final images have been referenced to 0. The kinetically resolved activation (KRA) migration barrier is defined as the highest energy along the reaction coordinate after the reference, which presents the relevant migration barrier in kinetic theories[43]. While all three uMLIPs are shown to capture the overall shape of the DFT energy along the path, we observe systematic energy underpredictions of uMLIPs resulting in underpredictions of KRA migration barriers. MACE achieves the best performance with a 0.34 eV MAE against DFT, followed by CHGNet (0.39 eV) and M3GNet (0.49 eV). The parity plot of uMLIP barriers *vs*. DFT barriers is provided in Supplementary Fig. 1 and shows that the majority of uMLIP barriers are underpredicted, similar to the result of the surface and defect benchmarks. Figure 3b presents the distribution of the energy barrier difference between uMLIPs and DFT, from which we observe that all three uMLIPs show negative shifts in barrier predictions.

Phonon Properties

Accurate descriptions of vibrational properties and phonon spectra are crucial for understanding a wide range of material characteristics, such as thermodynamic[44], mechanical[45], and thermal transport properties[46]. Predicting phonon frequencies represents a stringent test of the MLIPs' ability to capture the subtle energy and force landscape around equilibrium configurations. Compared to the previous modeling benchmarks, phonon properties assess the uMLIPs' accuracies in PES regions that are closer to training distribution. In this section, we benchmark the uMLIPs' performance on phonon frequencies by applying the finite displacement method[47] to calculate harmonic phonons.

Figure 4a shows an example of uMLIP and DFT calculated phonon frequency on CsF (Materials Project ID mp-1784), where the solid red lines represent DFT phonon frequencies without non-analytical corrections (NAC) taken from the PhononDB[48,49] and the dashed lines show uMLIP phonon frequencies. While the overall shapes of the phonon bands are generally well-captured by the uMLIPs, a systematic reduction of the vibrational frequencies (i.e., the frequency magnitude difference of the branches at a given wave vector) is observed across all models compared to the DFT reference, particularly for the optical modes predicted by M3GNet (blue dashed line). The reduced vibrational frequency is an indication that the forces described by uMLIPs are systematically lower than the DFT values.

To quantify this softening behavior, we evaluate the ratio between the maximum phonon frequencies predicted by the uMLIPs and the corresponding DFT value for a diverse set of 229 materials (see Supplementary Table 3) from the PhononDB[48,49]. The distribution of these ratios is shown in Fig. 4b, which demonstrates that the majority (>90%) of materials are found to be softened in uMLIPs compared to DFT, with the phonon frequency underpredicted. The result suggests that both the energy and force described by uMLIPs are softened for almost all chemical systems.

PES softening scale for high-energy states

By definition, a machine learning model with only random errors should have its prediction error distribution centered at 0. However, all three uMLIPs are shown to not satisfy such criterion in both OOD atomic configurations and PES regions that are closer to equilibrium like phonons. These consistent underpredictions can come from two possible causes: (1) Systematic underpredictions of energies and forces that soften the PES. (2) Ionic relaxations that optimize the output energy towards lower values due to modified PES minima created by random errors. While the latter cause arises from random errors that are challenging to eliminate, the former cause arises from systematic errors that can be easily quantified by separating out relaxations and directly benchmarking uMLIPs against DFT at fixed checkpoints in the PES.

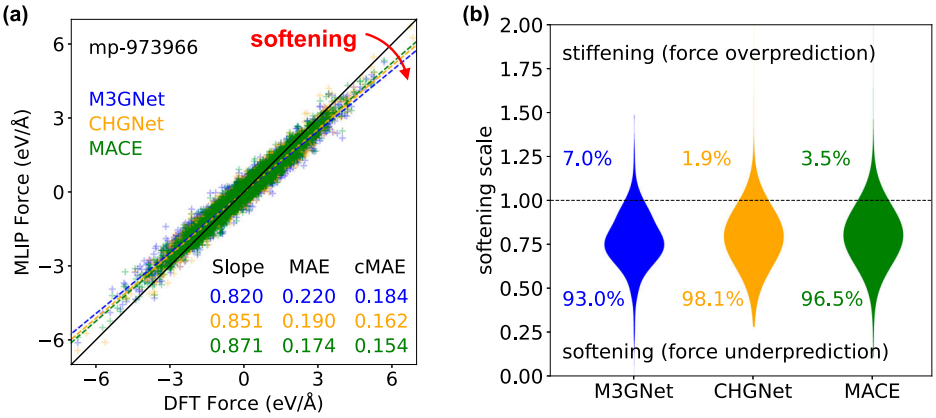

Fig. 5 | The PES softening scale from shifted force predictions. a uMLIP forces *vs.* DFT forces in high-energy states, sampled from high-temperature MDs of a Materials Project structure (`mp-973966`). Systematic softening of PES is indicated by the tilted distribution of forces from the diagonal. The softening scale is defined as the slope of the distribution, where softening is indicated by slope < 1. cMAE stands for corrected mean absolute error, which is the MAE if the softening scale is corrected to 1, equivalent to having the force distribution rotated back to diagonal. **b** Distribution of softening scales of 1000 compounds sampled from the WBM dataset, showing the PES softening behavior is universal across various chemical systems.

To quantify the extent of systematic softening in uMLIP PES, we propose the softening scale parameter, which is calculated as the linear fitted slope of uMLIPs *vs.* DFT forces in a material. As an example, Fig. 5a shows an exemplary parity plot of uMLIPs *vs.* DFT forces from sampled high-energy OOD atomic configurations of $Li_6Zn_2In_2(IO_3)_{16}$ (derived from Materials Project ID `mp-973966`). These OOD atomic configurations are sampled away from the energy minimum in the PES, by applying high-temperature molecular dynamics (MD) simulations(see method section). The corresponding forces of each sampled state are subsequently evaluated using static calculations with uMLIPs and DFT.

The systematic PES softening effect shows up in Fig. 5a by the clockwise tilting of the distribution away from the diagonal. The slope of this distribution, extractable by linear regression, can be defined to be the PES *softening scale*. In Fig. 5a, we provide the fitted slopes and force MAEs of the three uMLIPs. When the softening scale is 1, the MLIP's force distribution aligns with the diagonal, indicating that the curvature of the MLIP-PES systematically agrees with DFT with only random errors present. A softening scale smaller than 1.00 indicates a systematic underprediction of energy and forces that leads to an overall smoother PES curvature as illustrated in Fig. 1.

To investigate how broadly across chemistry the PES softening occurs, we collected 1000 different compounds from the WBM materials dataset by Wang et al.[50], which was generated by elemental substitution of Materials Project compounds and therefore contains only crystalline structures that are not included in the pre-training dataset of the three uMLIPs. For each of these compounds, 10 high-energy states away from the PES energy minimum are sampled with a 1000K MD simulation, and the softening scale is extracted from a linear fit with uMLIP and DFT forces. Figure 5b presents the distribution of the PES softening scale for these 1000 WBM compounds, and shows that for the majority (>90%) of the compounds, the softening scale is smaller than 1 for all 3 uMLIPs we have tested. This result indicates the systematic softening behavior is universal across all chemical systems in current uMLIP models.

Data-efficient fine-tuning

Within a local PES region of a specific chemical system, the softening issue appears as a tilted distribution of forces in the parity plot as shown in Fig. 5a for `mp-973966`. Intuitively, one can rotate the distribution back to the diagonal to reset the softening scale to 1 hereby reducing the prediction error. In this scenario, we define cMAE as the linearly corrected mean absolute error if the uMLIP force distributions were rotated back to align with the diagonal. As shown in Fig. 5a, the cMAEs are considerably reduced from the original MAE from 0.220/0.190/0.176 eV/Å to 0.184/0.162/0.155 eV/Å for M3GNet/CHGNet/MACE, respectively. This observation suggests that a considerable fraction of force errors from uMLIP are likely to be systematic and can be easily corrected locally to reduce force errors.

Mathematically, rotating the force distribution is equivalent to multiplying every force value by a scalar, which can be realized by multiplying the MLIP energy by a scalar term

$$\begin{aligned} E^{\text{corr}} &= c * \mathbf{MLIP}(\{\vec{r}_i\}, \{C_i\}), \\ f_i^{\text{corr}} &= -\frac{\partial E^{\text{corr}}}{\partial \vec{r}_i} = c * f_i. \end{aligned} \tag{4}$$

It is noted that the above formulation is equivalent to fine-tuning a MLIP by fixing all model weights except a scalar linear layer, which essentially modifies only the scalar parameter c in Equation (4). Since only a scalar parameter requires modification, only one single label (1 force component) is needed for the training. Since the crystal cell typically consists of multiple atoms, with each atom carrying three force components, a single training structure already contains enough information for the proposed linear correction. In the left part of Fig. 6a, we show the result when pre-trained CHGNet is fine-tuned with an added hypothetical scalar linear layer (see Methods), trained on only a ***single*** high-energy configuration of `mp-973966`. The test forces, which originate from the same set of atomic arrangements in Fig. 5a, are labeled in orange and the training forces from the single additional configuration are labeled in red. The linear corrected CHGNet exhibits a softening scale of 0.965 and a force MAE of 0.166 eV/Å, improved from 0.859 and 0.190 eV/Å in the pre-trained CHGNet as shown in Fig. 5. The estimated cMAE is 0.162 eV/Å when the softening is corrected to 1, which is close to the force MAE of 0.166 eV/Å that is achieved by fine-tuning the scalar linear layer. Hence, a linear correction with one high-energy OOD configuration indeed operates as a rotation of the force distribution back to the diagonal, substantially eliminating the systematic softening error and considerably reducing the force MAE.

We propose that the cMAE derived from the linear correction serves as an approximate lower bound for the expected error reduction from fine-tuning uMLIPs. In Supplementary Fig. 5, we show that the errors in materials modeling tasks, such as surface calculation, can be similarly reduced after a linear correction with one label. Consequently, the proposed linear correction serves as a baseline for fine-tuning error reduction in uMLIPs. In practice, a typical fine-tuning process involves hundreds and thousands of structure labels that can further reduce the MAE of the model. We tested fine-tuning the pretrained CHGNet by optimizing all model parameters with 10 training structures, and the resulting force parity plot is shown on the right of Fig. 6a. Compared to the linear correction with only one configuration, the right panel in Fig. 6a shows that a very small dataset of 10 training structures further reduces the MAE to 0.125 eV/Å, which proves the linear-corrected cMAE approximates a safe lower bound. By statistically evaluating the distribution of force MAEs and cMAEs for the 1000 WBM structures, we present their fine-tuning error-reduction lower-bounds in Fig. 6b. From the observed distribution, considerable error reduction (~15%) can be adequately achieved with a simple linear correction.

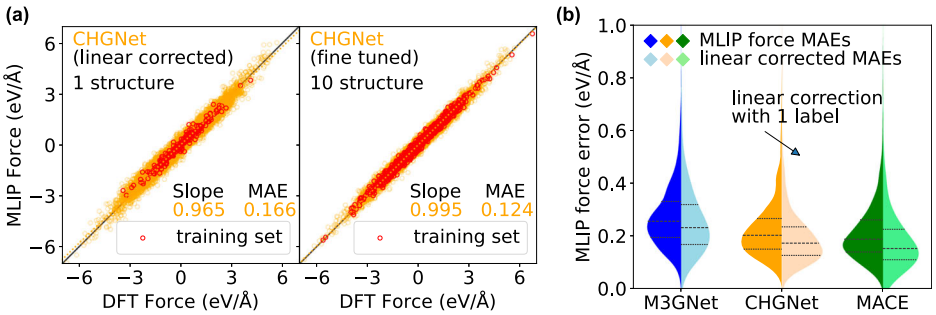

Fig. 6 | Data efficient fine-tuning demonstrated by linear correction. a Parity plots of fine-tuned CHGNet predictions on `mp-973966`, with the training force labels plotted in red and pre-excluded test force labels plotted in orange. Left: fine-tuned CHGNet with a linear correction and a single DFT label solves the softening issue and greatly reduces force MAE from 0.190 eV/Å to 0.166 eV/Å. Right: a more realistic fine-tuning example that optimizes all model parameters with 10 DFT labels, which further decreases the force MAE. **b** Distribution of force MAEs and linear corrected MAEs (cMAEs) for 1000 WBM compounds, showing uMLIP force errors can be greatly reduced by fine-tuning with a single data point. Quartiles are labeled by dashed lines.

These results suggest a theoretical explanation for the commonly observed data-efficient performance boost that is achievable by fine-tuning foundational uMLIPs compared to training randomly initialized MLIPs. The data efficiency arises from the observation that a significant part of the MAEs in pre-trained uMLIPs are highly systematic, which can be efficiently amended by optimizing a fraction of model parameters with a small amount of data. The linear correction demonstrates the fine-tuning mechanism in the extreme case of one structure label and one trainable parameter. In practice, one doesn't necessarily need to fine-tune only a linear layer since a realistic fine-tuning dataset is far larger and richer than a single structure label. In Supplementary Figs. 3 and 4, we present a comparison between the force error in the fine-tuned CHGNet models to those trained from scratch. The result demonstrates that the fine-tuning process can achieve significantly higher data efficiency compared to training MLIPs from scratch. Our theory of systematic error correction provides a mechanistic explanation of these advantages of foundational MLIPs.

Discussion

The design and discovery of novel materials raises the need for advanced simulation tools capable of efficiently and accurately describing the intricate details of atomic interactions. MLIPs offer a potential solution to bridge the gap between quantum mechanical accuracy and affordable computation cost by learning and emulating complex atomic interactions. Recent work on pre-training foundational MLIPs with comprehensive material datasets has opened up the possibility for out-of-box use of robust universal interatomic potentials[7,8,10,12,13].

Unlike DFT, MLIPs cannot by default be expected to perform well in a configurational space where they have not been trained. We therefore benchmark the performance of three uMLIPs for multiple modeling tasks including surfaces, defects, solid-solution energetics, phonon vibration modes, ion migration barriers, and more general high energy states. These states are under-represented in the widely-used pre-training dataset[7,8,17] that only consists of bulk crystalline materials. For the properties tested in this work, we observe a universal softening of the PES, characterized by the uMLIPs' underprediction of energies and forces.

The uMLIP datasets are primarily drawn from Materials Project[17] ionic relaxation trajectories and are therefore largely distributed around the energy minima of the PES. Consequently, the uMLIPs are exposed to a limited range of atomic configurations and force gradients, leading to difficulties in accurately capturing the energy landscapes and steep gradients associated with OOD states and processes like ion migrations and phase transformations.

We found similar signs of softening in the published literature, though less attention was dedicated to an in-depth examination of the softening issue. Pandey et al.[51] and Bartel[52] presented an extrapolation issue arising from a distribution shift when training a CGCNN[18] energy predictor with ICSD data[53]. The CGCNN model trained with only experimental stable materials experienced a six-fold increased prediction MAE when applied to hypothetical crystal structures in the Materials Project[17]. Furthermore, the Google DeepMind's GNoME uMLIP exhibited pronounced softening tendencies when trained on the M3GNet dataset[7], as evidenced in Supplementary Information of ref. 12, similar to our observation in Fig. 5a. After being trained on the expanded dataset of 89 million structures, the softening issue in GNoME was shown to be mitigated but not fully eliminated, which is shown in Supplementary Figs. S34–S37 from ref. 12. These examples underscore the universality of the PES softening issue across various models and datasets, highlighting the importance of the systematic benchmark and analysis undertaken by our study to address this challenge.

Another possible cause of underpredicted energies arises from modified PES minima. Even if the uMLIP would be unbiased with only random errors, uMLIP ionic relaxation may further relax the atoms into positions that are at lower energy, resulting in underpredictions of the relaxed energies. To illustrate this, consider an unbiased uMLIP. When comparing the result of static energy calculations the uMLIP energy predictions would be unbiased compared to the DFT energies. However, when ionic relaxations are performed with the uMLIP on these DFT-relaxed structures, the relaxation may further displace the atoms away from the DFT minimum and reduce the energy. As a result, the expected uMLIP relaxed energy may show a bias to be lower than the DFT energy, even when only random error is present. Compared to the systematic softening discussed in the current manuscript, these erroneous relaxations are much more challenging to resolve as multiple factors are involved: optimization algorithm, relaxation convergence criterion, PES training and validation error, etc. As most physical properties are determined by the energy difference of various configurations, this error arising from the "re-optimization" of the atomic position will affect both states from which the energy difference is derived. For example, in an NEB calculation, the re-optimization error can affect the initial state of the ion as well as the saddle point. Nonetheless, if the random error is larger in high-energy configurations, one expects the re-optimization error to be larger in the high-energy configurations, effectively showing up as softening.

The observed limitations of current uMLIPs raise questions about the effect of model size and expressive capacity on their ability to capture the intricate details of the PES[54]. The MACE model with 4.69 Million parameters, which is around 11 times the size of the CHGNet and 21 times the size of M3GNet, shows improved MAE and decreased softening compared to the smaller uMLIPs. The better performance of larger uMLIPs aligns with the previous study by Frey et al.[55] on the scaling of model performance as a function of MLIP capacity. The observed relationship between model capacity and performance prompts further inquiry into the extent to which the parameter size of current uMLIPs influences the PES softening issue, and whether the softening can be minimized by scaling to a larger, yet reasonable model size without expanding the dataset. In Supplementary Figs. 6 and 7,

we show the distribution of softening scale and force MAEs for two additional uMLIPs: `CHGNet-matgl` and `M3GNet-matgl`, which were also pre-trained using Materials Project database. The `CHGNet-matgl` with increased model size and `M3GNet-matgl` with enhanced sampling[56] demonstrate decreased softening effect and improved force predictions. Furthermore, previous investigations by Xu et al.[57] showed the extrapolation behavior of neural networks tends to be linear, which aligns with our observation of underpredicted curvatures in the uMLIP PES. While the scope of current work does not explicitly investigate the effect of model size and design, further studies could provide a better explanation of the number of model parameters needed to describe a universal potential energy surface.

Fortunately, we demonstrate the softening issue can be effectively resolved by including a minimal amount of high-energy OOD training points in fine-tuning. Our result not only provides a guideline to avoid softening issues when applying uMLIPs to atomic modeling, but more importantly, derives an explanation for the frequently observed data-efficient fine-tuning of foundational MLIPs. Our result suggests that a significant portion of the MAE in uMLIPs is highly systematic and therefore can be efficiently corrected by a small amount of data. In addition to the robustness of uMLIPs that has been acknowledged as an advantage obtained from pre-training[8,12], our study elucidates another benefit of fine-tuning foundational MLIPs – the data-efficient systematic error correction that is unavailable for training a randomly initialized MLIP. Our study serves as a guideline for researchers attempting to fit interatomic potentials for their systems of interest.

In summary, our work presents an in-depth analysis of the softening effect of uMLIPs observed in a series of materials benchmarks, from which we provide guidelines for the fine-tuning effects of uMLIPs. With the observed limitation of current uMLIPs, we advocate the need for an improved next-generation dataset for training foundational atomic models, and more investigation in the role of model complexity. Despite significant efforts dedicated to model design and training strategies, less emphasis has been placed on constructing comprehensive and well-curated open-source materials datasets[58]. Most current foundational models still rely on datasets that were not originally generated for machine learning purposes. Apart from diversifying the chemical space, our findings highlight the importance of ensuring a comprehensive sampling of the PES in generating a reliable MLIP dataset. We believe a next-generation foundational atomic dataset with improved sampling will be pivotal for the development of MLIP and atomistic simulations.

Methods

uMLIP versions

The table below shows the details and versions of the uMLIPs tested Table 1.

Materials modeling tasks

For surface energy calculations, stoichiometric and symmetric slabs are generated with up to a maximum Miller index of 2 in three directions. Minimum slab thickness of 10 Å and minimum vacuum length of 10 Å are used for DFT to ensure convergence of surface energy[59]. When relaxing the slab, in-plane lattice vectors are fixed to their bulk value. The ionic relaxations are converged to a maximum interatomic force criteria of 0.05 eV/Å for all uMLIPs.

For defect energy calculations, defects in elemental phases as well as binary metallic compounds are considered. The defect structures are fully relaxed and referenced to the bulk energy. The off-stoichiometric defect energies (ex: vacancy defect) are referenced to the chemical potential of the pure elemental phase, instead of any chemical potential corresponding to multi-phase equilibria in the phase diagram. This is done deliberately to avoid additional errors associated with calculating the phase diagram using the uMLIPs. For all uMLIPs, the ionic relaxations are converged while a maximum interatomic force is 0.05 eV/Å.

For solid-solution calculation in $Ca_xMg_{2-x}O_2$, we randomly select different Ca-Mg orderings (up to 52 number of configuration) at each Ca concentration and evaluate the energy of the configuration with ionic relaxation with DFT or uMLIPs.

For phonon calculations, we use the `phonopy` workflow as implemented in `atomate2`[60] with relaxation convergence and supercell settings identical to those used in Batatia et al.[10]. The DFT referenced data are taken from the PhononDB[48,49]. We restrict benchmarking materials without magnetism and U-corrections. Moreover, we removed the non-analytic corrections (NAC) from the PBEsol phonons which are derived from the Born effective charges as these are unavailable from uMLIPs which have no concept of electronic structure. In practice, a future hybrid uMLIP-DFT workflow could perform a single DFT static at the uMLIP relaxed structure to obtain Born charges and post-hoc apply non-analytic corrections to the uMLIP phonon spectrum. However, such a hybrid workflow while necessary in practice, would not affect the results of this benchmark concerned specifically with the ML-obtainable parts of the spectrum.

The ion migration barrier DFT data are collected from the work of Rutt et al.[40], in which the ApproxNEB algorithm[42] was used to evaluate Mg^{2+} ion migration barriers. The key difference between ApproxNEB with regular NEB[61] is that ApproxNEB relaxes each image along the migration path independently, while NEB relaxes the migration path collectively. In the ApproxNEB method, an initial guess of the ion migration path is interpolated based on the charge density of the host structure. The energies associated with suggested image structures are calculated by the constrained relaxation that fixes the moving ion and lattice vectors. The ApproxNEB method was shown to provide a comparable barrier within 20 meV error of NEB and reduce the computational time significantly for materials where the path is not too complex[42].

The high-energy states are sampled by high-temperature molecular dynamics. The atomic configurations in Fig. 5a are sampled from a 1000 K ab-initio MD run, and the 1000 materials in Fig. 5b are selected from the WBM dataset[50] and sampled with CHGNet MD run. For each structure selected, a 20 ps, 1000 K molecular dynamics simulation is performed under constant number of particles, volume, and temperature (NVT) ensemble with the pre-trained CHGNet, and 10 structures are subsequently sampled from each MD trajectory[56]. + 3% strain and a − 3% strain are applied along three lattice dimensions for 4 out of the 10 structures to sample strained configurations. All the force MAEs and fine-tuning are calculated with the three-dimensional force components rather than the absolute magnitude of forces.

Fine-tuning

Every fine-tuning and linear correction experiment in the current manuscript is trained separately for each material system. For the fine-tuning of CHGNet uMLIP, the models are trained with energy, force, and stress labels with 0.1-100-0.1 loss fractions under the mean squared error (MSE) loss criterion. The structures and labels are taken from a DFT ab-initio MD trajectory data of $Li_6Zn_2In_2(IO_3)_{16}$ (`mp-973966`) from Jun et al.[62], where 100 structures are reserved for the test set, as shown by the orange points in Fig. 6a. The train-validation ratio is set to be 9:1. As a result, 9 out of the 10 training structures in the right panel of Fig. 6a are actually used for gradient back-propagations. The `Adam` optimizer[63] is used with a learning rate of 1e-3 that cosine decays to 1e-5 in 100 epochs. The model checkpoint of best validation force MAE is collected for test set predictions. For the model trained with only 1 structure, the last-epoch checkpoint is used instead.

The linear correction of CHGNet is realized by adding a hypothetical scalar linear before the energy prediction. The weight of the scalar linear layer is initialized to be 1, therefore not influencing the energy prediction before being optimized. During the linear correction, all CHGNet model parameters are frozen except for the added scalar linear layer.

Table 1 | uMLIP Model Specifications

Model	Version	ModelSize	DataSet	DataSize
M3GNet[7]	2021.2.8	227.5K	MPF[7]	188.3K
CHGNet[8]	v0.3.0	412.5K	MPtrj[8]	1.58M
MACE[10]	2023.12.03	4.69M	MPtrj	1.58M

DFT calculations

DFT calculations were performed with the *Vienna ab initio simulation package* (VASP) using the projector-augmented wave method[64,65]. All calculation settings are generated using `pymatgen MPRelaxSet` to ensure all DFT results are compatible with Materials Project DFT calculations[66]. All the calculations were converged to at least 10^{-5} eV in total energy for electronic steps and 0.02 eV/Å in interatomic forces for ionic steps.

Data availability

The dataset used to extract the softening scales of uMLIPs is available at https://doi.org/10.6084/m9.figshare.27307776[67].

Received: 23 May 2024; Accepted: 20 December 2024;
Published online: 10 January 2025

References

1. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
2. Wang, H., Zhang, L., Han, J. & Weinan, E. Deepmd-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228**, 178–184 (2018).
3. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
4. Guo, Z. et al. Extending the limit of molecular dynamics with ab initio accuracy to 10 billion atoms. In *Proc. of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* 205–218 (2022).
5. Musaelian, A., Johansson, A., Batzner, S. & Kozinsky, B. Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size. In *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–12 (2023).
6. Gasteiger, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. *ICLR*, *arXiv preprint arXiv:2003.03123* (2020).
7. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
8. Deng, B. et al. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).
9. Choudhary, K. et al. Unified graph neural network force-field for the periodic table: solid state applications. *Digit. Discov.* **2**, 346–355 (2023).
10. Batatia, I. et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*. https://arxiv.org/abs/2401.00096 (2023).
11. Takamoto, S. et al. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nat. Commun.* **13**, 2991 (2022).
12. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* 1–6 (2023).
13. Zhang, D. et al. Pretraining of attention-based deep learning potential model for molecular simulation. *npj Comput. Mater.* **10**, 94 (2024).
14. Riebesell, J. et al. Matbench discovery – an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920*. https://arxiv.org/abs/2308.14920 (2023).
15. Focassio, B., Freitas, L. P. M. & Schleder, G. R. Performance assessment of universal machine learning interatomic potentials: Challenges and directions for materials' surfaces. *ACS Appl. Mater. Interfaces*. https://doi.org/10.1021/acsami.4c03815 (2024).
16. Yu, H., Giantomassi, M., Materzanini, G., Wang, J. & Rignanese, G.-M. Systematic assessment of various universal machine-learning interatomic potentials. *Mater. Genome Eng. Adv.* **2**, 3 (2024).
17. Jain, A. et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
18. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
19. Luo, S., Chen, T. & Krishnapriyan, A. S. Enabling efficient equivariant operations in the fourier basis via gaunt tensor products. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=mhyQXJ6JsK (2024).
20. Cheng, B. Cartesian atomic cluster expansion for machine learning interatomic potentials. *npj Comput. Mater.* **10**, 157 (2024).
21. Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. Accurate fourth-generation machine learning potentials by electrostatic embedding. *J. Chem. Theory Comput.* **19**, 3567–3579 (2023).
22. Chanussot, L. et al. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **11**, 6059–6072 (2021).
23. Obot, I., Macdonald, D. & Gasem, Z. Density functional theory (DFT) as a powerful tool for designing new organic corrosion inhibitors. part 1: An overview. *Corros. Sci.* **99**, 1–30 (2015).
24. Han, Y. et al. Surface energies, adhesion energies, and exfoliation energies relevant to copper-graphene and copper-graphite systems. *Surf. Sci.* **685**, 48–58 (2019).
25. Sun, W., Kitchaev, D. A., Kramer, D. & Ceder, G. Non-equilibrium crystallization pathways of manganese oxides in aqueous solution. *Nat. Commun.* **10**, 573 (2019).
26. Fichthorn, K. A. & Scheffler, M. Island nucleation in thin-film epitaxy: A first-principles investigation. *Phys. Rev. Lett.* **84**, 5371–5374 (2000).
27. Gurylev, V. & Perng, T. P. Defect engineering of zno: Review on oxygen and zinc vacancies. *J. Eur. Ceram. Soc.* **41**, 4977–4996 (2021).
28. Broberg, D. et al. High-throughput calculations of charged point defect properties with semi-local density functional theory—performance benchmarks for materials screening applications. *npj Comput. Mater.* **9**, 72 (2023).
29. Ahangari, M. G. et al. Effect of various defects on mechanical and electronic properties of zinc-oxide graphene-like structure: A DFT study. *Vacuum* **165**, 26–34 (2019).
30. Kang, K. & Ceder, G. Factors that affect li mobility in layered lithium transition metal oxides. *Phys. Rev. B* **74**, 094105 (2006).
31. van de Walle, A. Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit. *Calphad* **33**, 266–278 (2009).
32. Cation-disordered rocksalt-type high-entropy cathodes for Li-ion batteries. *Nat. Mater.* **20**, 214–221 (2021).
33. Barroso-Luque, L. et al. Cluster expansions of multicomponent ionic materials: Formalism and methodology. *Phys. Rev. B* **106**, 144202 (2022).
34. Zhong, P., Xie, F., Barroso-Luque, L., Huang, L. & Ceder, G. Modeling intercalation chemistry with multiredox reactions by sparse lattice models in disordered rocksalt cathodes. *PRX Energy* **2**, 043005 (2023).
35. Ceder, G. A derivation of the ising model for the computation of phase diagrams. *Comput. Mater. Sci.* **1**, 144–150 (1993).
36. DOMAN, R. C., BARR, J. B., McNALLY, R. N. & ALPER, A. M. Phase equilibria in the system cao—mgo. *J. Am. Ceram. Soc.* **46**, 313–316 (1963).
37. Jung, I.-H., Decterov, S. A. & Pelton, A. D. Critical thermodynamic evaluation and optimization of the CaO–MgO–SiO2 system. *J. Eur. Ceram. Soc.* **25**, 313–333 (2005).
38. Deng, Z., Radhakrishnan, B. & Ong, S. P. Rational composition optimization of the Lithium-Rich Li3OCl1–x Br x anti-perovskite superionic conductors. *Chem. Mater.* **27**, 3749–3755 (2015).
39. Du, P. et al. Cooperative origin of proton pair diffusivity in yttrium substituted barium zirconate. *Commun. Phys.* **3**, 200 (2020).
40. Rutt, A. et al. Expanding the material search space for multivalent cathodes. *ACS Appl. Mater. Interfaces* **14**, 44367–44376 (2022).
41. Shen, J.-X., Horton, M. & Persson, K. A. A charge-density-based general cation insertion algorithm for generating new li-ion cathode materials. *npj Comput. Mater.* **6**, 161 (2020).

42. Rong, Z., Kitchaev, D., Canepa, P., Huang, W. & Ceder, G. An efficient algorithm for finding the minimum energy path for cation migration in ionic materials. *J. Chem. Phys.* **145**, 074112 (2016).
43. Ven, A. V. D., Ceder, G., Asta, M. & Tepesch, P. D. First-principles theory of ionic diffusion with nondilute carriers. *Phys. Rev. B* **64**, 184307 (2001).
44. Walle, A. V. D. & Ceder, G. The effect of lattice vibrations on substitutional alloy thermodynamics. *Rev. Mod. Phys.* **74**, 11–45 (2002).
45. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
46. Yue, S. et al. Phonon softening near topological phase transitions. *Phys. Rev. B* **102**, 235428 (2020).
47. Parlinski, K., Li, Z. Q. & Kawazoe, Y. First-principles determination of the soft mode in cubic ZrO2. *Phys. Rev. Lett.* **78**, 4063–4066 (1997).
48. Togo, A. First-principles Phonon calculations with Phonopy and Phono3py. *J. Phys. Soc. Jpn.* **92**, 012001 (2023).
49. Togo, A., Chaput, L., Tadano, T. & Tanaka, I. Implementation strategies in phonopy and phono3py. *J. Phys.: Condens. Matter* **35**, 353001 (2023).
50. Wang, H.-C., Botti, S. & Marques, M. A. L. Predicting stable crystalline compounds using chemical similarity. *npj Comput. Mater.* **7**, 12 (2021).
51. Pandey, S., Qu, J., Stevanović, V., John, P. S. & Gorai, P. Predicting energy and stability of known and hypothetical crystals using graph neural network. *Patterns* **2**, 100361 (2021).
52. Bartel, C. J. Data-centric approach to improve machine learning models for inorganic materials. *Patterns* **2**, 100382 (2021).
53. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr. Sect. B* **58**, 364–369 (2002).
54. Ko, T. W. & Ong, S. P. Recent advances and outstanding challenges for machine learning interatomic potentials. *Nat. Comput. Sci.* 1–3 (2023).
55. Frey, N. C. et al. Neural scaling of deep chemical models. *Nat. Mach. Intell.* 1–9 (2023).
56. Qi, J., Ko, T. W., Wood, B. C., Pham, T. A. & Ong, S. P. Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling. *npj Comput. Mater.* **10**, 43 (2024).
57. Xu, K. et al. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848* (2020).
58. Barroso-Luque, L. et al. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771* (2024).
59. Sun, W. & Ceder, G. Efficient creation and convergence of surface slabs. *Surf. Sci.* **617**, 53–59 (2013).
60. Ganose, A. et al. atomate2. https://github.com/materialsproject/atomate2 (2024).
61. Henkelman, G., Uberuaga, B. P. & Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **113**, 9901–9904 (2000).
62. Jun, K. et al. Lithium superionic conductors with corner-sharing frameworks. *Nat. Mater.* **21**, 924–931 (2022).
63. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
64. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
65. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
66. Ong, S. P. et al. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
67. Deng, B. Wbm high energy states. figshare. Dataset (2024). https://figshare.com/articles/dataset/WBM_high_energy_states/27307776.

Acknowledgements

This work was funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC0205CH11231 (Materials Project program KC23MP). The work was also supported by the computational resources provided by the Extreme Science and Engineering Discovery Environment (XSEDE), supported by National Science Foundation grant number ACI1053575; the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory; and the Swift Cluster resource provided by the National Renewable Energy Laboratory (NREL). The authors would also like to thank Tsz Wai Ko and Yuanqi Du for helpful discussions.

Author contributions

B.D. and G.C. conceived the initial idea. Z.L. and B.D. performed benchmarks on surface energies. S.A. and B.D. performed benchmarks on defect energies. P.Z. and B.D. performed benchmarks on solid solutions. Y.C. performed benchmarks on ion migration barriers. J.R. performed benchmarks on phonons. B.D. performed benchmarks on high-energy states. B.D. conceived and analyzed the softening scale and linear corrections. K.P. and G.C. offered insight and guidance throughout the project. All authors contributed to discussions and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01500-6.

Correspondence and requests for materials should be addressed to Gerbrand Ceder.