Check for updates

Received: 29 February 2024 | Accepted: 6 July 2024

DOI: 10.1002/mgea.58

RESEARCH ARTICLE

MGE ADVANCES
Materials Genome Engineering

Systematic assessment of various universal machine-learning interatomic potentials

Haochen Yu[1] | **Matteo Giantomassi**[1] | **Giuliana Materzanini**[1] | **Junjie Wang**[2] | **Gian-Marco Rignanese**[1,2,3]

[1]Institute of Condensed Matter and Nanosciences, Université catholique de Louvain, Louvain-la-Neuve, Belgium

[2]State Key Laboratory of Solidification Processing, Northwestern Polytechnical University, Xi'an, Shaanxi, China

[3]WEL Research Institute, Wavre, Belgium

Correspondence
Gian-Marco Rignanese.
Email: gian-marco.rignanese@uclouvain.be

Abstract

Machine-learning interatomic potentials have revolutionized materials modeling at the atomic scale. Thanks to these, it is now indeed possible to perform simulations of ab initio quality over very large time and length scales. More recently, various universal machine-learning models have been proposed as an out-of-box approach avoiding the need to train and validate specific potentials for each particular material of interest. In this paper, we review and evaluate four different universal machine-learning interatomic potentials (uMLIPs), all based on graph neural network architectures which have demonstrated transferability from one chemical system to another. The evaluation procedure relies on data both from a recent verification study of density-functional-theory implementations and from the Materials Project. Through this comprehensive evaluation, we aim to provide guidance to materials scientists in selecting suitable models for their specific research problems, offer recommendations for model selection and optimization, and stimulate discussion on potential areas for improvement in current machine-learning methodologies in materials science.

KEYWORDS

formation energy, geometry optimization, machine learning, phonons, universal machine-learning interatomic potentials, verification

1 | INTRODUCTION

Materials simulations at the atomic scale are the backbone of computational materials design and discovery. They rely on the Born–Oppenheimer approximation, in which the electrons follow the nuclear motion adiabatically, so that the potential governing the nuclei consists of the electronic energies as a function of the nuclear positions, called "potential energy surface" (PES). Knowledge of the PES allows the identification of stable and metastable atomic configurations from minimum energy search, or the determination of materials properties as thermodynamical averages from molecular dynamics simulations.[1] The utility of atom-based materials simulations is thus intimately related to the generation of accurate PESs, which has been possible in the last few decades thanks to the advent of density-functional theory (DFT).[2–5] Nonetheless, this ab initio approach relies on the quantum mechanical solution of the electronic problem whose computational cost scales cubically with system size and can therefore become unaffordable in various significant cases of technological interest such as amorphous solids, interfaces, surfaces, etc. At the other end of the simulation approaches, parametrized approximations of the Born–Oppenheimer PES, known as empirical analytical potentials, or force fields, or "classical" interatomic potentials, have been widely used especially for

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Materials Genome Engineering Advances* published by Wiley-VCH GmbH on behalf of University of Science and Technology Beijing.

large-scale material studies.[6] Unfortunately, in particular when complex electron interactions are involved (as in chemical reactions or phase transitions) these approaches cannot usually achieve DFT accuracy, and in addition they have limited applicability and transferability. They cannot therefore be considered as a drop-in replacement for standard ab initio methods. In this context, machine-learning interatomic potentials (MLIPs) have emerged as an in-between solution with computational cost similar to the empirical analytical potentials, but with the promise of achieving an accuracy comparable to DFT hence enabling accurate simulations over very large time and length scales.[7] The key difference with respect to the empirical potentials is that the interatomic potentials are now directly obtained via a highly nonlinear fit of a set of input/target data (in general, at the DFT accuracy), without any a priori assumption on their analytical form.[8] In the original formulation, building an MLIP consists of generating a dataset of atomic configurations for the specific material under study and training (and subsequently validating) the MLIP on these data based on the accurate prediction of some target metrics as, for example, energies, forces, and stresses.[8–16] This process is highly material dependent, and usually requires a significant human and computational effort.[17] Obtaining an accurate description of the PES without the need for costly DFT computations, but also covering all possible chemical and structural spaces, would be the holy grail of MLIPs. In this framework, graph Neural Networks (GNN) have revealed to be very effective, in particular in message passing and equivariant[15,18] NN.[19] In general, NN-based methods can be particularly useful for generalizability thanks to their property of "learning locally" that makes the resulting potential less material dependent.[8,9,20–22] Indeed, the first "universal" MLIP (uMLIP), the MEGNet[23] model, exploited a graph network architecture. It was trained on ~60,000 inorganic crystals in their minimum energy configurations from the Materials Project (MP)[24] database, which covers 89 elements of the periodic table and is based on the Perdew–Burke–Enzerhof (PBE) exchange-correlation functional.[25] This model could provide the formation energy as well as a number of other properties. In order to predict forces and stresses, various other models were subsequently developed relying on different datasets. In particular, the M3GNet[26] and CHGNet[27] models relying on equivariant graph neural network architectures were both developed using snapshots from DFT relaxations of the MP structures. The publication introducing CHGNet was also the opportunity for releasing the Materials Project Trajectory (MPtrj) dataset,[27] including the DFT calculations for more than 1.5 million atomic configurations of inorganic structures. By including magnetic moments in the training properties, CHGNet aimed at a better description of chemical reactions as charged states influence how atoms connect with others through chemical bonds.[27] A re-implementation of M3GNet, MatGL, has been built on the Deep Graph Library and PyTorch in order to improve usability and scalability.[28] Recently, the M3GNet-DIRECT model has been released[29] in the MatGL package[28] with 1.3 million structures and 89 elements, aiming to predict more reliably unseen structures thanks to a new sampling scheme.[29] The ALIGNN-FF model[30] was developed to model a diverse set of materials with any combination of 89 elements from the periodic table but relying on a different database of inorganic crystals, JARVIS-DFT,[31] which is based on the optB88vdW exchange-correlation functional.[32] Recently, the MACE-MP-0[33] relying on the MACE architecture[19] was also proposed. It was trained on the MPtrj dataset and showed outstanding performance on an extraordinary range of examples from quantum-chemistry and materials science.[33] Two proprietary models relying on very large databases have also been developed: the GNoME[34] model exploiting the NequIP architecture[15] and the PFP model[35,36] exploiting the TeaNet architecture.[37] On the one hand, GNoME was trained on a database obtained from a complex active learning workflow of the original MP data, resulting in a number of inorganic structures ~100 times larger than MPtrj. On the other hand, the PFP was trained on a large dataset which initially included ~10^7 DFT configurations covering 45 elements[35] and was further extended to cover 72 elements at the moment of the publication,[36] with an expected rise to 94 including rare-earth elements and actinides. It is worth mentioning that uMLIPs have also been developed specifically for organic molecules[38–40] and for metal alloys.[41] While this rapid growth of uMLIP models promises an extraordinary impact in the materials science community, benchmark efforts are crucially needed to assess and control their performance and effective usefulness.[42–44]

In this paper, we conduct a comprehensive review and evaluation of four different GNN-based uMLIPs: M3GNet-DIRECT,[29] CHGNet,[27] MACE-MP-0,[33] and ALIGNN-FF.[30] They have demonstrated the possibility of universal interatomic potentials that may not require re-training for new applications. The evaluation uses three different datasets: for the equation of state, we use the set of theoretical structures employed in Ref.[45], for the phonon calculations, we use the crystalline structures considered in Ref.[46] that have been relaxed with ABINIT,[47,48] norm-conserving pseudopotentials,[49,50] and the PBEsol exchange-correlation functional[51] while for all the other tests, we use the VASP-relaxed structures from the Materials Project (MP).[24,52]

The paper is organized as follows. After a brief overview of the techniques employed in the generation of the different uMLIPs examined in this work (Section 2.1), we test and discuss the quality and transferability of the chosen uMLIPs in different types of calculations, and using the three different datasets as explained above. The tests include the calculation of the equation of state (Section 2.2), the evaluation of formation energies and optimization of structural parameters (Section 2.3), and the calculation of phonon bands (Section 2.4). The conclusions are provided in Section 3 and the details of the methods employed are given in Section 4.

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliotheque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2 | DISCUSSION

2.1 | Overview of the uMLIP models

In what follows, we give a brief overview of the basic ingredients underlying the four GNN-based uMLIPs examined.[27,29,30,33] Depending on the uMLIPs, they might differ for the specific GNN architecture employed, for the choice of the training set, for the sampling scheme, or for the properties included in the training. A comprehensive and detailed explanation is beyond the scopes of the present work, and we refer the interested reader to the respective original papers[27,29,30,33] and related literature.[19,23,26,31,53–56] GNN are reviewed, among others, in Refs.[18,57,58], while Refs.[8–17,59] address specifically the construction of MLIP in materials science.

The pre-trained M3GNet-DIRECT[29] model starts from the Materials 3-body Graph Network (M3GNet) architecture,[26] which conjugates the flexibility of typical graph techniques with the physically based many-body features (here 3-body) of traditional interatomic potentials like the Tersoff bond-order potential.[60] The many-body interactions (angles) are aggregated to bonds in standard graph convolution steps to update the bond, atom, and global state (properties) information.[26] The main improvement of M3GNet-DIRECT[29] with respect to the original M3GNet[26] is in the sampling of the materials space. This is done starting from the 128-element vector outputs from the M3GNet[26] model trained on the MP formation energies database,[24] to which dimensionality reduction, characteristics-sharing-based clustering, and stratification, are subsequently applied, giving the name of DImensionality-Reduced Encoded Clusters with sTratified (DIRECT) sampling to the approach.[29] Static DFT calculations on the so-obtained 1.3 million structures produce the final training set.[29] The result is an improved performance with respect to the original M3GNet,[26] with a better transferability and prediction of configurations with large energies, forces, and stresses.[29]

The Crystal Hamiltonian Graph Neural Network (CHGNet[27]) is a model pre-trained on the energies, forces, stresses, and magnetic moments from the MPtrj[27] dataset. The GGA/GGA + U mixing compatibility corrections[61–63] are applied to the energies.[27] In the CHGNet architecture, the angle information is drawn as a pairwise message passing convolution between bonds (bond graph, where bonds are nodes and edges are angles), on top of the bond information which is drawn as a pairwise convolution between atoms (atom graph, where atoms are nodes and edges are bonds).[27] To constrain the atom features used to predict energy, forces, and stresses by their charge-state information, the latter is inferred from the magnetic moments and atomic orbital theory before the last convolution layer.[27] In incorporating the site-specific magnetic moments as the input charge states into the GNN, this model aims to capture the chemical interaction variability across different valence states.[27] This should be especially important for transition metal elements, with variable valence states.[64]

The pre-trained MACE-MP-0 model[33] relies on MACE,[19] an equivariant message passing neural network potential using messages of an order higher than two body. In combining high body order with message passing, MACE achieves superior accuracy as a result of a trade-off between using more layers and increasing the local body order within a single layer.[55] Using 4-body messages,[19] the required number of message passing iterations is reduced to two and the steepness of the learning curve is improved.[19] The high-order equivariant message passing neural network model is combined with the atomic cluster expansion (ACE[13,65]), a framework for deriving an efficient body-ordered symmetric polynomial basis to represent functions of atomic neighborhoods. ACE naturally extends to equivariant features and to include variables beyond geometry, such as charges or magnetic moments.[66] MACE-MP-0 is pre-trained on the MPtrj[27] dataset, including a variety of magnetic orders.[33] The fitting does not explicitly account for the GGA/GGA + U mixing compatibility corrections.[61–63]

The pre-trained ALIGNN-FF model[30] is the extension to derivative prediction and force field (FF) formalism[30] of the Atomistic LIne Graph Neural Network (ALIGNN) models,[56] able to capture two and three body interactions and to output more than 70 materials properties.[30] The ALIGNN[56] models were implemented using the deep graph library (DGL)[67] which allows efficient line graph construction and neural message passing. As in the case of CHGNet, the message passing occurs both in an atom graph, where atoms are nodes and interatomic bonds are edges, and in a line graph, where bonds are nodes and bond angles are edges.[56] For the atom graph, nine input node features are assigned to each node, based on the atomic species involved: electronegativity, group number, covalent radius, valence electrons, first ionization energy, and electron affinity, block, and atomic volume. ALIGNN is a part of the Joint Automated Repository for Various Integrated Simulations (JARVIS) infrastructure,[31] a set of databases and tools for materials design. The ALIGNN-FF model was trained on the JARVIS-DFT dataset which contains around 75,000 materials and 4 million energy-force entries, out of which 307,113 were used in the training.[30]

For the sake of simplicity, in what follows we use a shorter nomenclature for three out of the four uMLIPs benchmarked: M3GNet-DIRECT[29] will be called M3GNet, MACE-MP-0[33] will be referred to as MACE, and ALIGNN-FF[30] will be termed ALIGNN. Some key features of the four uMLIPs are listed in Table 1.

2.2 | Equation of state and comparison with all-electron results

As a first test, we use the protocol for the equation of state (EOS) detailed in Ref.[45], where a high-quality reference

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliotheque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

dataset of EOS for 960 cubic crystal structures is generated by employing two all-electron (AE) codes. This dataset includes all elements from $Z = 1$ (hydrogen) to $Z = 96$ (curium). For each element X, four mono-elemental cubic crystals (*unaries structures*) are considered, in the face-centered cubic (FCC), body-centered cubic (BCC), simple cubic (SC), and diamond crystal structure, respectively. Besides, for each element, six binary cubic oxides (*oxides structures*) are included, with chemical formula X_2O, XO, X_2O_3, XO_2, X_2O_5, and XO_3, respectively. While in Ref.[45] this dataset is used to gauge the precision and transferability of nine pseudopotential-based ab initio codes, we use it here to assess the four tested uMLIPs.

In Figure 1, the relative errors on the uMLIPs predictions for the equilibrium volume V_0, the bulk modulus B_0, and its derivative with respect to the pressure B_1 obtained from a fit[45] of the EOS, are compared with the analogous errors from selected ab initio methods employed in Ref.[45]. FLEUR[76] and WIEN2K[77] are AE codes while VASP[52] implements the projector augmented-wave method (PAW) method[78] with a planewave basis set. The errors reported in Figure 1 are calculated with respect to the average of the AE methods considered here, that is, FLEUR[76] and WIEN2K.[77]

Furthermore, as in Ref.[45], the EOS computed with two different computational approaches a and b ($E_a(V)$ and $E_b(V)$) are compared through two metrics $\epsilon(a, b)$ and $v(a, b)$. The first is a renormalized dimensionless version of the metrics $\Delta(a, b)$ introduced in Ref.[79]:

$$\epsilon(a,b) = \sqrt{\frac{\sum_i [E_a(V_i) - E_b(V_i)]^2}{\sqrt{\sum_i [E_a(V_i) - \langle E_a \rangle]^2 \sum_i [E_b(V_i) - \langle E_b \rangle]^2}}}, \quad (1)$$

where the index i runs over the explicit calculations of $E_{a,b}(V)$ for the different methods and $\langle E_{a,b} \rangle$ is the average of $E_{a,b}(V)$ over the considered volume range. The second metrics, dependent directly on the physically measurable quantities V_0, B_0, and B_1, captures the relative deviation for each of these three parameters between the two computational approaches a and b:

$$v_{w_{V_0}, w_{B_0}, w_{B_1}}(a,b) = 100 \sqrt{\sum_{Y=V_0,B_0,B_1} \left[w_Y \frac{Y_a - Y_b}{(Y_a + Y_b)/2} \right]^2}, \quad (2)$$

where w_{V_0}, w_{B_0}, and w_{B_1} are appropriately chosen weights (see Ref.[45]). In Equations (1) and (2), a is a uMLIP and b is the average of the AE methods considered here, that is, FLEUR[76] and WIEN2K.[77] Heatmaps of the periodic table with the values of the comparison metrics ε and v obtained with all the different uMLIPs are reported in Figures S1–S4.

It should be noted that most structures in the dataset used in this EOS test are not stable in nature, so that they were not

TABLE 1 Specifications of the uMLIPs benchmarked. Model sizes refers to the number of parameters used for training from the different models.[27,43,44]

uMLIP label	Model	Version	Model size	Data set	$N_{elements}$	Data size
'CHGNet'	CHGNet[27]	0.3.0[68]	400K[27]	MP-trj[69]	89	1.58M[27]
'M3GNet'	M3GNet-DIRECT[29]	2021.2.8-DIRECT-PES[70]	1.1M[43]	MPF[24] + DIRECT[29]	128	1.31M[29]
'MACE'	MACE-MP-0[33]	2023-12-03-mace-128-L1_epoch-199[71,72]	4.7M[44]	MP-trj[69]	89	1.58M[27]
'ALIGNN'	ALIGNN-FF[30]	alignnff_wt10[73,74]	4.0M[75]	JARVIS-DFT[31]	89	75K[30,31]

Note: Data Size refers to the number of structures included in the training Data Set.[29] $N_{elements}$ refers to the number of chemical species (elements) covered.

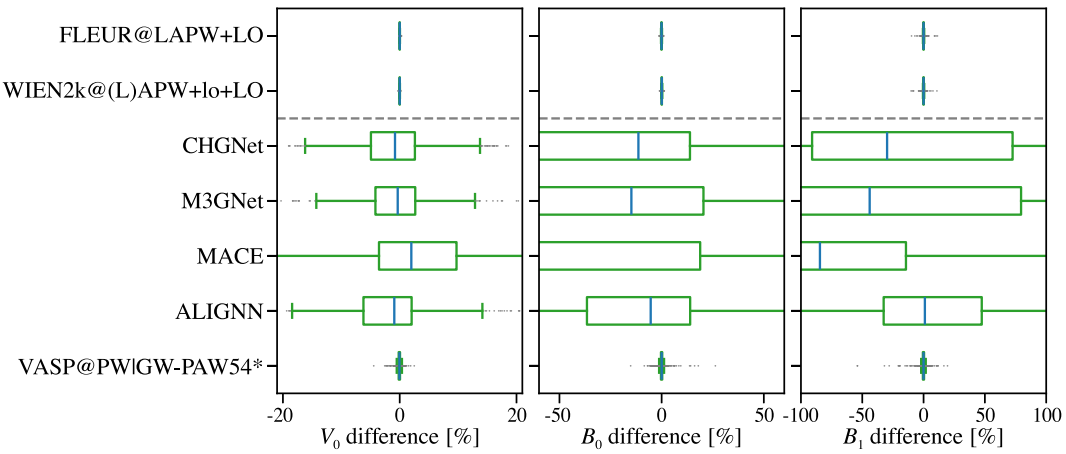

FIGURE 1 Boxplot showing the relative error in V_0, B_0, and B_1 for the different uMLIPs analyzed in this work and some of the ab initio codes considered in Ref. [45] with respect to the average of the AE reference results.

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliotheque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

likely included in the dataset used to train the uMLIPs. As a consequence, it is not surprising that the uMLIPs are not able to predict the correct energy versus volume curve for a significant fraction of the systems (Figures S1–S4). Nevertheless, even for those systems for which a physical EOS is obtained, the precision and transferability is still far from the one that can be achieved with state-of-the-art pseudopotential-based ab initio techniques, as can be observed in Figure 1. This is a very stringent test for uMLIPs, especially given the effort made in Ref.[45] to improve the precision and the transferability of pre-existent pseudopotential tables used for ab initio calculations. Yet these results suggest that uMLIPs predictions should be taken with some caution and, if possible, validated a posteriori via ab initio calculations, especially if the chemical/physical environment under study is not properly included in the training dataset. The precision of uMLIPs can be improved after retraining the model by including additional ab initio data capturing the chemical/physical configurations under investigation.[43,44] However, despite their relevance, these topics are beyond the scope of the present work and are left for future investigation.

2.3 | Structural optimization and formation energies

To test the accuracy of the different uMLIPs further, we prepare a dataset with 19998 materials consisting of unary and binary (with 6903 element combinations) phases in MP. It is worthwhile to highlight that the MP database contains structures relaxed with VASP[52] with the PBE functional.[25] The distribution of the chosen dataset on the chemical elements is reported in Figure 2. We use the different uMLIPs to perform energy calculations both with no structural optimization (we call these "one-shot" calculations) and with structural optimization. The latter are of two types: (i) only the atomic positions are relaxed (we call these "ion-relax" calculations) and (ii) both atomic positions and cell parameters (*i.e.*, lattice parameters and angles) are relaxed (we call these "cell-relax" calculations). For the structural optimizations, in some cases the calculation stops due to errors while building the graph representation (e.g., given that isolated atoms are found in the structure). Calculations for which this problem appears are tagged as *problematic*. In another nonnegligible number of cases, which we tag as *unconverged*, the relaxation algorithm is not able to reach the stopping criterion before 150 steps. Finally, we tag as *converged* the calculations where the stopping criterion is met in less than 150 steps. Additional details on the relaxation algorithm are given in Section 4.

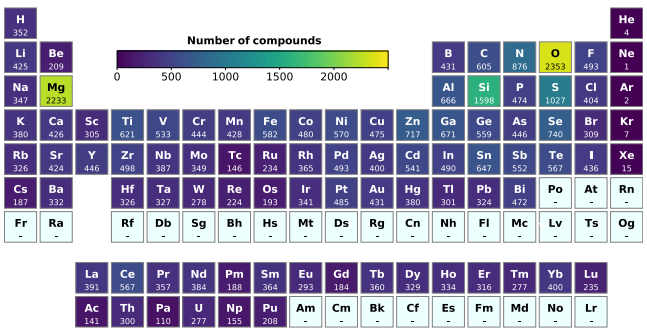

FIGURE 2 Heatmaps of the periodic table for the dataset chosen for the structural relaxations (19998 unary and binary compounds from the Materials Project), expressing the data distribution over the chemical composition space. The figure has been produced with the pymatviz tool.[80]

In Figure 3, we report the number of occurrences of the *problematic*, *unconverged*, and *converged* tags from the different uMLIPs for the two types of structural optimization. When only the atomic positions are relaxed (Figure 3a), the fraction of *unconverged* cases is very limited (about 0.3 −0.4% of the total for all the uMLIPs), while only ALIGNN presents some *problematic* cases worth to be noted, but still very small (0.2%). The fraction of *unconverged* calculations increases significantly when both atomic positions and cell parameters are optimized (Figure 3b), especially for M3GNet (2.5%) and ALIGNN (4.4%). CHGNet and MACE perform more robustly, with lower fractions of *unconverged* results (both 1.2%). In the following, we discuss only results from the *converged* geometry optimizations.

To compare the predictive performance of these uMLIPs on the energetics, we use as a benchmark the formation energy, defined as follows:

$$E_{\text{form}}[A_aB_b] = E[A_aB_b] - x_aE[A] - x_bE[B], \tag{3}$$

where $E[A_aB_b]$ is the total energy per atom of the phase of interest with A, B the constituent elements, $x_a = a/(a+b)$ and $x_b = b/(a+b)$ are the fractions of A and B, respectively, and $E[A]$ and $E[B]$ are the lowest possible energies per atom of A and B, respectively. The ability of a uMLIP in predicting formation energies is evaluated as a difference from the MP values, *that is* as follows:

$$\Delta E_{\text{form}} = E_{\text{form}}^{\text{MP}} - E_{\text{form}}^{\text{uMLIP}}, \tag{4}$$

with $E_{\text{form}}^{\text{MP}}$ and $E_{\text{form}}^{\text{uMLIP}}$ being the formation energy (Equation (3)) from the MP and calculated with the uMLIP model, respectively.

We choose to work with the one-shot energies as this allows to avoid spurious effects due to the different relaxed geometries predicted by the uMLIPs. It is important to note that all our calculations incorporate energy corrections. For CHGNet and M3GNet, these corrections are already encoded in the uMLIPs, as both models were trained using corrected energies.[61–63] Conversely, for MACE, the corrections are applied as a post-processing step (see Section 4 and Refs.[81,82] for details) since this model was trained on uncorrected entries.[33] ALIGNN, on the other hand, was trained using the JARVIS-DFT dataset with OptB88vdW

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliotheque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

functional,[32] and does not use any corrected energies, but rather direct DFT outputs. In this case, we compute Equation (4) with and without energy corrections, and we select the best-agreement set of results.

Figure 4 shows the distribution of the absolute mean values of ΔE_{form} on the chemical space for the different uMLIPs, that is, the average of the absolute value of ΔE_{form} over all the chemical systems containing a given element. CHGNet and MACE have the best performance over all the periodic table. For what concerns the elements in the transition metal series, CHGNet and MACE perform significantly better than M3GNet and ALIGNN, especially for elements like V, Mo, and W. In the case of CHGNet, this might be related to the inclusion of the magnetic moments in the training process. CHGNet and MACE outperform M3GNet and ALIGNN also for what concerns the calchogens (O, S, Se, and Te) and some halogens (F, Cl, and I). ALIGNN is also less performant for actinide elements Np and Pu compared to other uMLIPs. In Figure 5, we report the distribution of ΔE_{form} (Equation (4)) for the one-shot calculations using the different uMLIPs, while the performance of each model is summarized in Table 2. MACE shows better performance than the other uMLIPs, with the smallest MAE and RMSE and the highest R^2 (0.044, 0.101, and 0.989, respectively). In turn, MACE outperforms the other uMLIPs in the one-shot calculations (Figure 5 and Table 2),

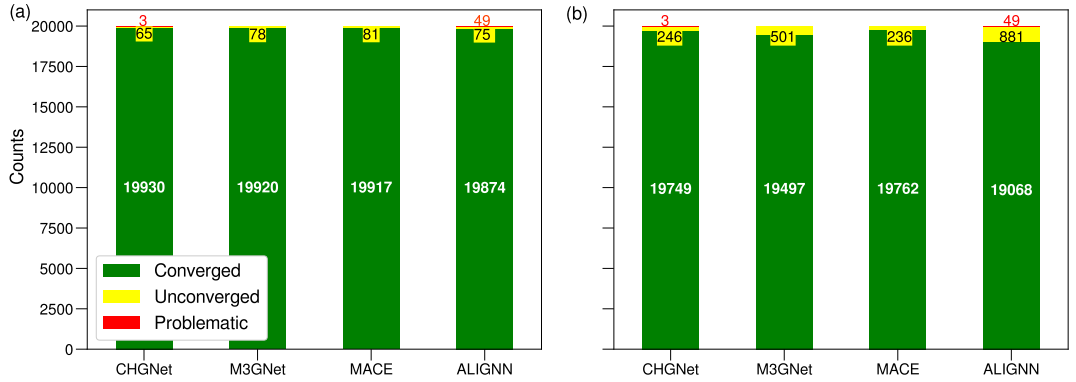

FIGURE 3 Analysis of the convergence of the different uMLIPs for geometry optimization. In panel (a), only the atomic positions are relaxed, while the cell parameters are kept fixed at the original values. In panel (b), both the atomic positions and the cell parameters are relaxed. As explained in the text, a calculation is considered to be converged when the relaxation criteria (in terms of forces and stresses) are met within 150 steps.

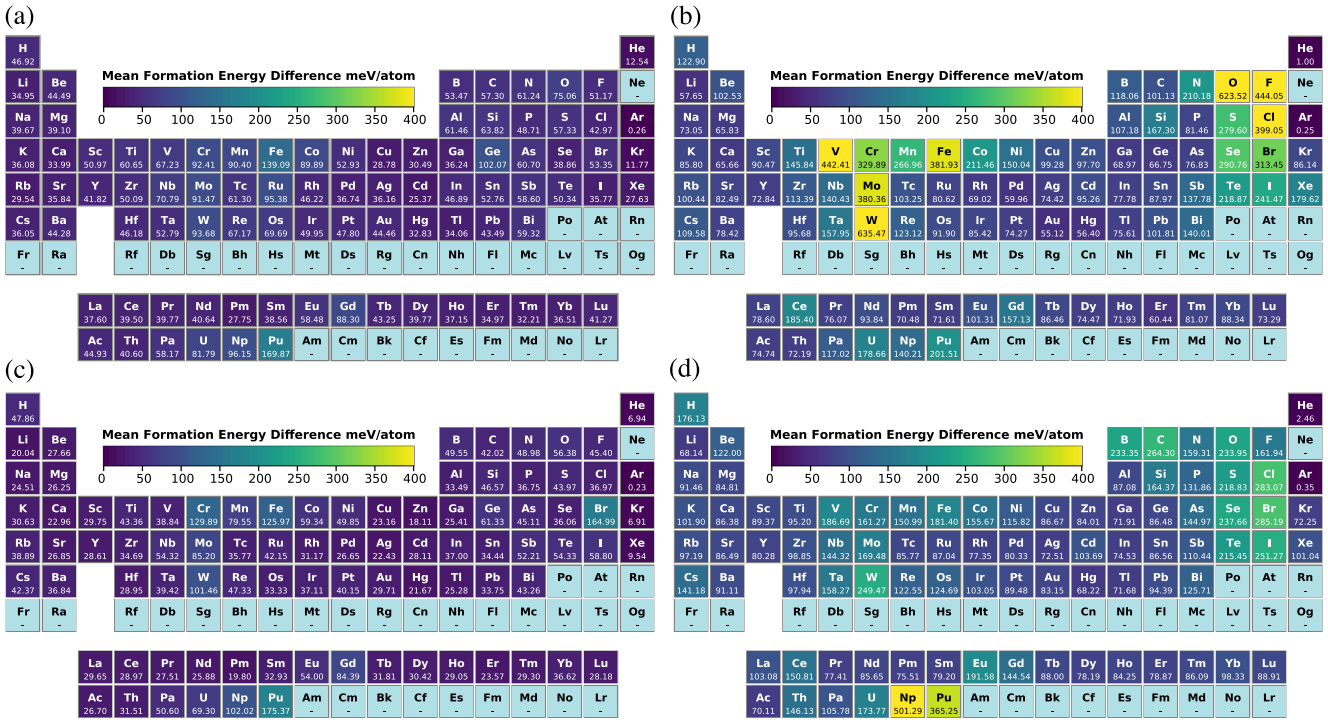

FIGURE 4 Heatmaps of the periodic table for the absolute values of ΔE_{form} for the one-shot calculations from the different uMLIPS: (a) CHGNet, (b) M3GNet, (c) MACE, and (d) ALIGNN. The figure has been produced with the pymatviz tool.[80]

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliothèque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

while displaying, together with CHGNet, the highest number of successful structural optimizations (or *converged* results), *that is*, 88.8%, while M3GNet and ALIGNN display 87.5% and 85.6%, respectively (Figure 3).

To examine the ability of the different uMLIPs to predict the cell geometry (*i.e.*, lattice parameters, angles, and volume) in the cell-relax calculations, we compare the uMLIPs results with the MP values, as done for the formation energy (Equation (4)). For example, for the volume we consider the relative difference as follows:

$$\Delta_{\text{rel}}V = 1 - \frac{V^{\text{uMLIP}}}{V^{\text{MP}}}, \tag{5}$$

where V^{MP} and V^{uMLIP} are the cell volume from MP and from the uMLIP cell-relax calculations, respectively. The distribution of $\Delta_{\text{rel}}V$ is reported in Figure 6. CHGNet and M3GNet outperform the other uMLIPs, with a narrower distribution of $\Delta_{\text{rel}}V$. ALIGNN performs the worst, while MACE has an intermediate performance. In Table 3, we report the Mean Absolute Relative Error (MARE) on the predicted lattice parameters, angles, and volumes. We observe that MACE shows similar performance with respect to CHGNet and M3GNet for a, b, c and α, β, γ, but its performance on the volume is significantly worse than the other two uMLIPs, as also shown in Figure 6. One possible origin of this behavior might be the compensation of the errors of the cell parameters in the case of M3GNet and CHGNet, which would not occur in MACE. Another reason could be the presence of more outliers (extreme large deviations) in the case of MACE, as documented from the histograms of the relative volume differences and lattice parameters differences provided in Figure S6 of the Supplemental Material.

A similar argument can be applied to the slightly higher MARE on the volume for CHGNet as compared to M3GNet (Table 3). ALIGNN shows larger errors compared to the other uMLIPs (Table 3). To further test the ability of uMLIPs to treat multi-element compounds, we perform 'one-shot' and 'cell-relax' calculations on 100 randomly chosen quinary materials. There are 4 *unconverged* cases (4%) for CHGNet and M3GNet, while 2 *unconverged* cases (2%) for MACE and ALIGNN. This seems to indicate that these uMLIPs can still be used for structural relaxations even for quinaries although additional investigations would be needed to assess the transferability in this part of the chemical space. In the Supplemental Material we report the comparison of formation energies and cell geometries of quinaries with the respective MP reference data (Figures S10 and S12, respectively), as done for the unary and binary compounds (Figures 5 and 6). This comparison indicates that MACE still gives the most reliable prediction on formation energies, while CHGNet shows the best performance for cell geometry predictions.

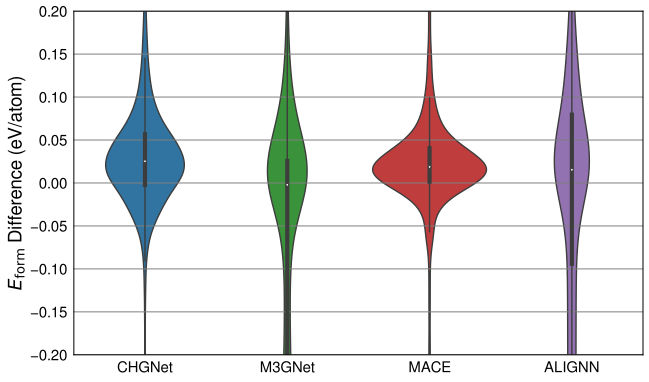

FIGURE 5 Violin plots showing the distribution of ΔE_{form} (Equation (4)) from one-shot calculations for the different uMLIPs.

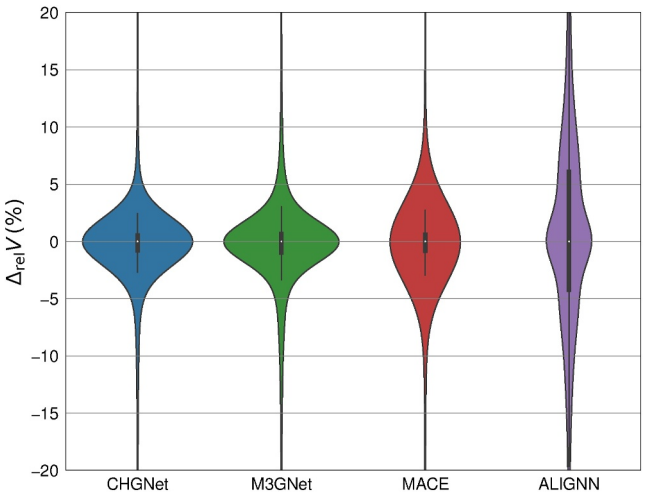

FIGURE 6 Violin plots of the relative difference between the MP cell volume and the one predicted by the different uMLIPs (Equation (5)). Only *converged* runs are included in the plot.

TABLE 2 Performance of the uMLIPs in predicting the formation energy for one-shot calculations.

uMLIP	MAE	RMSE	R^2
CHGNet	0.054	0.105	0.988
M3GNet	0.172	0.316	0.896
MACE	0.044	0.101	0.989
ALIGNN	0.137	0.2226	0.947

Note: As in Figure 5 for the energies, here MAE and RMSE are in eV/atom.

TABLE 3 Mean Absolute Relative Error (MARE in %) of the different uMLIPs in predicting the volume, lattice parameters, and angles.

uMLIP	Volume	a	b	c	α	β	γ
CHGNet	3.16	2.03	2.07	2.44	0.75	0.62	1.19
M3GNet	2.97	2.04	2.09	2.46	0.89	0.73	1.24
MACE	5.22	2.01	2.11	2.58	0.73	0.59	1.13
ALIGNN	7.85	3.42	3.42	3.61	0.94	0.86	1.32

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliotheque de l'EPFL , Wiley Online Library on [04/09/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.4 | Vibrational properties

In this section we analyze the capability of uMLIPs to reproduce the vibrational properties of crystalline materials using accurate ab initio results reported in a previous work[46] as a reference. From the structures in Ref.[46] we select those whose energy above hull is zero both in the MP database and from the uMLIP cell-relax calculations, leading to 101 structures. Phonons are computed using the finite displacement method as implemented in the PHONOPY package.[83,84] Further details on the protocol employed to compute phonons with uMLIPs are provided in Section 4. By exploiting the finite displacement method to compute phonons, this study indirectly probes the quality of the uMLIPs-calculated forces when atoms are slightly displaced away from the equilibrium positions. From a methodological point of view we note that, already when ab initio engines are used, the accuracy of phonon calculations from the finite displacement method is rather sensitive to the quality of the force calculations in the supercell. Since by construction these forces are less accurate when calculated from uMLIP models than from ab initio methods, it is reasonable to expect some nonnegligible discrepancy between uMLIPs and ab initio phonon calculations. Also, as discussed in more detail in Section 4, the MLIP architectures at the basis of the universal models considered in this study cannot predict the long-range dipolar contributions to the interatomic force constants, needed to obtain a reliable Fourier phonon interpolation in the case of polar materials.[85,86] The uMLIP-calculated forces should therefore be augmented with the ML electronic dielectric tensor and the Born effective charges, which are not available at present, leading us to augment the uMLIPs forces with ab initio[47,48] values for these quantities (see Section 4 for further details).

For the above reasons, when comparing uMLIPs results with ab initio data, we choose a rather generous metrics, *that is*, the MAE between the uMLIPs and the ab initio phonon band structures computed along a high-symmetry **q**-path[87]:

$$\mathrm{MAE} = \frac{1}{N_{\mathbf{q}}} \sum_{\mathbf{q}\nu} |\omega_{\mathbf{q}\nu}{}^{\mathrm{uMLIP}} - \omega_{\mathbf{q}\nu}{}^{\mathrm{DFPT}}|. \qquad (6)$$

In Equation (6), $\omega_{\mathbf{q}\nu}$ is the phonon energy in meV, ν is the branch index, $N_{\mathbf{q}}$ is the number of wavevectors used to sample the **q**-path, and $\omega_{\mathbf{q}\nu}^{\mathrm{uMLIP}}$, $\omega_{\mathbf{q}\nu}^{\mathrm{DFPT}}$ are the phonon energies computed using the uMLIP model and the density-functional perturbation theory (DFPT) with ABINIT,[46] respectively. Table 4 reports the minimum, maximum, and average values of the MAE in the phonon band structures obtained from the different uMLIPs (Equation (6)), while the MAE distribution is given in panel (a) of Figure 7.

From panel (a) of Figure 7 and Table 4, we observe a nonideal agreement between uMLIPs-calculated and ab initio-calculated vibrational properties. In addition to the fundamental reasons discussed above, in this particular study a source of discrepancy could come from the fact that Ref.[46] exploits the PBEsol exchange-correlation functional while the uMLIPs have been trained using PBE[25] data with the inclusion of U-corrections[61–63] for certain systems (in the case of

TABLE 4 Minimum, maximum, and average MAE in meV for the phonon band structures computed from different uMLIPs.

uMLIP	MIN_MAE	MAX_MAE	MEAN_MAE
CHGNet	0.82	37.34	8.12
M3GNet	0.74	40.20	10.41
MACE	0.31	17.22	3.71
ALIGNN	5.60	75.38	29.36

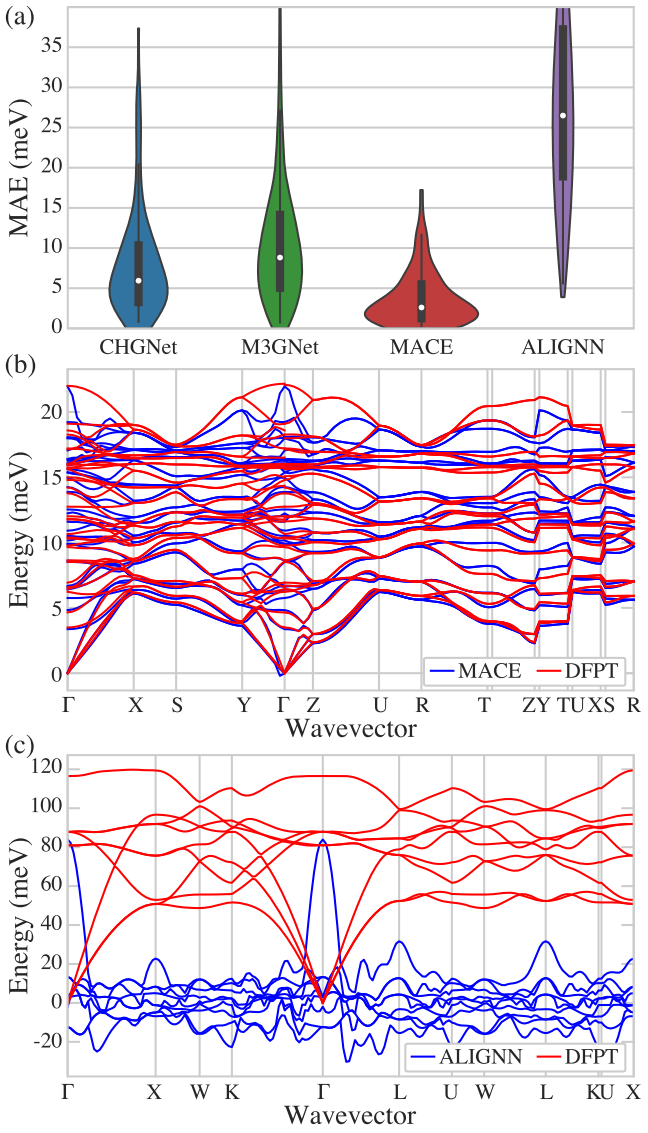

FIGURE 7 (a) Violins plots of the MAE (Equation (6)) on the computed phonon band structures from uMLIPs and DFPT. (b) Comparison of the phonon band structures computed with DFPT and MACE for the compound (mp-567744: $SrBr_2$) with the smallest MAE (0.3 meV). (c) Comparison of the phonon band structures computed with DFPT and ALIGNN for the compound (mp-1569: Be_2C) with the largest MAE (75.4 meV).

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliothèque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

CHGNet and M3GNet) or PBE[25] data without the inclusion of U-correction (in the case of MACE) or optB88vdW[32] data (in the case of ALIGNN). In panel (b) of Figure 7, we report the phonon band structure for the system with the lowest MAE (0.312 meV, obtained from MACE, Table 4), which is $SrBr_2$ (mp-567744). Conversely, in panel (c) of Figure 7, we report the phonon band structure for the system with the highest MAE (75.375 meV, obtained from ALIGNN, Table 4), which is Be_2C (mp-1569). For the latter, the uMLIP predicts vibrational instabilities that are not observed in the ab initio results. However, we also note that even for the lowest MAE material (panel (b) of Figure 7), the high-energy region of the phonon band structure is still far from being accurate.

To summarize, our results indicate that presently available uMLIPs can predict ab initio vibrational properties with a typical error of 3.71 meV in the best-case scenario. This value should be considered as a lower bound as the discrepancy is expected to increase if the electronic dielectric tensor and the Born effective charges were predicted with ML techniques. In the opinion of the authors, the predictive behavior of uMLIPs for vibrational properties can be improved by training new uMLIPs with larger weights for the forces loss function but it is also clear that for accurate ML-based predictions in polar materials, one needs uMLIPs capable of inferring the long-range part of the dynamical matrix. All this being said, we believe that uMLIPs represent an efficient and promising approach to perform an initial screening for vibrational and thermodynamic properties, especially in a high-throughput context in which high accuracy is not necessarily needed.

3 | CONCLUSIONS

We present a systematic assessment of various universal machine-learning interatomic potentials (uMLIPs) by investigating their capability to reproduce ab initio results for several important physical properties such as equation of state, relaxed geometries, formation energies, and vibrational properties of an extensive set of crystalline materials. Among the considered uMLIPs, we find that MACE shows superior accuracy in predicting formation energies and vibrational properties, and CHGNet and M3GNet are outstanding for relaxed geometry predictions. MACE and CHGNet show superior performance on the formation energy prediction along the periodic table, especially when considering systems containing transition metal elements, chalcogens, and halogens (with Br being accurately reproduced only by CHGNet). M3GNet exhibits relatively high errors, leading to a lower R^2 value compared to the other uMLIPs in predicting formation energies. Despite this, its ability to accurately predict volume and lattice parameters remains high, and it has intermediate performance in predicting vibrational properties. For what concerns ALIGNN, despite being trained with a different dataset, it still shows relatively good results when predicting formation energies. However, it also reveals to be the most problematic model when performing geometry optimization at variable cell and for predicting vibrational properties. These results underscore the need for further optimization and training of the currently available uMLIPs to fully exploit the capability of ML techniques across a broader range of applications. The choice of a particular uMLIP for specific applications should take into account an appropriate balance between accuracy and computational efficiency. Future work should aim at enhancing the performance of these potentials further, particularly focusing on areas where current uMLIPs exhibit limitations such as more accurate prediction of forces and stresses or the capability of learning Born effective charges and electronic dielectric tensors that are crucial for the vibrational properties of polar materials. Our work will hopefully pave the way toward a more systematic assessment of uMLIPs in different scenarios and the establishment of a standardized benchmark set that can be used to gauge the precision and transferability of uMLIPs.

4 | METHODS

The calculations are performed with the AbiPy package,[47,48] more specifically the `abiml.py` script providing a unified interface that allows one to perform different types of calculations such as structural relaxations, molecular dynamics, or NEB using the algorithms implemented in ASE[88] and different uMLIPs as calculators. The following package versions are used to produce the results reported in this work. Python: 3.11, Pymatgen: 2023.7.17, AbiPy: 0.9.6, ABINIT: 9.8.4, CHGNet: 0.3.2, ALIGNN-FF: alignn 2023.10.1, MatGL: 1.0.0, MACE-MP-0: mace-torch 0.3.4, and ASE: 3.22.1.

Structural relaxations are performed using the ASE optimizer, employing the BFGS algorithm. To ensure good trade-off between accuracy and computational cost, the stopping criterion fmax is set to 0.1 eV/Å. If the stopping criterion is not met at a maximum number of iterations set to 150, the calculation is considered unconverged. This is justified by the fact that all the initial structures in our dataset are taken from the Materials Project,[24] where they have been already relaxed with VASP[52] with the PBE functional,[25] and the same structures are supposed to be included in the set used to train the uMLIPs (with the exception of ALIGNN). It should be pointed out that for each compound the total energy obtained from the Materials Project[24] is obtained with energy corrections.[61–63] Therefore, we apply the same correction according to the methods in `compatibility.py` from the Pymatgen package[81,82] to MACE since it was trained with Materials Project uncorrected energies.

For the phonon calculations we first perform a structural relaxation with the uMLIPs from the crystalline structures employed in Ref.[46]. The atomic positions are relaxed at fixed cell parameters in order to avoid spurious effects due to

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliothèque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the change of the lattice parameters with respect to the reference DFT results. The relaxed configuration is then used to compute vibrational properties using the finite displacement method implemented in PHONOPY[83,84] with a displacement of 0.01 Å. In each calculation, the real-space supercell is matched with the **q**-mesh employed in Ref.[46] to compute the dynamical matrix with the DFPT part of ABINIT.[47,48] It is noteworthy that the uMLIPs employed in this work lack the capability to predict the electronic dielectric tensor ϵ^{∞} and the Born effective charge tensor $\mathbf{Z}^*_{\kappa}$ where κ is the index of the atom in the unit cell. As discussed in Refs.[85,86], these quantities are needed to model the long-range dipolar contribution to the interatomic force constants in polar materials. This term is indeed responsible for the LO-TO splitting as well as for the nonanalytical behavior of the vibrational spectrum near $\mathbf{q} = 0$. Its correct numerical treatment is therefore crucial to obtain an accurate Fourier interpolation of the dynamical matrix at arbitrary **q**-points. For this reason, in all PHONOPY calculations we use the ab initio values of ϵ^{∞} and $\mathbf{Z}^*_{\kappa}$ obtained with ABINIT to model long-range interactions. The ab initio phonon dispersions are obtained using the `anaddb` post-processing tool[46–48] using the DDB files retrieved from the MP database to compare the different phonon dispersions between uMLIPs and ab initio DFPT results.

Figure 1 and the heatmaps of the periodic table in the Supplemental Material (Figures S1-S4) have been produced using the ab initio results and the open source python scripts available in the acwf-verification-scripts github repository.

AUTHOR CONTRIBUTIONS

Gian-Marco Rignanese designed the research topic and coordinated the whole project. All authors provided the ideas underlying this work, contributed to its development, and discussed the findings reported in the paper. All authors took part to the writing and reviewing of the paper, and to the final approval of its completed version. Haochen Yu and Matteo Giantomassi executed the calculations presented. Matteo Giantomassi developed and maintained the software infrastructure.

ACKNOWLEDGMENTS

The authors acknowledge useful discussions about the correct use of the different uMLIPs with Yuan Chiang, Kamal Choudhary, Bowen Deng, Tsz Wai Ko, and Shyue Ping Ong. Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL), and the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI). This work is supported by the National Key Research and Development Program of China (2022YFE0141100 and 2023YFB3003005).

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in mlpot_benchmarks at https://github.com/modl-uclouvain/mlpot_benchmarks.

ORCID

Giuliana Materzanini https://orcid.org/0000-0001-5845-4679
Gian-Marco Rignanese https://orcid.org/0000-0002-1422-1205

REFERENCES

1. Frenkel D, Smit B. *Understanding Molecular Simulation: From Algorithms to Applications*. Elsevier; 2023.
2. Hohenberg P, Kohn W. Inhomogeneous electron gas. *Phys Rev*. 1964;136(3B):B864-B871.
3. Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. *Phys Rev*. 1965;140(4A):A1133-A1138.
4. Payne MC, Teter MP, Allan DC, Arias T, Joannopoulos AJ. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Rev Mod Phys*. 1992;64(4):1045-1097.
5. Marzari N, Ferretti A, Wolverton C. Electronic-structure methods for materials design. *Nat Mater*. 2021;20(6):736-749.
6. Goddard I. *Classical Force Fields and Methods of Molecular Dynamics*. Springer International Publishing; 2021:1063-1072.
7. Zuo Y, Chen C, Li X, et al. Performance and cost assessment of machine learning interatomic potentials. *J Phys Chem A*. 2020;124(4):731-745.
8. Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. *Adv Mater*. 2019;31(46).
9. Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys Rev Lett*. 2007;98(14):146401.
10. Bartók AP, Payne MC, Kondor R, Csányi G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys Rev Lett*. 2010;104(13):136403.
11. Thompson A, Swiler L, Trott C, Foiles S, Tucker G. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J Comput Phys*. 2015;285:316-330.
12. Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet – a deep learning architecture for molecules and materials. *J Chem Phys*. 2018;148(24).
13. Drautz R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys Rev B*. 2019;99(1):014104.
14. Lilienfeld VOA, Burke K. Retrospective on a decade of machine learning for chemical discovery. *Nat Commun*. 2020;11(1):4895.
15. Batzner S, Musaelian A, Sun L, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun*. 2022;13(1):2453.
16. Ko TW, Ong SP. Recent advances and outstanding challenges for machine learning interatomic potentials. *Nat Comput Sci*. 2023;3(12):998-1000.
17. Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. *Chem Rev*. 2021;121(16):10073-10141.
18. Geiger M, Smidt T. e3nn: Euclidean neural networks. arXiv preprint arXiv:2207.09453. 2022.
19. Batatia I, Kovacs DP, Simm G, Ortner C, Csanyi G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022). 35*. NeurIPS. Curran Associates, Inc.; 2022:11423-11436.

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliotheque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

20. Behler J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys Chem Chem Phys*. 2011;13(40):17930-17955.
21. Behler J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew Chem Int Ed*. 2017;56(42):12828-12840.
22. Behler J. Four generations of high-dimensional neural network potentials. *Chem Rev*. 2021;121(16):10037-10072.
23. Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater*. 2019;31(9):3564-3572.
24. Jain A, Ong SP, Hautier G, et al. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater*. 2013;1(1).
25. Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. *Phys Rev Lett*. 1996;77(18):3865-3868.
26. Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. *Nat Comput Sci*. 2022;2(11):718-728.
27. Deng B, Zhong P, Jun K, et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat Mach Intell*. 2023;5(9):1031-1041.
28. https://github.com/materialsvirtuallab/matgl. Accessed: 2024-02-29.
29. Qi J, Ko TW, Wood BC, Pham TA, Ong SP. Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling. *NPJ Comput Mater*. 2024;10(1):43.
30. Choudhary K, DeCost B, Major L, Butler K, Thiyagalingam J, Tavazza F. Unified graph neural network force-field for the periodic table: solid state applications. *Digit Discov*. 2023;2(2):346-355.
31. Choudhary K, Garrity KF, Reid ACE, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *NPJ Comput Mater*. 2020;6(1):173.
32. Klimeš J, Bowler DR, Michaelides A. Chemical accuracy for the van der Waals density functional. *J Phys Condens Matter*. 2009;22(2):022201.
33. Batatia I, Benner P, Chiang Y, et al. A foundation model for atomistic materials chemistry. arXiv preprint arXiv:2401.00096. 2023.
34. Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. *Nature*. 2023;624(7990):80-85.
35. Takamoto S, Shinagawa C, Motoki D, et al. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nat Commun*. 2022;13(1):2991.
36. Takamoto S, Okanohara D, Li QJ, Li J. Towards universal neural network interatomic potential. *J Materiomics*. 2023;9(3):447-454.
37. Takamoto S, Izumi S, Li J. TeaNet: universal neural network interatomic potential inspired by iterative electronic relaxations. *Comput Mater Sci*. 2022;207:111280.
38. Smith JS, Isayev O, Roitberg AE. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem Sci*. 2017;8(4):3192-3203.
39. Smith JS, Nebgen BT, Zubatyuk R, et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun*. 2019;10(1):2903.
40. Zubatyuk R, Smith JS, Leszczynski J, Isayev O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci Adv*. 2019;5(8).
41. Lopanitsyna N, Fraux G, Springer MA, De S, Ceriotti M. Modeling high-entropy transition metal alloys with alchemical compression. *Phys Rev Mater*. 2023;7(4):045802.
42. Riebesell J, Goodall RE, Jain A, Benner P, Persson KA, Lee AA. Matbench Discovery–An evaluation framework for machine learning crystal stability prediction. arXiv preprint arXiv:2308.14920. 2023.
43. Focassio B, Freitas LPM, Schleder GR. Performance assessment of universal machine learning interatomic potentials: challenges and directions for materials' surfaces. arXiv preprint arXiv:2403.04217. 2024.
44. Deng B, Choi Y, Zhong P, et al. Overcoming systematic softening in universal machine learning interatomic potentials by fine-tuning. arXiv preprint arXiv:2405.07105. 2024.
45. Bosoni E, Beal L, Bercx M, et al. How to verify the precision of density-functional-theory implementations via reproducible and universal workflows. *Nat Rev Phys*. 2023;6(1):45-58.
46. Petretto G, Dwaraknath S, Miranda H PC, et al. High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci Data*. 2018;5(1):180065.
47. Gonze X, Amadon B, Antonius G, et al. The Abinit project: impact, environment and recent developments. *Comput Phys Commun*. 2020;248:107042.
48. Romero AH, Allan DC, Amadon B, et al. ABINIT: overview and focus on selected capabilities. *J Chem Phys*. 2020;152(12).
49. Hamann DR. Optimized norm-conserving Vanderbilt pseudopotentials. *Phys Rev B*. 2013;88(8):085117.
50. Setten vM, Giantomassi M, Bousquet E, et al. The PseudoDojo: training and grading a 85 element optimized norm-conserving pseudopotential table. *Comput Phys Commun*. 2018;226:39-54.
51. Perdew JP, Ruzsinszky A, Csonka GI, et al. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys Rev Lett*. 2008;100(13):136406.
52. Kresse G, Joubert D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys Rev B*. 1999;59(3):1758-1775.
53. Chen C, Zuo Y, Ye W, Li X, Ong SP. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat Comput Sci*. 2021;1(1):46-53.
54. Choudhary K, DeCost B, Chen C, et al. Recent advances and applications of deep learning methods in materials science. *NPJ Comput Mater*. 2022;8(1):59.
55. Batatia I, Batzner S, Kovács DP, et al. The design space of e (3)-equivariant atom-centered interatomic potentials. arXiv preprint arXiv:2205.06643. 2022.
56. Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. *NPJ Comput Mater*. 2021;7(1):185.
57. Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY. A comprehensive survey on graph neural networks. *IEEE Transact Neural Networks Learn Syst*. 2020;32(1):4-24.
58. Asif NA, Sarker Y, Chakrabortty RK, et al. Graph neural network: a comprehensive review on non-euclidean space. *IEEE Access*. 2021;9:60588-60606.
59. Mueller T, Hernandez A, Wang C. Machine learning for interatomic potential models. *J Chem Phys*. 2020;152(5).
60. Tersoff J. New empirical approach for the structure and energy of covalent systems. *Phys Rev B*. 1988;37(12):6991-7000.
61. Wang A, Kingsbury R, McDermott M, et al. A framework for quantifying uncertainty in DFT energy corrections. *Sci Rep*. 2021;11(1):15496.
62. Jain A, Hautier G, Ong SP, et al. Formation enthalpies by mixing GGA and GGA calculations. *Phys Rev B*. 2011;84(4):045115.
63. Materials Project. Accessed: 2024-06-15. https://docs.materialsproject.org/methodology/materials-methodology/calculation-details/gga+u-calculations/hubbard-u-values
64. Reed J, Ceder G. Role of electronic structure in the susceptibility of metastable transition-metal oxide structures to transformation. *Chem Rev*. 2004;104(10):4513-4534.
65. Dusson G, Bachmayr M, Csányi G, et al. Atomic cluster expansion: completeness, efficiency and stability. *J Comput Phys*. 2022;454:110946.
66. Drautz R. Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer. *Phys Rev B*. 2020;102(2):024104.
67. Wang M, Zheng D, Ye Z, et al. Deep graph library: a graph-centric, highly-performant package for graph neural networks. arXiv preprint arXiv:1909.01315. 2019.
68. CHGNet pretrained. Accessed: 2024-06-15. https://github.com/CederGroupHub/chgnet/blob/main/chgnet/pretrained/0.3.0/chgnet_0.3.0_e29f68s314m37.pth.tar
69. Materials Project Trajectory (MPtrj) Dataset. figshare. Accessed:2024-06-15.

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliotheque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

70. M3GNet-DIRECT pretrained. Accessed: 2024-06-15. https://github.com/materialsvirtuallab/matgl/tree/main/pretrained_models/M3GNet-MP-2021.2.8-DIRECT-PES
71. ACEsuit/mace-mp. Accessed: 2024-06-15. https://github.com/ACEsuit/mace-mp/releases/tag/mace_mp_0
72. ACEsuit/mace-mp. Accessed: 2024-06-15. https://github.com/ACEsuit/mace/blob/main/mace/calculators/foundations_models.py
73. ALIGNN pretrained. Accessed: 2024-06-15. https://github.com/usnistgov/alignn/blob/main/alignn/ff/ff.py
74. ALIGNN pretrained. Accessed: 2024-06-15. https://figshare.com/ndownloader/files/41583594
75. ALIGNN pretrained. Accessed: 2024-06-15. https://github.com/usnistgov/alignn/blob/main/alignn/models/alignn_atomwise.py
76. Wortmann D, Michalicek G, Hilgers R, et al. FLEUR; 2023.
77. Blaha P, Schwarz K, Tran F, Laskowski R, Madsen GKH, Marks LD. WIEN2k: an APW+lo program for calculating the properties of solids. *J Chem Phys*. 2020;152(7):074101.
78. Blöchl PE. Projector augmented-wave method. *Phys Rev B*. 1994;50(24):17953-17979.
79. Lejaeghere K, Bihlmayer G, Björkman T, et al. Reproducibility in density functional theory calculations of solids. *Science*. 2016;351(6280):aad3000.
80. Riebesell J. Pymatviz: visualization toolkit for materials informatics; 2022.
81. Ong SP, Richards WD, Jain A, et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput Mater Sci*. 2013;68:314-319.
82. Pymatgen Compatibility corrections entry. Accessed: 2024-06-24. https://github.com/materialsproject/pymatgen/blob/master/pymatgen/entries/compatibility.py
83. Togo A, Chaput L, Tadano T, Tanaka I. Implementation strategies in phonopy and phono3py. *J Phys Condens Matter*. 2023;35(35):353001.
84. Togo A. First-principles phonon calculations with phonopy and phono3py. *J Phys Soc Jpn*. 2023;92(1):012001.
85. Gonze X, Lee C. Dynamical matrices, Born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory. *Phys Rev B*. 1997;55(16):10355-10368.
86. Baroni S, Gironcoli dS, Dal Corso A, Giannozzi P. Phonons and related crystal properties from density-functional perturbation theory. *Rev Mod Phys*. 2001;73(2):515-562.
87. Setyawan W, Curtarolo S. High-throughput electronic band structure calculations: challenges and tools. *Comput Mater Sci*. 2010;49(2):299-312.
88. Hjorth Larsen A, Jørgen Mortensen J, Blomqvist J, et al. The atomic simulation environment—a Python library for working with atoms. *J Phys Condens Matter*. 2017;29(27):273002.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Yu H, Giantomassi M, Materzanini G, Wang J, Rignanese G-M. Systematic assessment of various universal machine-learning interatomic potentials. *MGE Advances*. 2024;e58. https://doi.org/10.1002/mgea.58

29409497, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mgea.58 by Gian-Marco Rignanese - Bibliothèque de l'EPFL , Wiley Online Library on [04/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License