Systematic atomic structure datasets for machine learning potentials: Application to defects in magnesium

Marvin Poul,[*] Liam Huber,[†] Erik Bitzek,[‡] and Jörg Neugebauer[§]
Department of Computational Materials Design, Max-Planck-Institut für Eisenforschung GmbH, D-40327 Düsseldorf, Germany

(Received 26 July 2022; revised 15 December 2022; accepted 8 February 2023; published 13 March 2023)

We present a physically motivated strategy for the construction of training sets for transferable machine learning interatomic potentials. It is based on a systematic exploration of all possible space groups in random crystal structures, together with deformations of cell shape, size, and atomic positions. The resulting potentials turn out to be unbiased and generically applicable to studies of bulk defects without including any defect structures in the training set or employing any additional active learning. Using this approach we construct transferable potentials for pure magnesium that reproduce the properties of hexagonal closed packed (hcp) and body centered cubic (bcc) polymorphs very well. In the process we investigate how different types of training structures impact the properties and the predictive power of the resulting potential.

DOI: 10.1103/PhysRevB.107.104103

I. INTRODUCTION

A key concept in materials science to design materials with tailored properties is defect engineering. In order to successfully employ this concept, one needs a detailed understanding of the relationship between crystal defects on the atomistic scale and their influence on macroscopic materials properties. Until now this understanding has been provided to a large extent by density functional theory (DFT) calculations especially when investigating, e. g., the thermodynamic stability of materials phases and simple, isolated defects such as vacancies [1], dislocation arrays [2], or high-symmetry planar defects [3,4]. However, successful defect engineering must include most of the macroscopic and microscopic degrees of freedom of the defects—or risk missing potential candidate states. Especially in extended defects such as grain boundaries this defect phase space is very large, making it unfeasible to scan with DFT due to its high computational cost and system size restrictions. Together with recent interest in defect phase diagrams [5,6] this motivates us to develop a machine learning potential specifically aimed at a transferable description of defects. To this end, we will apply the moment tensor potential (MTP) methodology [7], and rigorously examine the impact of training data on the quality and performance of the resulting potentials. The approach and the detailed analysis and discussion are however general and can be applied to any machine learning (ML) potential methodology.

Classical potentials are often trained on a set of properties that they ought to reproduce, e.g., relative phase stabilities, surface energies, and elastic properties. The more data hungry machine learning potentials instead use large sets of reference structure with energies, forces, and potentially stresses calculated with quantum mechanical models like DFT. These reference structures are generally constructed starting from equilibrium structures of interest, which are then perturbed in various ways to sample the energy landscape. This approach can work very well, but can lead to failure of the potential when relevant structures are missing. Another approach recently presented is to combine active learning and some form for structure generation (randomly, by molecular dynamics or Monte Carlo simulations) [8–11]. By starting from random environments, bias is removed from the training data and then a given active learning algorithm is in control of selecting structures to add to the training set. For example Smith *et al.* [9] demonstrates that this works very well for aluminum and it allowed them to obtain a robust potential that predicted the correct relative phase stabilities in a wide temperature and pressure window without any human guidance. Of these approaches Bernstein *et al.* [10] appears to be most closely related to our approach. The major difference is the starting point of the generation procedure. Where they start with completely random distribution of atoms and enforce only a few symmetry operations, we will systematically include most space groups. Additionally they use an on the fly fitted potential in an iterative scheme to minimize cells whereas we will rely on DFT.

There are also parallels to the paper of Podryabinkin *et al.* [11]. The authors' objective there is to predict the stable crystal structures of elements and employ active learning to provide candidate structures that can be investigated with DFT in a reasonable time. For the explicit purpose of predicting crystal structures they show that this approach works very well.

A challenge in constructing potentials that accurately describe defects is that atoms at or near the defect can have

[*]poul@mpie.de
[†]huber@mpie.de
[‡]bitzek@mpie.de
[§]neugebauer@mpie.de

structures that are far away from any low-energy bulk structures. These atoms represent spatially highly localized regions of high energy that are not captured when including only low-energy structures. We will discuss in this paper how to construct and utilize structures that are not energetically near the equilibrium structure. An approach in that direction is the recent paper from M. Karabin *et al.* [12] and Montes de Oca Zapiain *et al.* [13]. Their paper aims to sample the descriptor space of the targeted potential model as widely and unbiased as possible. For this they define a *descriptor entropy* that favors structures with different local environments in the same cell and then maximize or minimize this entropy in a simple annealing procedure. They show that this yields significantly wider coverage of descriptor space than sets drawn from high temperature MD with a simple size exclusion potential or traditionally constructed training sets, but still includes crystallographic relevant bulk phases and phases. Since this scheme makes no reference to structure prototypes or crystallography it is a completely unbiased procedure in this sense. Instead it relies on the quality of the underlying descriptor set. This means that changing the descriptor set will produce a different training set. A potential drawback of their method is also the large number of structures that are generated: up to 200 000 structures with 32–40 atoms each. In contrast to this we aim here to provide a method for a smaller, descriptor, or potential agnostic training set, generated purely based on the constraints placed on atomic positions by the space group symmetries that still accurately captures the necessary structures.

We structure the paper as follows: First the construction of the data set in Sec. II A, which is a key concept in this study. Then a brief review of moment tensor potentials in Sec. II D. In Secs. II E and III A we discuss choosing cutoffs and the fitting of the potentials. Afterwards we verify the potentials on defects and analyze in detail the influence of training data in Secs. III C and III E. This analysis demonstrates the performance of our main idea. We close with a brief comparison with active learning in Sec. III F before we summarize our findings in Sec. IV.

II. METHODS

A. Training set construction

We construct several different training subsets, each of which explores slightly different regions of phase space that have clear physical interpretation. We generate these sets in an iterative, multistage process. The foundational dataset for this process, which feeds into all further subsets, consists of random (periodic) crystal structures obtained from RANDSPG [14]. We generate these structures with one to ten magnesium atoms per cell, asking for all possible space groups, 1–230, and allowing volumes ±10% around the equilibrium atomic volume of hcp Mg of $\Omega_0 = 22.87\,\text{Å}^3/\text{atom}$, obtained from DFT at $T = 0\,\text{K}$. We label this the RANDSPG set. Note that this approach requires as only input parameters the volume range and number of atoms considered in super cells for training. An automation and extension to other materials is therefore straightforward. Applying this approach, not all space groups are present, because some symmetries are not consistent with the allowed volume range or lead to structures with very inhomogeneous particle distribution. As a check the space groups have been determined with SPGLIB [15]. Figure 1 shows the frequency of each space group and crystal system in our initial data set. While not all systems are equally present, there is a sizable number of structures available for each.

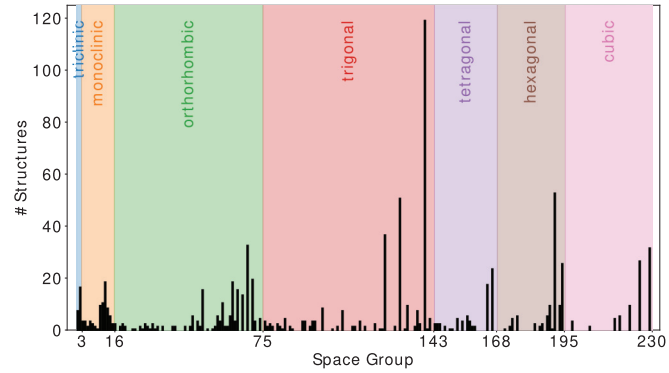

FIG. 1. Frequency of crystal systems in the RANDSPG set (see text).

From this starting point, we then successively minimized the volume, cell shape, and the internal coordinates independently using VASP [16,17]. These calculations are done at low convergence parameters, since they only serve to bring the structures near the equilibrium structures and the energies from these runs do not enter the fitting routines. We call these sets VOLMIN, CELLMIN, and INTMIN, respectively.

Naturally the minimization generally leads to higher symmetry structures exploring a reduced phase space. In fact, in the fully internally relaxed set some space groups are no longer present. This reduces the inherent dimensionality of the minimized training sets, but we have not attempted to filter structures that relaxed into the same minima. It is also noteworthy that volume minimization—particularly for the structures with more atoms per unit cell—can lead to quasi 1D and 2D structures. This gives the potentials the opportunity to see structures resembling surfaces and isolated atoms even though in the construction setup we do not explicitly enforce such structures [18].

As a final step in our process for creating training data, the structures from INTMIN are disturbed by either a random triaxial (TRIAX, up to 80%) strain, a combination of random shear strains (SHEAR, up to 80%), or by random displacements of atoms combined with a small random strain tensor (RATTLE, 0.5 Å mean displacement and up to 5% strain) [19]. For each structure in INTMIN these modifications are applied five times, resulting in five times more structures for the respective derived training sets, TRIAX, SHEAR and RATTLE than in the INTMIN set.

Figure 2 gives a conceptual overview of this procedure. During each step some structures resist DFT calculations due to excessively close atoms or deformed cells, which we then discard. Also shown in Fig. 2 are the number of structures (top number) and atoms (i. e., atomic environments, bottom number) in each structure set. We will examine the performance

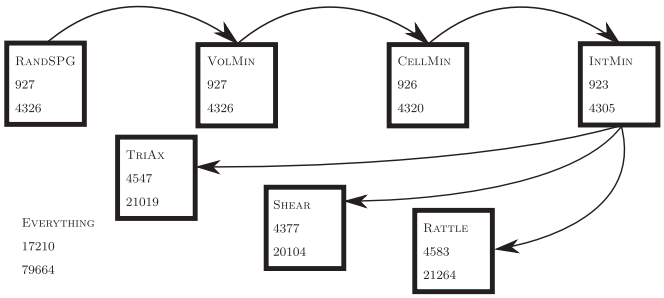

FIG. 2. Schematic procedure to generate the training sets. First number gives the number of structures in each set, the second the total number of atoms.

of potentials fitted to each of these sets compared to potentials fitted to the set of all structures, EVERYTHING.

Training sets for multi component potentials

The procedure can be extended to binary or ternary compounds by using RANDSPG [14] to generate structures with various concentrations. This would naturally increase the number of structures substantially. It is not clear at this point how dense in concentration space such training sets would need to be or how well potentials would be able to interpolate or extrapolate between concentrations. In this case it might be necessary to combine the data generation procedure shown here with data selection strategies from active learning schemes. However, for now we focus on unaries and leave the exploration of alloys to future work.

B. Test Data

In general in machine learning it is custom to reserve some percentage of the reference data in a hold-out set on which to test the final model. We explicitly decide against doing this and will use the full data set for fitting and report only the errors on this (training errors). The reason is that to test the potentials we construct completely fresh data closer to the application domain of the potential, such as defect structures, phonon, and elastically strained calculations. The results of this testing are discussed in Sec. III C. We chose to do this because the meaningfulness of test errors depends to a large extent on the sampling of reference structures. If the reference structures are not drawn evenly from the full space that the potential will be applied on, the train-test errors can give the impression that the potential is fitted well, even though there are gaps in the potential, simply because the relevant structures never entered the reference structure set. This is discussed in more detail, e. g., in a review from J. Behler [[8], Fig. 10 and the discussion Sec. 4]. Since we now rely on these separate completely out-of-fold structures for testing we opt to include the full reference data set in the fitting to provide more learning opportunities to the potential.

C. DFT data generation

All training data is generated using VASP [16,17] using the projector augmented wave (PAW) method [20,21] and the PBE [22] functional with the standard s-valent pseudopotential from the VASP [16,17] distribution. Γ-centered k meshes with $27\times27\times27$ k points and plane wave energy cutoff of 550 eV are used. While the structures vary in volume, we keep the k points constant to avoid discontinuities in the potential energy surface. The chosen k-point setting corresponds to a k-mesh spacing of 0.06 Å^{-1} [23]. All calculations used the Methfessel-Paxton occupation smearing scheme of order 1 with a smearing parameter of 0.2 eV [24]. By convergence testing we find the energies to be converged to 0.6 meV and the forces to 7×10^{-5} eV/Å. These values represent the mean error of the training calculations with respect to a sample of 50 structures of each training set (350 in total) calculated at a $37\times37\times37$ k-point mesh and a plane wave cutoff of 687.5 eV. DFT data for the verification calculations is generated with the same parameters except for large grain boundary and surface structures where we use a k-mesh spacing of 0.05 Å^{-1}.

A small number of calculations fail during the minimizations and the final training set generation. They are automatically discarded and do not enter subsequent steps. Figure 2 shows that their total number is small, however.

D. Moment tensor potentials

Moment tensor potentials (MTP) are machine learning interatomic potentials originally introduced by A. Shapeev [25]. We will briefly review this formalism here, but leave the details to to the original authors [7].

The total energy E^{MTP} of any atomic structure is constructed from contributions of neighbors around each atom, which are expanded in linear basis functions

$$E^{\text{MTP}} = \sum_i^N V(\mathfrak{n}_i) = \sum_i^N \sum_\alpha \xi_\alpha B_\alpha(\mathfrak{n}_i), \tag{1}$$

where $\mathfrak{n}_i$ is the atomic environment around atom i, N is the total number atoms, B_α are the descriptor basis functions, and ξ_α the linear expansion coefficients, which are determined during the fitting procedure, and α runs over all basis functions [26]. The descriptor basis functions are defined as contractions of the *moment tensors* $M_{\mu,\nu}$ defined in the single-component case as

$$M_{\mu,\nu}(\mathfrak{n}_i) = \sum_{j\in\mathfrak{n}_i} f_\mu(|r_{ij}|)\overbrace{r_{ij}\otimes\cdots\otimes r_{ij}}^{\nu \text{ times}} \tag{2}$$

where r_{ij} is the vector connecting the i th and j th atoms and $\otimes$ is the outer product on vectors and tensors. The radial functions f_μ are expanded in an orthogonal polynomial basis and contain an outer cutoff R_c, such that their derivatives go smoothly to zero. This encodes the locality assumption generally made in interatomic potentials. The polynomial expansion coefficients of the radial functions are also additional fitting parameters. The authors then *define* the *level* of an MTP as

$$\text{lev}\, M_{\mu,\nu} = 2 + 4\mu + \nu \quad . \tag{3}$$

The level of the basis functions B_α are then the sum of the levels of the tensors out of which they are contracted. Finally the potentials are constructed by including all basis functions below a given level l_{max}. This implicitly defines up to what

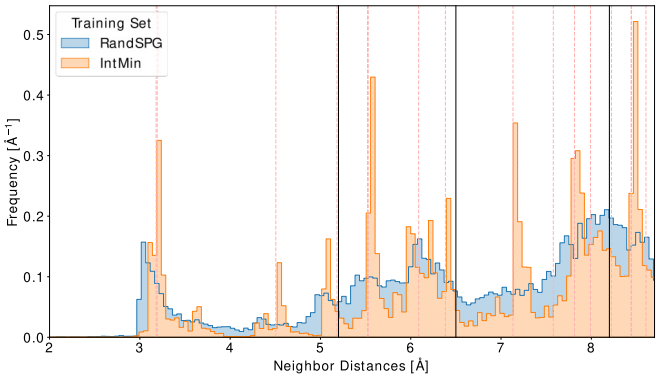

FIG. 3. Histogram of neighbor distances in RANDSPG and INTMIN training sets. Black lines are the considered cutoffs; red-dashed lines the shells of hcp Mg at equilibrium volume of Ω_0.

values of μ and ν the moment tensors $M_{\mu,\nu}$ are included in the final potential. The number of fitting parameters in a potential goes exponentially with its level.

E. Cutoff radius determination

MTPs are local potentials, i. e., they separate the total energy of a structure into individual contributions of each atom or, more specifically, to spatially localized environments that are atom centered. This environment is defined by a lower and upper cutoff radius R_c, such that all the individual regions of space considered are shell shaped. The first task in fitting potentials then is to determine appropriate cutoffs. To this end, we first calculate nearest neighbor vectors for all structures in the training sets. This task requires a detailed and explicit analysis. As the lower cutoff we pick 1.8 Å for all potentials, which is the pseudopotential cutoff used in our VASP [16,17] calculations. We select a set of upper cutoffs, which we thoroughly investigate to determine their impact on the potential's accuracy. Figure 3 shows the distribution of neighbor distances in the RANDSPG and INTMIN training sets. Also drawn are the hcp Mg shell distances at Ω_0 (dashed-red lines) and the three considered cutoffs (black-solid lines). Between 2 Å to 3 Å only very few structures are present due to the constraints we have put on the structure generation. While the nearest-neighbor distances of the RANDSPG structures are mostly evenly distributed, the INTMIN distributions shows distinct peaks. This is expected after energy minimization and gives important clues what cutoffs are physically meaningful. The peaks tend to align with the hcp shell distances (red-dashed lines), but additional peaks from other structures are also present. It can be seen that $R_c = 5.2$ Å includes the first three shells, $R_c = 6.5$ Å the first six, and 8.2 Å the first ten shells. We will pick these cutoffs for the rest of the paper. The choice of the cutoff has important consequences on the quality of the potentials as will be seen in Sec. III C 1 and it is therefore important to explicitly check what cutoffs may reasonably be considered without depriving the model of physically relevant information. Finally the fact that RANDSPG has a fairly smeared out distribution is also important as it gives the potentials critical information on out-of-equilibrium configurations.

F. Fitting procedure

We fit MTP models for each of the data sets at different model complexities, choosing levels from 8 to 24 with the MLIP [7] program, which performs energy, force, and stress matching in a least-squares optimization. All potentials are fitted with respect to energies, forces, and stresses from DFT, with weights of 1, 0.01, and 0.001 respectively.

III. RESULTS AND DISCUSSION

A. Fitting results

We obtain energy, force, and stress root mean square errors (RMSE) values after each fit. Energy RMSE are plotted in Fig. 4(a) as a function of potential level for the three cutoffs. They follow a systematic improvement, but interestingly the different structure sets follow a different convergence. Since the training sets contain progressively minimized structures their structural complexity decreases and they appear to become easier to capture for the potentials. The trend only reverses with the strained and displaced sets, which add complexity again. The lowest RMSE at the highest level also follow this trend, from which we conclude that potentials fitted to larger, more diverse, structure sets naturally have a higher interpolation error than potentials with smaller training sets.

In Fig. 4(b) the same energy RMSE is plotted as a function of cutoff for three selected levels. It can be seen that the levels below 24 quickly saturate with respect to the cutoff, i. e., to the low level descriptors higher cutoffs do not necessarily include more information. Since the level characterizes both the body-order and the number of radial basis functions included in the potential, it is not clear which of them is the limiting factor [27]. Thus, to optimize the numerical performance of a potential one should carefully check whether for the given descriptor level the cutoff is appropriately chosen [28].

B. Error-Cost tradeoff

Once sets of potentials are obtained, practitioners must pick which to use for certain applications. One way of making that choice is to look at the trade-off between computational cost and their accuracy. Here we use the training RMSE as measure for the accuracy of the potentials fit to the EVERYTHING set. To provide a measure for the computational cost we run NVT-MD on $6\times6\times6$ primitive hcp Mg unit cells for 10 000 steps, or 10 ps, at 50 K. Additionally we calculated the RMSE for four classical potentials [29–32] on the EVERYTHING set and their runtime in the same MD setup. Figure 5 shows the Pareto front for these potentials where we compare the runtime to the fitting error. Drawn as horizontal lines are the aforementioned DFT convergence errors; once the mean error (0.6 meV/atom, solid line) and the maximum error (6.4 meV/atom, dashed line). Generally the MTPs are 1–2 orders of magnitude more accurate, while being 1–3 orders slower than the classical potentials. None of the classical potential achieve lower errors than 100 meV. At low cost and low accuracy, the Pareto front follows the potential fit with a cutoff $R_c = 5.2$ Å before switching over to $R_c = 6.5$ Å at around 10 meV/atom error. While the $R_c = 8.2$ Å potential is behind the Pareto front (with the exception of the highest

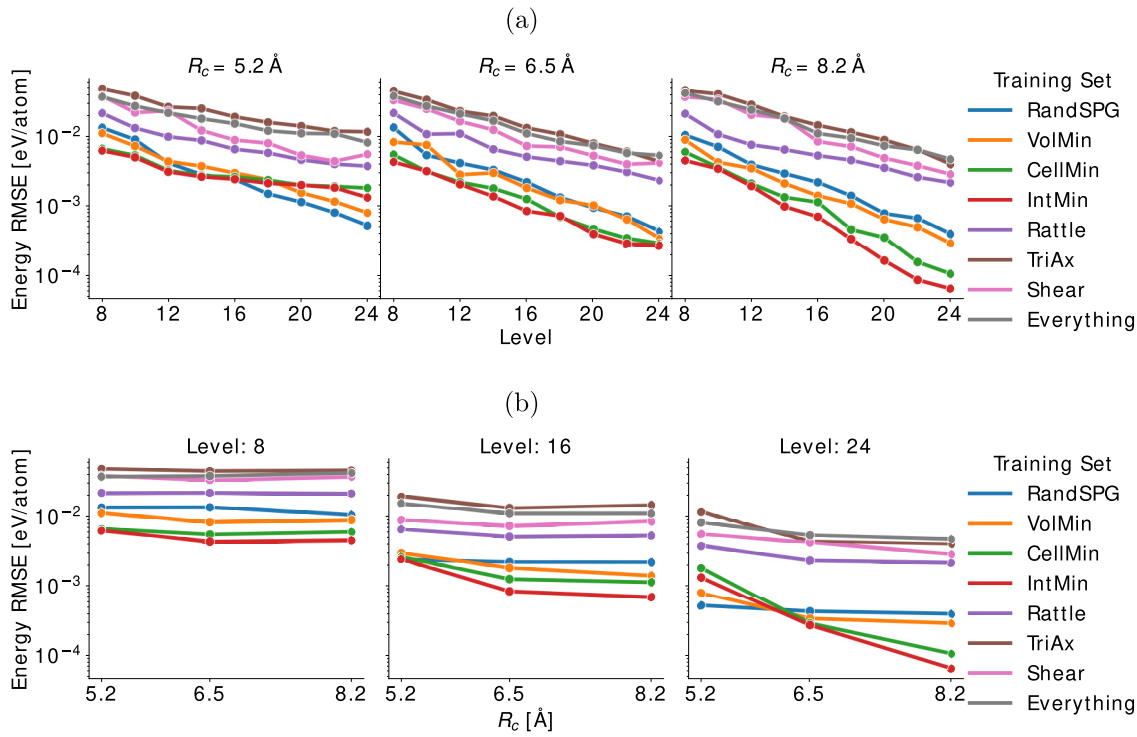

FIG. 4. Energy training RMSE of the potentials fitted to different training sets over potential level and cutoff. Subplots show potential level and cutoff radii. Higher cutoffs clearly show improved convergence rates at high level: (a) Energy RMSE over the level of the potentials. (b) Energy RMSE over the cutoff radius of the potentials.

level, where it is marginally more accurate than $R_c = 6.5$ Å), we will later see in Secs. III C 4 and III C 1 that it is still useful due to superior performance when treating defects and different closed packed structures. The plain computational cost of each potential as a function of level and cutoff are shown in Appendix A for all potentials fit to EVERYTHING.

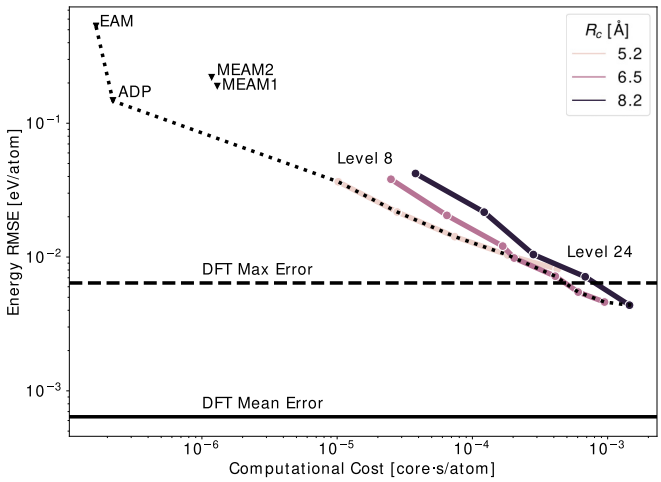

FIG. 5. Time per force call per atom versus root mean square error of the energy on the EVERYTHING set. Colors symbolize the cutoff radius. Each line is constructed by using potentials of increasing level, from 8 to 24. The dotted line marks the Pareto front. The horizontal-black lines indicate the mean (solid) and maximum (dashed) errors in the DFT training set from convergence testing. The classical potentials (ADP [31], EAM [29], MEAM1 [30], MEAM2 [32]) for comparison are shown with black triangles.

C. Verification

As mentioned in the methods, we do not split the fitting data into traditional train and test sets. Instead we perform calculations of various quantities that we can compare to independent (i. e., not entering the fitting) DFT calculations. In this section we will focus on results for the potentials fitted to EVERYTHING unless otherwise noted, deferring the discussion of the performance of the various data sets to Sec. III E. In total more than 1000 additional structures have been calculated with DFT for the verification calculations below. These structures are part of volumetric and uniaxial strain, phonon, and defect calculations are explained in more detail below.

1. Strain calculations

An important part of verifying machine learning potentials is checking that the stability of the bulk phases is correctly predicted over the volume range of interest. To this end we calculate the E-V curves of hcp-, fcc-, dhcp-, and bcc-Mg. First we strain the reference structures hydrostatically within ±80% and ±10% of the hcp equilibrium volume. Figure 6 shows the RMSE on these ranges as a function of potential level and cutoff. We compare here potentials fitted to RANDSPG and EVERYTHING. Not shown is the error of the RANDSPG set on the 80% range because this potential clearly failed, as the error exceeds 1 eV/atom. The EVERYTHING set achieves ≈5 meV/atom in the 10% range and <10 meV/atom in 80% range.

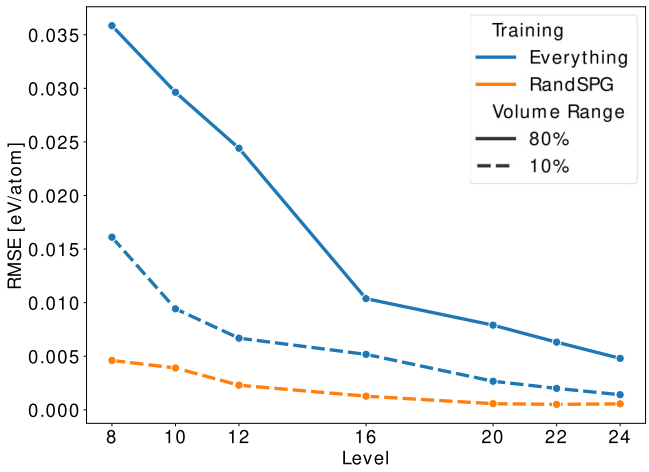

FIG. 6. Energy RMSE of potentials fitted to RANDSPG and EVERYTHING with $R_c = 8.2$ Å averaged over the volume ranges (10% and 80% respectively) and hcp, fcc, dhcp, and bcc as a function of potential level. While the simpler RANDSPG set achieves lower errors on the narrower range, it completely fails at the larger range, where as potentials trained on EVERYTHING still achieve less than 10 meV over the wide range. The RMSE on the 80% volume range for potentials fitted to RANDSPG is not shown as it exceeds 1 eV/atom indicating clear failure of these potentials on larger volume ranges. Potentials with $R_c = 5.2$ Å and $R_c = 6.5$ Å show the same qualitative behavior.

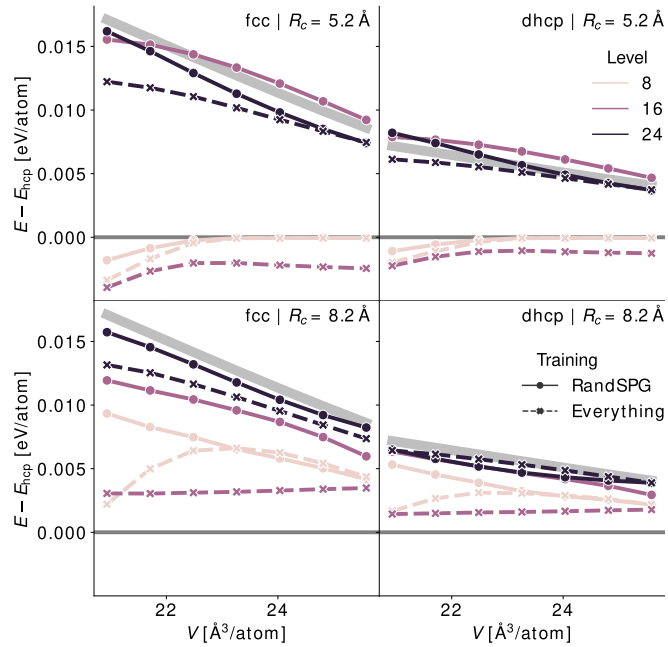

FIG. 7. Energy difference to HCP per atom vs atomic volume for FCC and DHCP (columns); thick-gray lines are DFT reference energies; figures in the top row correspond to potentials with $R_c = 5.2$ Å and in the bottom row $R_c = 8.2$ Å. The potentials are fit to the RANDSPG (solid) and EVERYTHING (dashed). While both sets and cutoffs eventually converge to the DFT phase stabilities, low cutoffs and low ranks clearly fail in differentiating close packed structures.

Figure 7 shows the energy of dhcp and fcc relative to the predicted hcp energy for potentials of level 8, 16, and 24 with cutoff 5.2 Å and 8.2 Å, and training sets RANDSPG and EVERYTHING. All potentials eventually converge close to the DFT values, but note the pronounced failure of level 8 potentials at the lower cutoff. Even more interestingly, for the larger training set also the level 16 potential fails to distinguish the three structures. At higher cutoffs all levels are able again to differentiate the structures, though again the larger training set has a harder time correctly describing all structures. The first observation implies that interpolation errors become relevant. Since the EVERYTHING set is much broader in phase space, we interpret the failure of the level 16 potential as still having too few basis functions to span the large configurations space covered by EVERYTHING. We will return to this in the context of active learning in Sec. III F. This additional interpolation error for larger training sets comes at the advantage of a larger applicability as can be seen referring back to Fig. 6. While the smaller RANDSPG set outperforms EVERYTHING in the smaller volume range, it is not at all usable on the larger range.

The bottom row in Fig. 7 shows the value of larger cutoffs, even though Fig. 5 did not seem to indicate that earlier. Here potentials with larger cutoff do no fail at differentiating the close packed structures (although quantitative agreement is achieved only at higher levels). The larger cutoff appears to make lower basis functions more efficient at differentiating closed packed structures. This also helps in the computational efficiency in a round-about way, as e.g., a level 12 potential with cutoff 8.2 Å takes as much time per force call as a level 20 potential with cutoff 5.2 Å while being less prone to overfitting, see Fig. 15. Thus, to construct potentials with an optimal balance between computational efficiency and accuracy the two parameters—levels and cutoff—should be simultaneously optimized.

Additionally we strained the prototype structures along each of the six possible axes (three strain, three shear) also within ±60% and compared again DFT and MTP. For space reasons we do not show the results here and defer to Sec. III E

2. *Phonons and force constants*

After checking static properties we now turn to dynamical properties. We have calculated phonon spectra and band structures for hcp and bcc cells at the minimum energy volume, as well as bcc cells compressed to 12 Å^3/atom where it is dynamically stable. All calculations were performed with LAMMPS [33] and PHONOPY [15] using an interaction cutoff of 10 Å, which corresponds to a 4×4×4 supercell for hcp and a 5×5×5 supercell for bcc.

Figure 8 shows the phonon band structure and density of states calculated with DFT and three MTPs fitted on EVERYTHING with cutoff 8.2 Å at three levels: 8, 16, and 24. The potentials fitted on EVERYTHING show very good agreement with the DFT results. Our validation results also indicate a significantly better description of bcc Mg, both in the compressed high pressure state as well as at the equilibrium volume, as compared to other recently reviewed Mg potentials. The band structure and density of states for bcc Mg are shown in Appendix F. Troncoso *et al.* [34] review this topic and find MEAM potentials are the best so far to study the dynamical behavior of bcc Mg, but also report that the same potentials are deficient in their elastic properties [35].

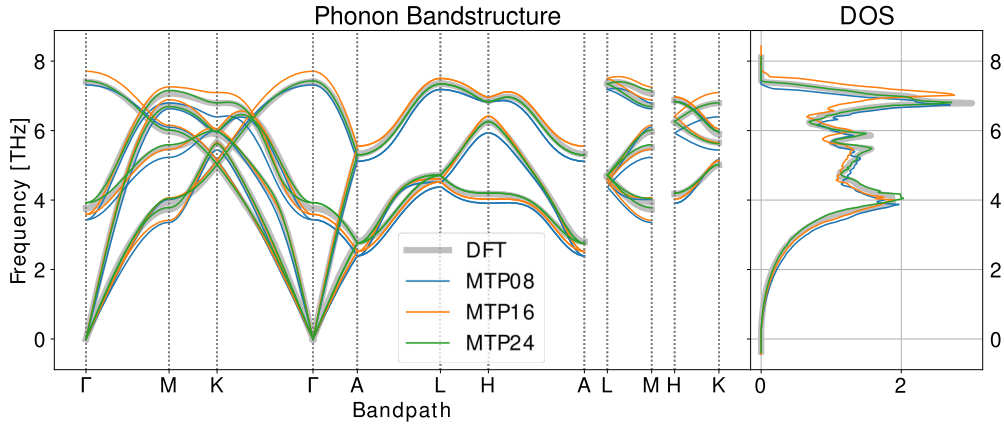

FIG. 8. Phonon band structure and density of states calculated with DFT (black line and dots) and three MTPs at levels 8, 16, and 24 (colored lines). The MTPs show very good agreement with DFT, increasing in accuracy with level.

Neural network potentials [36,37] reviewed in the same paper are found to be better for this application, but still predict wrong dynamical instabilities or predict them in the wrong part of the band structure. Based on these findings the authors concluded that these properties could be improved by specifically targeting bcc structures in the training set [34]. The fact that our potentials reproduce the dynamical as well as elastic properties so well without specifically targeting these values during the training set generation nicely illustrates the ability of our training approach to produce transferable potentials.

For a quantitative comparison between DFT and MTP phonon spectra we use the RMSE in the force constants computed on the same supercell. We refrain from directly comparing the resulting density of states, since such a comparison is ambiguous when the domain of the compared densities does not match. In contrast, errors in the force constants are naturally related to the errors in the forces themselves and are therefore an interesting quantity for validation. Figure 9 shows the force error as a function of potential level averaged over the structure prototypes. Shaded areas indicate the spread of the errors over the prototypes. For all cutoffs the accuracy naturally increases with level. We also see, however, that larger cutoffs aid the potentials in making more consistent predictions, i. e., yield lower spread of the errors over the structure prototypes. Assuming an average thermal displacement of atoms in the tenths of Å, these quantities mean that the average thermal force will have errors in around 0.1 eV/Å, which is larger than the DFT convergence error, but similar to the training force RMSE values.

3. *Vacancies*

The vacancies are constructed using 3×3×3 super cells and the reference bulk energies are calculated on the corresponding defect free primitive cells. We do this for each of the four structure prototypes mentioned previously: hcp, fcc, dhcp, and bcc. DFT calculations are run with a plane-wave cutoff of 550 eV and k-mesh spacing of 0.05 $Å^{-1}$ [38]. Structures are relaxed in their internal degrees of freedom and volume in DFT and MTP separately before the formation energy is calculated.

The mean absolute error (MAE) over the structure prototypes is shown in Fig. 10 as a function of MTP level and cutoff for the representative sets RANDSPG, RATTLE, and EVERYTHING. Appendix B shows the error on all other sets as well. The error decreases with level and saturates around 0.02 eV before increasing again in case of the RANDSPG set. As this set is smaller than EVERYTHING and RATTLE, it indicates overfitting on this set.

Shown in the bottom of Fig. 10 is the MAE as a function of cutoff at the highest fitted potential level. The smaller set RANDSPG shows increasing errors with cutoff and the (not

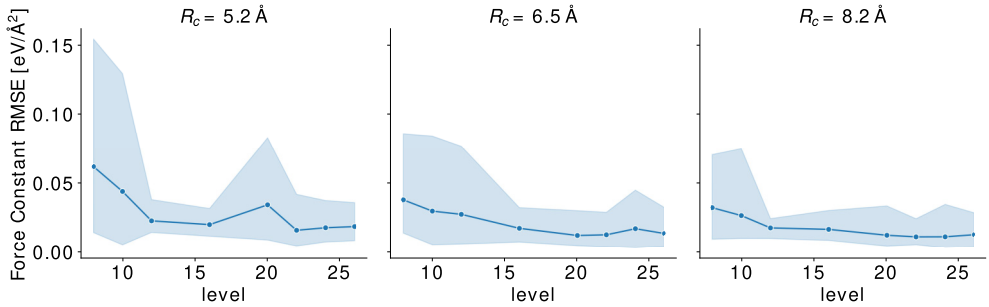

FIG. 9. Error in the force constants of potentials fit to EVERYTHING for different cutoffs. Lines are the averaged RMSE over all structure prototypes and shaded areas indicate their spread. Higher levels clearly improve accuracy, while higher cutoffs lead to a more consistent description, i. e., lower spread between structures. At highest levels the potentials reach force constant accuracy between 0.01 eV/Å^2 to 0.05 eV/Å^2.

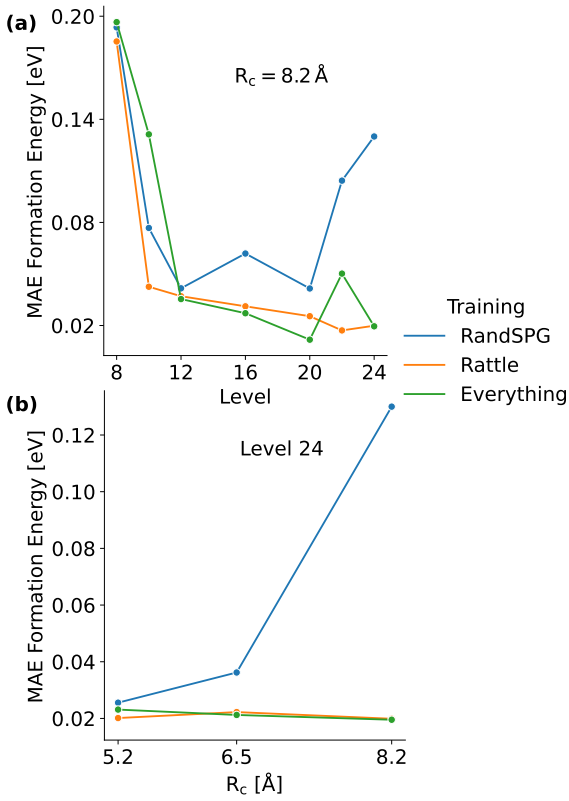

FIG. 10. Mean absolute error (MAE) in the vacancy formation energy E^f_{vacancy} averaged over the four structure prototypes: hcp, fcc, dhcp, and bcc. (a) MAE as function of potential level for potentials fitted with $R_c = 8.2$ Å for the RANDSPG, RATTLE, and EVERYTHING sets. The best potential achieve errors as low as 0.02 eV whereas the potential fitted RANDSPG shows increasing errors after level 20, a sign of overfitting due to its smaller size compared to RATTLE and EVERYTHING. (b) MAE as a function of R_c for potentials of level 24 for the same training sets. The error increases on the smaller training set RANDSPG, but is constant or decreases on the larger training sets RATTLE and EVERYTHING.

shown here) minimized sets VOLMIN and CELLMIN reproduce the same trend as the RANDSPG set. In contrast the larger sets RATTLE and EVERYTHING show constant or decreasing errors with cutoff. The other larger sets TRIAX and SHEAR follow this trend. We interpret this to mean that increasing the cutoff on smaller, less diverse training sets leads to the potentials only seeing more of the same environments, and hence overfitting, whereas it is more useful on larger, more diverse training sets, where potentials can pick up on more details that fall out of smaller cutoffs.

The absolute predictions of formation energy for just the potentials fit to the EVERYTHING set are also shown in the top row of Fig. 11 as a function of potential level. Horizontal lines are the DFT reference energies. Generally all three cutoff potentials manage to adequately (∼0.1 eV) describe the vacancy even at moderate potential levels (∼16), except for the bcc vacancy. However, the potentials fit with $R_c = 8.2$ Å manage good performance already at level 12. We therefore can conclude that at least potential cutoffs beyond 5.2 Å and potential levels beyond 16 are needed.

Next to the vacancy formation energies we also tested the performance of the potentials at predicting structural relaxations. The bottom row of Fig. 11 shows the maximum displacement (across all atoms and dimensions) during the minimization of the vacancy super cell. As in Fig. 11 already moderate potential levels manage adequate agreement and, again, lower cutoffs generally reduce predictive power. This suggest that the potentials with high basis set and cutoff also correctly relax the structure around the defect.

4. Planar defects

We calculate surface energies for the HCP (0001), $(10\bar{1}0)$, and $(1\bar{1}20)$ surfaces as well as a Σ13, two configurations of Σ19 (various rotations around [0001]) [39], Σ7*b* ($[12\bar{3}0]\,(001)$) grain boundaries and a basal reflection twin with lattice parameter $a = 3.195$ Å and $c/a = 1.624$. The $(10\bar{1}0)$ surface supports two different structures depending on which half plane terminates the surface and we have included both structures here. The Mg Σ7*b* also has a second realization, called *T* type [40], but they are very close in energy and they are not included in the calculations below, although we have verified similar accuracy on both structures independently. All surface slabs are at least 17 Å thick, more than twice the largest potential cutoff. The internal degrees of freedom are relaxed for each potential with the lateral lattice parameters fixed. The two grain boundary structures are relaxed normal to plane in DFT first. For each simulation we calculate the excess defect energy according to

$$E_{\text{defect}} = \frac{1}{A}\left(E_{\text{supercell}} - \frac{N_{\text{supercell}}}{N_{\text{reference}}}E_{\text{reference}}\right), \tag{4}$$

where the reference is bulk hcp Mg with the above mentioned lattice parameters and *A* is the cross section of the supercell.

Differences between DFT and each of the potentials are plotted in Fig. 12 for the final potentials fitted to EVERYTHING. All three cutoffs underestimate the surface energy at low potential levels, but show improvement with increasing level. Interestingly lower cutoffs seem to systematically underestimate the surface energy, whereas higher cutoffs have larger errors at low level, but get much more precise than the low cutoff at larger levels. The grain boundaries are well described at all potential levels and cutoffs. For the EVERYTHING set and $R_c = 8.2$ Å the error on the surface structures is generally below 100 meV/Å^2 and below 10 meV/Å^2 for levels larger than 16. The error on grain boundary structures is even an order of magnitude lower, except for the minimized training sets where the high-level potentials fail.

Additionally also generalized stacking fault energies and decohesion curves were tested with data provided by Stricker *et al.* [41] in Appendix C. We find very good agreement for all structures with errors below 5 meV/Å^2 even for the lowest level potentials, except for the surface energies as already indicated in Fig. 12.

This section shows that our potential can faithfully reproduce planar defects and surfaces without having seen them during training explicitly. We interpret this important fact as an indication that the RANDSPG set is complete in the sense that it contains most of the local environments that are present in planar defects.

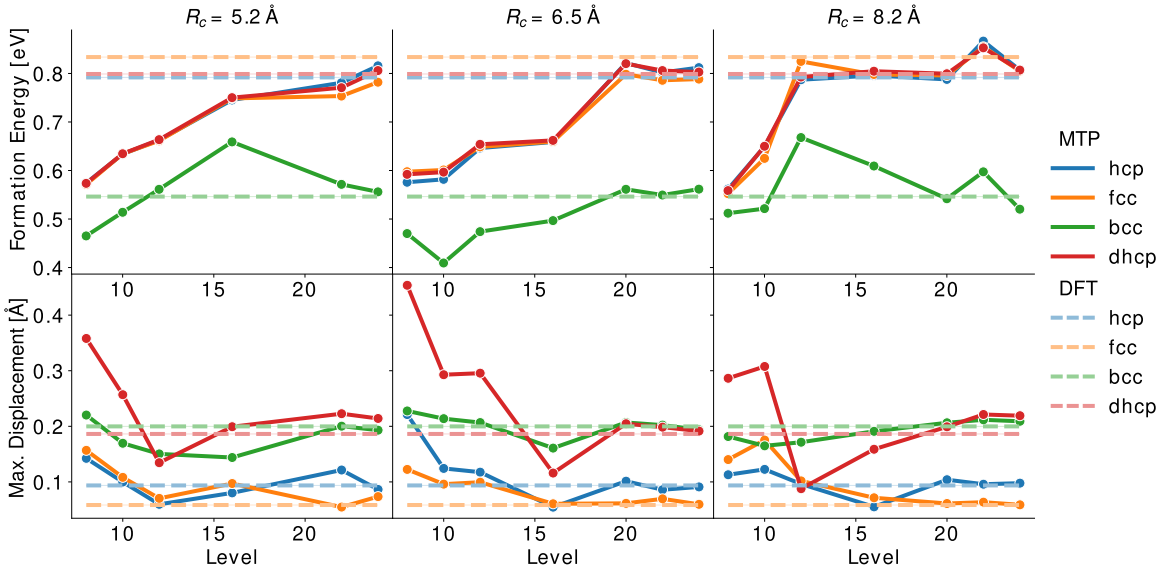

FIG. 11. Comparisons between DFT reference and MTP prediction in formation energy and maximum displacement during minimization of vacancies. Here only potentials fitted to EVERYTHING are shown, horizontal dashed lines are DFT references. (Top row) Predicted energy of vacancy formation. (Bottom row) Maximum displacement during minimization of the vacancy super cell across atoms and dimensions.

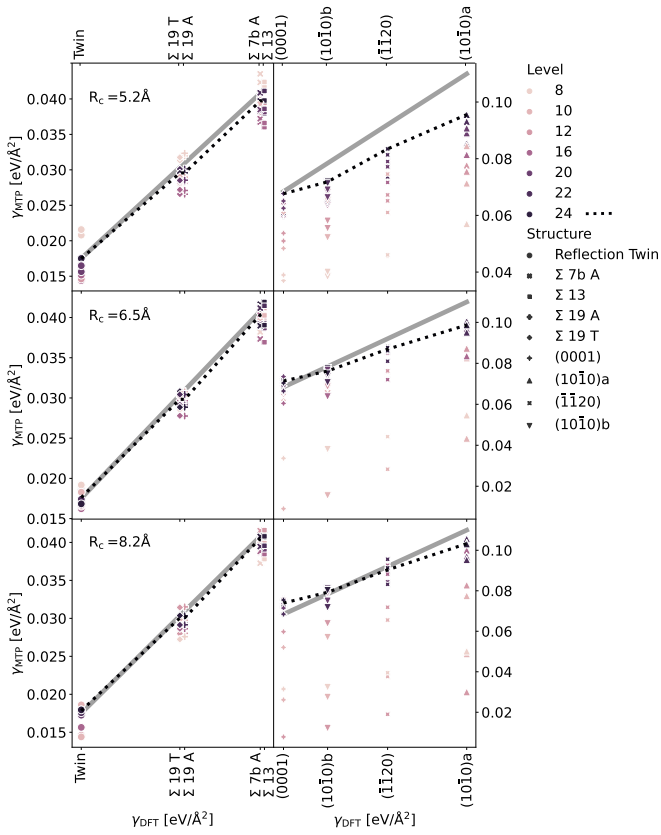

FIG. 12. Correlation plot of planar defect excess energies; DFT vs MTP predicted. Each point is a potential fit with the given cutoff and level highlighted by the hue. The grey line highlights perfect correlation; the dotted line connects the points of potentials with level 24. Compared to grain boundaries the initial accuracy of the potentials on surfaces is very poor, but increases substantially with higher levels. Note, however, that except for the highest cutoff, even the highest levels systematically underestimate the surface energies.

D. Testing at high temperature and pressure

As a final test of the stability of the fitted potentials over a wide range of thermodynamic conditions we calculate the thermal expansion and isothermal compressibility. Since hcp is the only thermodynamically stable configuration at zero or low pressures, results are shown here only for this structure. However, we also ran the same simulations for bcc, dhcp, and fcc and have verified that MD simulations are always stable. The simulations are run with $4\times4\times4$ unit cells for 1×10^6 MD steps.

Figure 13 shows the change of internal energy of hcp Mg with temperature at zero pressure and with pressure at $T = 300$ K. Simulation boxes remain stable until around 800 K to 900 K depending on potential level, after which melting occurs. While we did not carry out detailed calculations to precisely determine the melting point, this range is in good qualitative agreement with the experimental melting temperature of 923 K reported for Mg, see e.g., [42]. Figure 13 only shows potentials with a cutoff of 8.2 Å, but the cutoff appears to have no effect on the description of the simple thermodynamic state variables investigated here. On the other hand the potential level seems to induce a shift in the finite temperature energy of a few 10 meV. We speculate that this is due to slightly larger forces constant errors at lower levels, which means slightly different heat capacities.

We investigate a pressure window from -3 GPa to 12 GPa. Instabilities occur at large tensile pressures and elevated temperatures. The lowest pressure that leads to unstable simulation boxes for hcp is -0.5 GPa at 800 K, well above the tensile strength of pure magnesium. We note that higher level potentials seem to be able to bear more tensile pressure than lower level ones before becoming unstable.

E. Comparison of potentials

We now collect the errors against DFT calculated in the previous sections for all training sets and test domains in

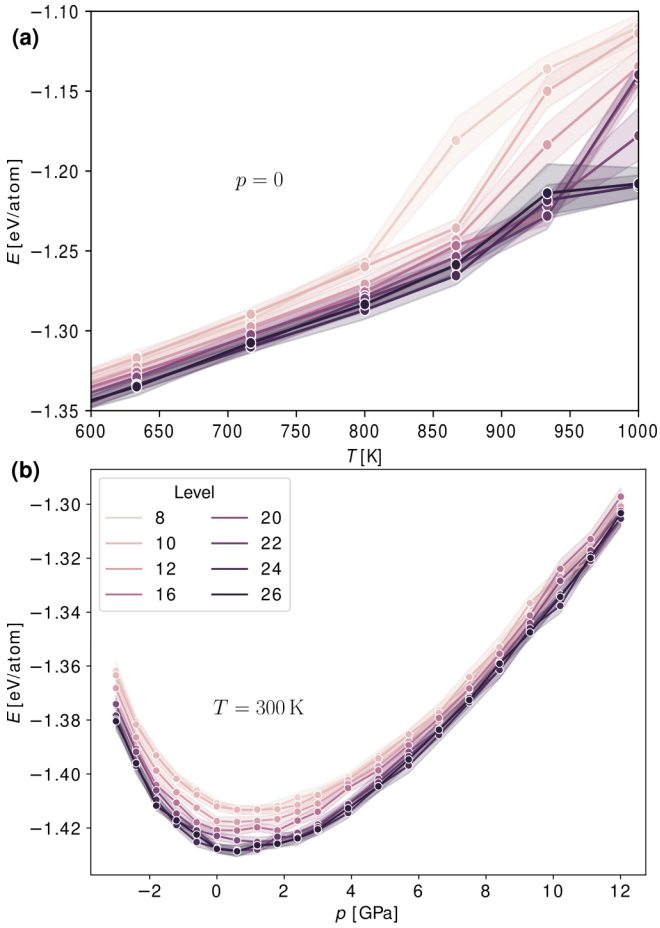

FIG. 13. Results from NPT MD simulations for potentials fit to EVERYTHING and $R_c = 8.2$ Å. Simulations for the other two cutoffs were run, but do not show significant differences from the ones shown. (a) Average energy as a function of temperature at $p = 0$ GPa. Melting is indicated by the abrupt change the temperature range 800 K to 900 K depending on the potential level. (b) Average energy as a function of hydrostatic pressure at $T = 300$ K.

Fig. 14. Generally potentials fit to RANDSPG show average performance overall, except on tests with large strains. Potentials fit to the minimized sets (VOLMIN, CELLMIN, INTMIN) tend to perform badly, due to their reduced inherent dimensionality after the minimization steps. While potentials fit to EVERYTHING are not always the best performing potentials in each test case, they consistently rank among the top potentials. For all defects at fixed potential level higher cutoffs give smaller errors, except for level 8, where it first increases for two of the minimized sets, CELLMIN and INTMIN. This aligns with our discussion of Fig. 4(b), that the low level descriptors are not flexible enough to make use of all the information available. The minimized sets are shown to be overfit, giving completely nonsensical results on some of the verification sets, worsening with increasing level. Comparing the two volume verification sets, we note that for the 10% set the training sets follow the same trend in accuracy as in Fig. 4(a), which we can understand since this is also the volume range present in each training set. On the substantially larger 80% set only TRIAX (as expected) and SHEAR perform well. It is not completely surprising, but interesting that the SHEAR set outperforms the other sets so strongly, not having seen any (uniaxially) strained structures either, but the shear structures seem to carry at least some information also on strain. In the single axis strained sets again, TRIAX and SHEAR perform well, but also RATTLE is in between both sets. Even though RATTLE only experiences strain and shear up to 5% during training, it still seems to give adequate results on the verification up to 60%. On the other hand, the SHEAR set does not substantially improve the performance of the EVERYTHING set anymore, but causes it to perform less well on different stackings in closed packed structures as discussed in Sec. III C 1 and Figs. 6 and 7. In the final potentials provided in the supplementary we have therefore excluded it again.

Vacancies appear least well described of all defects checked, but this is also due to that fact the errors for the planar defects and surfaces are given normalized to the area. Surfaces, however, are less well described than the grain boundaries, since they are included in the training set only indirectly due to the VOLMIN set, as indicated earlier. Still the potentials with highest level achieve less then 10 meV/Å^2 error on them. For grain boundaries the accuracy is even in the order of 1 meV/Å^2. We speculate that accuracy on the surfaces could be improved by cutting the randomly generated bulk structures of the RANDSPG set without compromising the unbiased sampling.

F. Comparison to active learning

We want to further show that potentials fitted to a wide range of physically inspired structures are able to be reliably transferred to structures outside their original training domain. The active learning scheme as implemented by MLIP [7,43] is reviewed in Appendix E 1. We present calculations that show the potentials fitted in this paper would remain unchanged under an active-learning scheme for applications investigated here.

We run MD for 100 ps on four different structures under the active learning regime provided by MLIP [7]. The structures are liquid Mg under high pressure ($T = 2500$ K and $p = 5$ GPa, to keep the simulation box stable); an HCP Mg vacancy in a 4×4×4 super cell at moderate temperatures and ambient pressure (T in 100, 400, 700 K); an HCP twin boundary and a Σ7b grain boundary at 600 K and ambient pressure; and finally fcc and dhcp structures at ambient temperature and pressure. The super cells are adjusted for each potential to allow for at least twice the potential cutoff radius to fall within the periodic boundaries. We set selection, $\gamma_{\text{select}} = 1.001$, and extrapolation thresholds, $\gamma_{\text{break}} = 5$, to catch any extrapolation. They define when a structure is selected for active learning or when the simulation is deemed to inaccurate, c.f. Appendix E 1 for their precise definitions. Within the given time frame *none* of the potentials trained on the EVERYTHING set encountered structures where $\gamma > \gamma_{\text{select}}$, i.e., there was no evidence of extrapolation. Appendix E 1 also shows exemplary calculations of the extrapolation grade over simulation time for potentials fitted on subset of EVERYTHING.

We can now verify our hypothesis explaining the success of the potential in Sec. III C 4; if the accuracy were due to good extrapolation, the active learning would still flag the structures

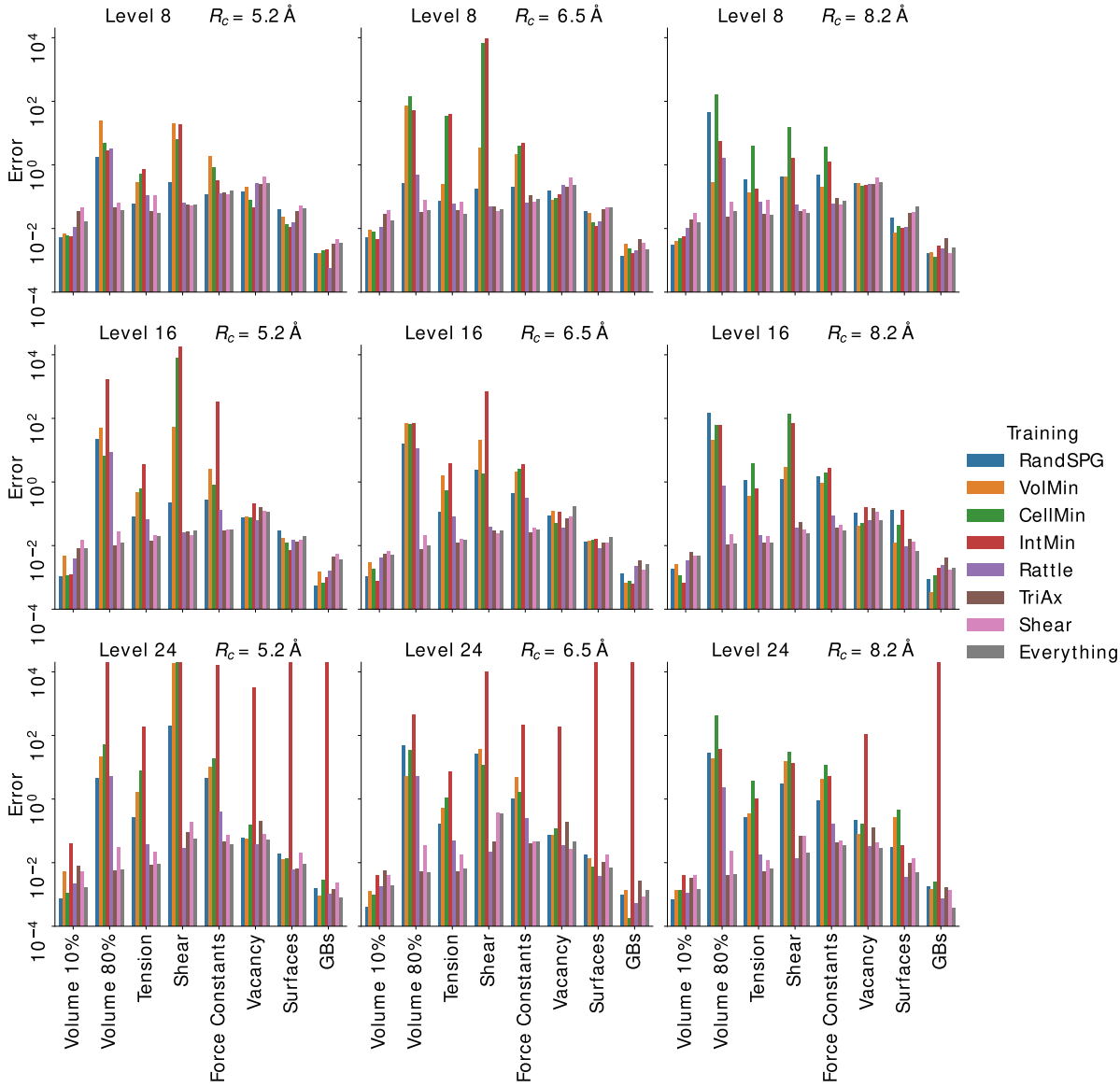

FIG. 14. Errors on verification sets for three selected potential levels (8, 16, 24) and cutoffs (5.2 Å, 6.5 Å, 8.2 Å). The x axis enumerates the verification set, while the y axis gives the logarithmic error on the set for the different training sets. For the sets "Volume", "Tension", and "Shear" the errors are the direct energy RMSE against DFT in eV; for "Force Constants" the errors the RMSE of the force constants in a super cell against DFT; and for the three defect sets "Vacancy", "Surface", "GBs" it is the RMSE of the respective energy excess of formation.

as outside of the active set. Since that did not happen, there must be local environments in the training set that are sufficiently close in descriptor space to allow interpolation of the defect structures.

With respect to the failure of low level potentials to differentiate closed packed structures as discussed in Sec. III C 1, we can now clearly state that this is an interpolation error, since dhcp and fcc structures were not identified as extrapolation by MLIP [7]. This is in contrast to Zeni *et al.* [44]. There, they fit atomic cluster expansion (ACE) [45] potentials to a variety of systems and e.g., show that potentials fit to liquid water often enter the extrapolative regime when applied to equilibrium ice structures. Since ACE and MTP potentials both completely span the space of local atomic environments [[46], Appendix B], we believe our results and approach are also transferable to ACE potentials. That our potentials do not extrapolate on the structures tested above, whereas they do in the study of Zeni *et al.* [44], shows that the choice of the structure set is critical.

On the other hand the original question they set out to answer and that has been discussed in the literature before is how to tell whether potentials are reliable or not in a given application. The distinction of extrapolation vs interpolation is then only a proxy for this. We do not provide an answer, but our results indicate that this question is not settled yet.

From this discussion we are confident that the construction of our training set is complete and the potentials can be directly used in applications with respect to the tests performed here. Still, users that wish to use the potentials on defects we have not verified, i. e., on isolated dislocation or cracks

may need to do additional testing. As for dislocations, given that the $\Sigma 7b$ interface is comprised of dislocation cores [47] we do not expect significant failure. In any case we expect our structure sets to be excellent starting points for additional active learning where necessary.

G. Transferability

For the purpose of this paper take a *transferable potential* to be one that describes not only structures it has been fitted to, but also any other structures it might be applied to or is at least well-behaved. This is true for most empirical potentials, but not for most machine learning potentials, and is often cited as their strength.

It is naturally not possible to exhaustively test this property on any given potential, though the preceding sections show that the potentials fitted here describe defects very well without being fitted on them. On the other hand, it is at least necessary that a transferable potential can adequately describe the structures of the training set of another potential. We have tested this condition here by applying two other machine learning potentials for Mg (a HDNNP by Stricker *et al.* [36] and a recent RANN by Barret *et al.* [48]) on the training set presented here and vice versa and find that our potentials achieve similar RMSE on other training sets as on their own sets.

The network based potentials however perform much worse on the broad EVERYTHING training set than on their own respective sets. This finding should not be mistaken to mean those potentials or their underlying formalism perform badly or are ill suited to the study of Mg and its defects. In fact they perform very well on the defects they were used to study [48,49]. It merely indicates that they are not (and were not meant to be) general potentials or at least do not transfer as well as potentials fitted to more extensive training sets as the one presented here. The numerical results of this comparison are summarized in the Appendix D.

IV. CONCLUSIONS

We demonstrate that an unbiased, systematic construction of the training set, which covers *all* bulk crystal symmetries instead of just low-energy structures, allows the successful construction of ML potentials that have a high degree of transferability, in particular to bulk defects. Importantly, the training database does not include any explicit defect structures or additional data from active learning, see Figs. 12 and 11.

In Sec. III A and Fig. 12 we show that the determination of the cutoff radius warrants more care than is sometimes paid in literature: A large enough cutoff radius is crucial for the transferability of the potential to structures not included in training (in this case surfaces). To fully utilize this benefit potentials of higher level (basis set) are required. In the future it may hence be worthwhile to separately increase the number of radial functions keeping the maximum tensor power in the basis functions (i. e., the body order) constant to inspect and utilize this effect.

In Sec. III F we further show that active learning does not provide additional benefits once a sufficiently diverse training set is considered. Additionally, we show that machine learning potentials can give nonoptimal results even when they are not extrapolating. This is an important statement as it is often implicitly or explicitly assumed by researchers developing machine learning and active learning formalisms that interpolation errors can be neglected. The question how to treat them hence continues to be an open question.

We expect our observations hold for general MLIPs with descriptors that form a complete basis, but at least for the ACE, since MTP descriptors can be expressed in the ACE basis as well.

In practical terms we provide a general purpose potential for Mg, that describes the equilibrium hcp and high-pressure bcc phases, and also planar and point defects. We recommend levels higher than 16 with $R_c = 8.2$ Å, to avoid wrong (zero) stacking energies predicted for low level potentials [50]. Lower levels and cutoffs may be used to save computational resources, although care has to be taken so that the interpolation errors do not jeopardize the application. For these cases it may be more advisable to custom fit lower level potentials to narrower training sets.

The approach to construct unbiased and physical data sets by systematically sampling bulk structures over all bulk symmetries can be straightforwardly applied to other materials systems. The construction approach outlined in this study facilitate the construction of such data sets in a systematic and automated fashion. We have shown that potentials fitted to training sets of this kind transfer better to training sets of other potentials than vice versa. This opens the route towards a largely automatized generation of general, transferable and accurate potentials.

The full DFT training set and resulting MTP potential files for the potentials fitted to EVERYTHING with and without the SHEAR set can be accessed online [51]. Under GitHub [52] we have also added example jupyter notebooks and a PYIRON [53] project that shows how to access the data, generate similar training sets and run simple simulations with LAMMPS [33].

ACKNOWLEDGMENTS

The authors acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Collaborative Research Center 1394 (SFB 1394, No. 409476157) and Project No. 405621160. E.B. further acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 725483). We also thank Ralf Drautz, whose potential fitting experience benefited us during this study, Prince Matthews for providing grain boundary structures and Markus Stricker for help with setting up their neural network potential and providing defect structures.

APPENDIX A: COST-ACCURACY TRADEOFF

Figure 15 shows the cost per force call per atom as a function of potential level and cutoff radius. There is a log-linear trend in the runtime vs level as expected, since the number of free parameters in the potential increases exponentially with

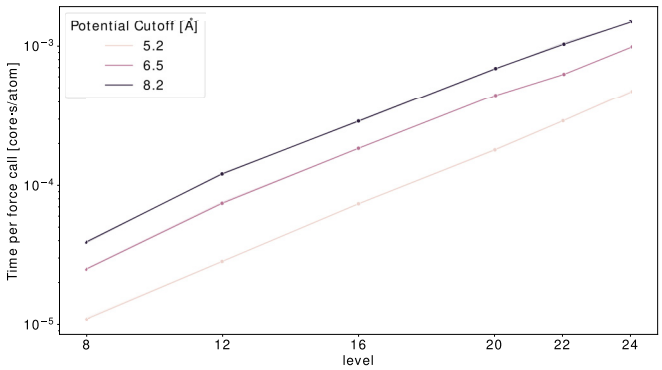

FIG. 15. Time per force per atom versus potential level for different cutoffs. Note the log-linear trend due to the exponential number of parameters as a function of level.

the level. Increments in radial cutoff only change the prefactor of the scaling, which is also expected of local models.

APPENDIX B: ERRORS ON VACANCY FORMATION ENERGY FOR ALL TRAINING SETS

In Figs. 16–18 the full data behind Fig. 10 is shown. The minimized sets VOLMIN, CELLMIN, and INTMIN follow the same trend as RANDSPG, whereas TRIAX, SHEAR follow the trend of RATTLE. The INTMIN set is clearly shown to be insufficient for potentials with high levels.

APPENDIX C: STACKING FAULTS AND DECOHESION CURVES

Based on the training data from Stricker *et al.* [41], we have compared also the generalized stacking fault energy (GSFE) and decohesion curves along various orientations in hcp. We picked two potentials fitted with $R_c = 8.2$ Å and level 8 and 24 as representative for the potentials fitted here. The results are summarized in Figs. 19 and 20. They show the energy difference along the stacking fault path or decohesion. Since DFT and the potentials predict slightly different cohesive energies for the bulk structures, we have subtracted this difference

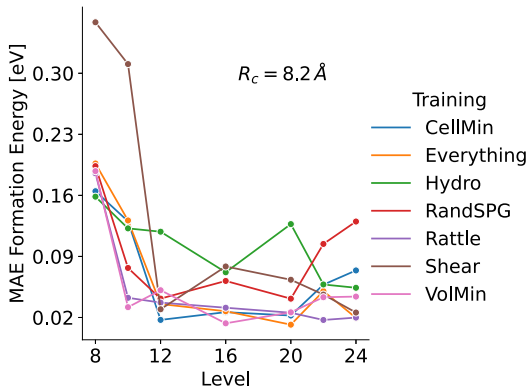

FIG. 16. Mean absolute error (MAE) in the vacancy formation energy E^f_{vacancy} averaged over the four structure prototypes: hcp, fcc, dhcp, and bcc, as function of potential level for potentials fitted with $R_c = 8.2$ Å for all sets except INTMIN.

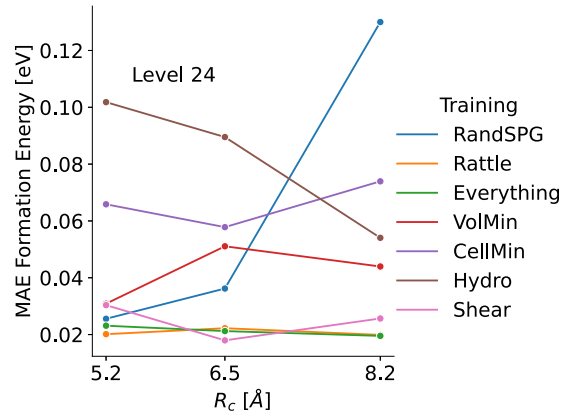

FIG. 17. Mean absolute error (MAE) in the vacancy formation energy E^f_{vacancy} averaged over the four structure prototypes: hcp, fcc, dhcp, and bcc, as a function of R_c for potentials of level 24 to all training sets, except INTMIN.

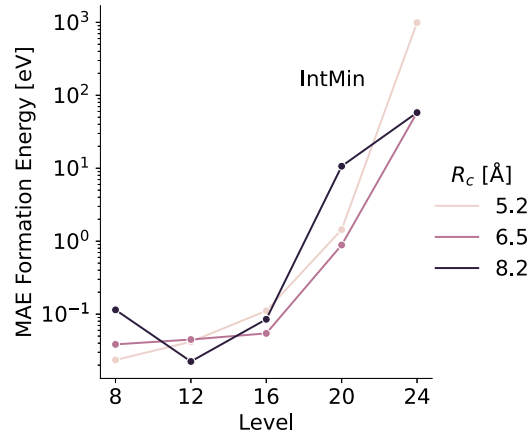

FIG. 18. Mean absolute error (MAE) in the vacancy formation energy E^f_{vacancy} averaged over the four structure prototypes: hcp, fcc, dhcp, and bcc, for the potentials fitted to INTMIN. At high levels the potentials clearly fail, indicated that the training set is not diverse enough for the model complexity.

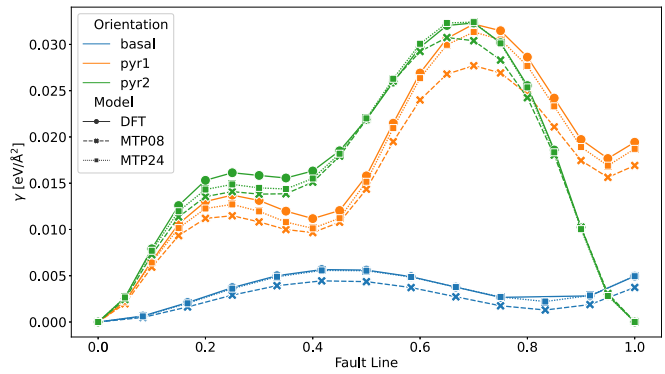

FIG. 19. GSFE curves for the basal, pyramidal 1 and pyramidal 2 stacking faults in hcp Mg calculated with DFT (solid lines) and two potentials with $R_c = 8.2$ Å and level 8 and 24 (dashed and dotted lines). All three description are qualitatively the same with the level 24 slightly better than level 8. The errors for all structures compared to DFT are below 5 meV/Å^2.

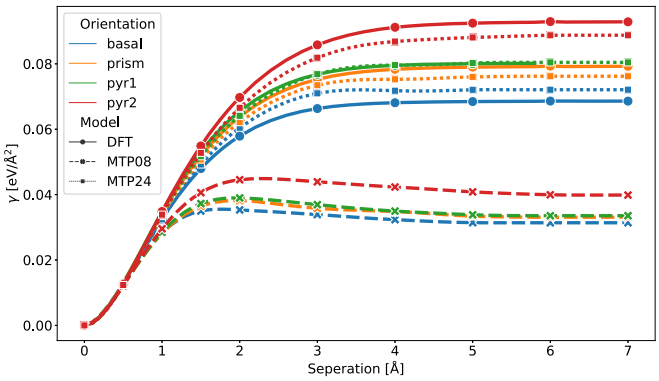

FIG. 20. Decohesion curves for the basal, pyramidal 1, 2, and prismatic surface in hcp Mg calculated with DFT (solid lines) and two potentials with $R_c = 8.2$ Å and level 8 and 24 (dashed and dotted lines). Markers are placed only as an aid to the eye, data points are denser. At short ranges both potentials give a correct description, but the level 8 potential (dashed lines with crosses) fails quickly due its severe underestimation of the surface energies.

as only the energy barriers along the path are of physical significance. For the level 24 potential all structures are in the range of a few meV/Å In contrast the level 8 potential does not give an adequate description of the decohesion. This is because they underestimate the surface energies as already noted in Fig. 12. On the other hand they also do well on the GSFE curves. For technical reasons [56] the DFT calculation for comparisons were carried out with the S/PHI/NX [54] DFT code, but using the VASP [16,17] provided pseudopotential files and equivalent settings as the rest of this paper.

APPENDIX D: TRANSFERABILITY COMPARISON

Table I compiles the results of the transferability tests between a MTP of level 24 with $R_c = 8.2$ Å and the HDNNP [36] and RANN [48] network potentials mentioned in the main text. For the calculation of the RMSE we have used the structures and corresponding DFT energies of each publication.

Since the HDNNP training set also used VASP [16,17], we have checked that our convergence settings and theirs produce matching DFT energies on a sub set of their training set. For the RANN training set we shifted the DFT energies to agree with the experimental cohesive energies of Mg, as done in their parametrization of the potential.

TABLE I. RMSE in meV/atom for three potentials mentioned in the main text tested on each training set. Asterisks (*) indicate errors in excess of 1 eV/atom.

	EVERYTHING		Stricker *et al.* [41]			Nitol *et al.* [48]	
	all	$E<0$	all	no dimer	no dimer no E–V	all	no dimer
MTP24	4.8	4.2	87.2	9.1	2.7	21.7	7.0
HDNNP	*	400.2	157.9	19.3	1.3	71.1	72.1
RANN	*	387.1	*	37.1	3.9	0.4	0.4

For the full EVERYTHING set the two network potentials fail, so we also checked a subset with only those structures that have negative energies to avoid those with atoms very close to each other. For the HDNNP training set, we found the errors of all potentials to be dominated by the dimer curve and the long distance part of the E–V curve, so we also checked subsets without those structures, and similarly also for the RANN training.

Errors on structures with very large separation between atoms are also affected by the fact that all three potentials give slightly different energies of the isolated atom (as they are fit to different references). The failure of the RANN potential on the set with the dimer structures is only due to some (unlikely to be relevant) structures with a Mg–Mg separation of less than 1 Å. The largest contribution to the RMSE of the potential fitted here is in surface structures, which could be remedied in future work by systematically including surfaces, as sketched in Sec. III C 4.

APPENDIX E: ACTIVE LEARNING

1. Active learning scheme

Suppose we have the list of m basis functions calculated for a given atomic neighborhood $\{b_i\}$ and we wish to know

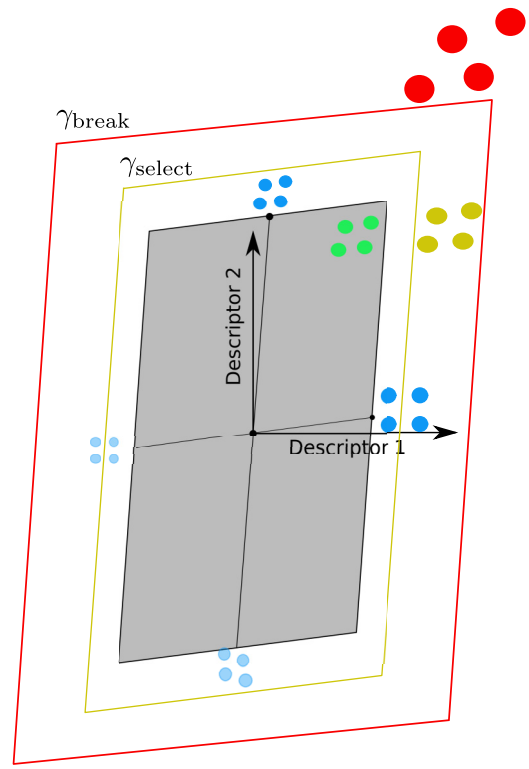

FIG. 21. Schematic illustration of phase space seen during training and active learning. Suppose our training set (blue structure) is made up of strained and sheared cubic structures. The phase space spanned by the active set then corresponds to the grey area. Structures inside the yellow borders (green) are assumed to be approximated well by the potential; between the yellow and the red border (yellow) are selected for further training and outside the red border (red) cause simulations to terminate.

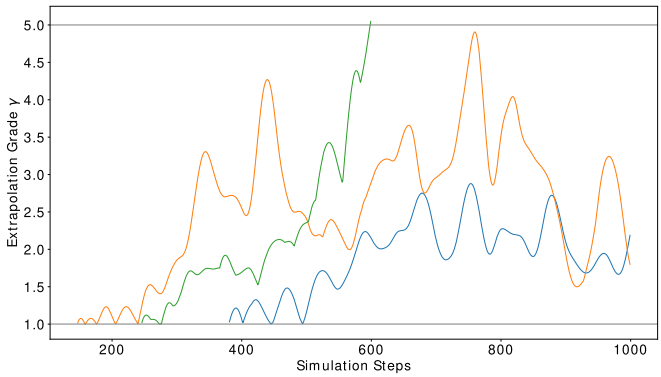

FIG. 22. Extrapolation grade as a function of time during three *NPT* simulations of liquid Mg at 2500 K and 5 GPa using potentials fitted to insufficient training sets. Potentials can be seen interpolating initially, then exceed the selection threshold and even the breaking threshold in one case.

whether this configuration can be safely approximated by the potential or not. Novikov *et al.* [7] answer this by defining an active set of configurations. We can think of this set being chosen from all training structures such that they cover the widest range of phase space seen during training [57]. Figure 21 shows a schematic illustration of the active set in solid blue, the covered phase space corresponds to the gray shaded area. They then define the active learning state A^{-1} as the matrix that projects our calculated coefficients $\{b_i\}$ into the space spanned by the active set, i.e. we can obtain the basis coefficients of the given atomic neighborhood by matrix multiplication

$$c = A^{-1}b. \tag{E1}$$

Now if all $\{c_i\}$ are smaller than unity [7], define this configuration to be within the interpolative region where small fitting errors can be assumed. If however one the of coefficients is larger than unity then the configuration is outside of the phase space region sampled during training and we might like to add it to the training set to improve the potential. They quantify this notion by introducing

$$\gamma(\mathrm{cfg}) = \max_i |c_i| = \max_i \left|A^{-1}_{ij} b_j(\mathrm{cfg})\right| \tag{E2}$$

where $\gamma \leqslant 0$ in the first case discussed above and $\gamma > 1$ in the second case. The active learning regime implemented in MLIP [7] then consists of running any desired simulation protocol but keeping tracking of γ for all neighborhoods in the considered structure. They define two thresholds, which we draw also in Fig. 21 with yellow and red lines. Exceeding the first (yellow), γ_{select}, causes the whole structure to be written out for later consideration. Exceeding the second (red), γ_{break}, causes the whole simulation to be aborted. When the simulation aborts due the latter case, the structures written in the first case can be used to enrich the training set and then rerun the simulation. This is repeated until the full simulation runs without exceeding γ_{break}.

2. Example simulation for extrapolation grade

To illustrate the concept of the extrapolation grade we run three simulations of liquid Mg in an *NPT* ensemble at 2500 K and 5GPa for 1 ps and track the extrapolation grade over time. We use $\gamma_{\mathrm{select}} = 1.001$ and $\gamma_{\mathrm{break}} = 5$ as in the main text. The resulting extrapolation grade as a function of time is shown in Fig. 22. The potentials used for this illustration were fitted to a random sample of 10% of EVERYTHING with level 16 and $R_c = 8.2$ Å to simulate an insufficient training set.

APPENDIX F: PHONON RESULTS FOR BCC MG

Figures 23 and 24 show the phonon band structure of bcc Mg at $\Omega = 22.83$ Å^3/atom and $\Omega = 12$ Å^3/atom respectively. It can be seen that also the phonons of the bcc polymorph are well described, even at volumes far from the equilibrium volume Ω_0. Note that in contrast to the potentials reviewed by Troncoso *et al.* [34] the potentials shown here correctly describe all the essential features as well as the negative part of the phonon spectrum in case of the minimized bcc structure. The only exception is the level 8 potential that does not capture the dynamical instability of bcc Mg at ambient pressure.

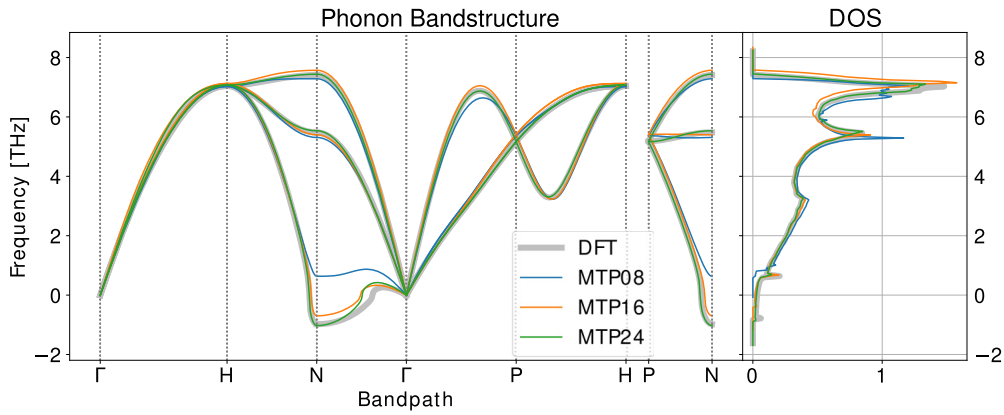

FIG. 23. Phonon band structure and density of states of bcc Mg ($\Omega = 22.83$ Å^3/atom) calculated with MTPs fitted to EVERYTHING with $R_c = 8.2$ Å at levels 8, 16, and 24. Colored lines indicate the results from MTPs; the thick-gray line indicates results from DFT.

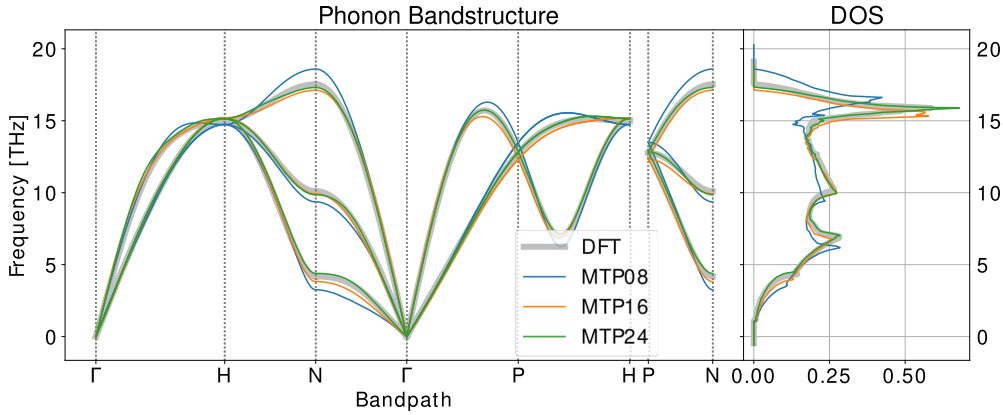

FIG. 24. Phonon band structure and density of states of compressed bcc Mg ($\Omega = 12\,\text{Å}^3/\text{atom}$) calculated with MTPs fitted to EVERYTHING with $R_c = 8.2\,\text{Å}$ at levels 8, 16, and 24. Colored lines indicate the results from MTPs; the thick gray line indicates results from DFT.

[1] R. Nazarov, T. Hickel, and J. Neugebauer, First-principles study of the thermodynamics of hydrogen-vacancy interaction in fcc iron, Phys. Rev. B **82**, 224104 (2010).

[2] E. Clouet, L. Ventelon, and F. Willaime, Dislocation Core Energies and Core Fields from First Principles, Phys. Rev. Lett. **102**, 055502 (2009).

[3] D. Finkenstadt and D. D. Johnson, Interphase energies of hcp precipitates in fcc metals: A density-functional theory study in Al-Ag, Phys. Rev. B **81**, 014113 (2010).

[4] L. Huber, J. Rottler, and M. Militzer, Atomistic simulations of the interaction of alloying elements with grain boundaries in Mg, Acta Mater. **80**, 194 (2014).

[5] P. R. Cantwell, T. Frolov, T. J. Rupert, A. R. Krause, C. J. Marvel, G. S. Rohrer, J. M. Rickman, and M. P. Harmer, Grain boundary complexion transitions, Annu. Rev. Mater. Res. **50**, 465 (2020).

[6] S. Korte-Kerzel, T. Hickel, L. Huber, D. Raabe, S. Sandlöbes-Haut, M. Todorova, and J. Neugebauer, Defect phases—Thermodynamics and impact on material properties, Int. Mater. Rev. **67**, 89 (2022).

[7] I. S. Novikov, K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, The MLIP package: Moment tensor potentials with MPI and active learning, Mach. Learn.: Sci. Technol. **2**, 025002 (2021).

[8] J. Behler, First principles neural network potentials for reactive simulations of large molecular and condensed systems, Angew. Chem. Int. Ed. **56**, 12828 (2017).

[9] J. S. Smith, B. Nebgen, N. Mathew, J. Chen, N. Lubbers, L. Burakovsky, S. Tretiak, H. A. Nam, T. Germann, S. Fensin *et al.*, Automated discovery of a robust interatomic potential for aluminum, Nat. Commun. **12**, 1257 (2021).

[10] N. Bernstein, G. Csányi, and V. L. Deringer, *De novo* exploration and self-guided learning of potential-energy surfaces, npj Comput. Mater. **5**, 99 (2019).

[11] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning, Phys. Rev. B **99**, 064114 (2019).

[12] M. Karabin and D. Perez, An entropy-maximization approach to automated training set generation for interatomic potentials, J. Chem. Phys. **153**, 094110 (2020).

[13] D. Montes de Oca Zapiain, M. A. Wood, N. Lubbers, C. Z. Pereyra, A. P. Thompson, and D. Perez, Training data selection for accuracy and transferability of interatomic potentials, npj Comput. Mater. **8**, 189 (2022).

[14] P. Avery and E. Zurek, Randspg: An open-source program for generating atomistic crystal structures with specific space-groups, Comput. Phys. Commun. **213**, 208 (2017).

[15] A. Togo and I. Tanaka, First principles phonon calculations in materials science, Scr. Mater. **108**, 1 (2015).

[16] G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. **6**, 15 (1996).

[17] G. Kresse and J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).

[18] For materials with dynamically unstable ground states or phases that are stable only at elevated temperatures, this minimization procedure might appear to under sample or neglect them. However even in these cases as long as the initial RANDSPG set is sufficiently diverse to find the structures once, their local environment will be sampled by the minimizing training set.

[19] For the TRIAX and SHEAR sets the respective entries in the strain tensor are chosen from uniform random distribution in range $\pm 80\%$.

[20] P. E. Blöchl, Projector augmented-wave method, Phys. Rev. B **50**, 17953 (1994).

[21] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B **59**, 1758 (1999).

[22] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. **77**, 3865 (1996).

[23] Less then 10% of the structures have k mesh spacing larger then 0.09Å and very few less than the above quoted.

[24] M. Methfessel and A. T. Paxton, High-precision sampling for Brillouin-zone integration in metals, Phys. Rev. B **40**, 3616 (1989).

[25] A. V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials, Multiscale Model. Simul. **14**, 1153 (2016).

[26] Readers acquainted more with classical interatomic potentials might expect a factor $\frac{1}{2}$ in front of the per atomic contribution to the total energy, to avoid double counting. In the usual machine learning formalism, this factor is absorbed into the basis coefficients ξ_α.

[27] The MLIP code allows to train potentials where body order and number of radial functions are independently varied, but we have not investigated in this paper.

[28] Comparisons of the runtime cost as a function of cutoff radius are explicitly given in Fig. 15.

[29] D. Y. Sun, M. I. Mendelev, C. A. Becker, K. Kudin, T. Haxhimali, M. Asta, J. J. Hoyt, A. Karma, and D. J. Srolovitz, Crystal-melt interfacial free energies in hcp metals: A molecular dynamics study of Mg, Phys. Rev. B **73**, 024116 (2006).

[30] D. E. Dickel, M. I. Baskes, I. Aslam, and C. D. Barrett, New interatomic potential for Mg–Al–Zn alloys with specific application to dilute Mg-based alloys, Modell. Simul. Mater. Sci. Eng. **26**, 045010 (2018).

[31] D. Smirnova, S. Starikov, and A. Vlasova, New interatomic potential for simulation of pure magnesium and magnesium hydrides, Comput. Mater. Sci. **154**, 295 (2018).

[32] Y.-M. Kim, N. J. Kim, and B.-J. Lee, Atomistic modeling of pure Mg and Mg–Al systems, Calphad **33**, 650 (2009).

[33] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen *et al.*, LAMMPS—A flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, Comput. Phys. Commun. **271**, 108171 (2022).

[34] J. Fernandez Troncoso and V. Turlo, Evaluating applicability of classical and neural network interatomic potentials for modeling body centered cubic polymorph of magnesium, Modell. Simul. Mater. Sci. Eng. **30**, 045009 (2022).

[35] While we have not shown the elastic constants calculated from our potential here, it is implicit in the very good agreement of the energies for large volume strains shown in Sec. III C 1.

[36] M. Stricker, B. Yin, E. Mak, and W. A. Curtin, Machine learning for metallurgy II. A neural-network potential for magnesium, Phys. Rev. Mater. **4**, 103602 (2020).

[37] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Phys. Rev. Lett. **98**, 146401 (2007).

[38] This corresponds to Monkhorst-Pack meshes of $15\times15\times8$ for hcp defect supercell, $15\times15\times4$ for dhcp and $16\times16\times16$ for fcc and bcc.

[39] Y. C. Wang and H. Q. Ye, On the tilt grain boundaries in hcp Ti with [0001] orientation, Philos. Mag. A **75**, 261 (1997).

[40] J. Wang and I. J. Beyerlein, Atomic structures of symmetric tilt grain boundaries in hexagonal close packed (hcp) crystals, Modell. Simul. Mater. Sci. Eng. **20**, 024002 (2012).

[41] B. Yin, M. Stricker, and W. A. Curtin, Pure magnesium DFT calculations for interatomic potential fitting, https://doi.org/10.24435/materialscloud:8f-1s (2020), accessed=30.11.2022

[42] A. Nayeb-Hashemi and J. Clark, *Phase Diagrams of Binary Magnesium Alloys* (ASM International, Metals Park, Ohio, 1988).

[43] E. V. Podryabinkin and A. V. Shapeev, Active learning of linearly parametrized interatomic potentials, Comput. Mater. Sci. **140**, 171 (2017).

[44] C. Zeni, A. Anelli, A. Glielmo, and K. Rossi, Exploring the robust extrapolation of high-dimensional machine learning potentials, Phys. Rev. B **105**, 165141 (2022).

[45] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, Phys. Rev. B **99**, 014104 (2019).

[46] G. Dusson, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord, and C. Ortner, Atomic cluster expansion: Completeness, efficiency and stability, J. Comput. Phys. **454**, 110946 (2022).

[47] I. Shin and E. A. Carter, Simulations of dislocation mobility in magnesium from first principles, Int. J. Plast. **60**, 58 (2014).

[48] C. Barrett, M. Nitol, and D. Dickel, Unraveling Mg ⟨c+a⟩ slip using neural network potentials, in *Magnesium Technology 2022* (Springer, New York, 2022), pp. 273–279.

[49] X. Liu, M. R. Niazi, T. Liu, B. Yin, and W. Curtin, A low-temperature prismatic slip instability in Mg understood using machine learning potentials, Acta Mater. **243**, 118490 (2023).

[50] See Fig. 7.

[51] https://doi.org/10.17617/3.A3MB7Z.

[52] https://github.com/eisenforschung/magnesium-mtp-training-data.

[53] J. Janssen, S. Surendralal, Y. Lysogorskiy, M. Todorova, T. Hickel, R. Drautz, and J. Neugebauer, pyiron: An integrated development environment for computational materials science, Comput. Mater. Sci. **163**, 24 (2019).

[54] S. Boeck, C. Freysoldt, A. Dick, L. Ismer, and J. Neugebauer, The object-oriented DFT program library S/PHI/nX, Comput. Phys. Commun. **182**, 543 (2011).

[55] S. A. Goreinov, I. V. Oseledets, D. V. Savostyanov, E. E. Tyrtyshnikov, and N. L. Zamarashkin, How to find a good submatrix, in *Matrix Methods: Theory, Algorithms And Applications: Dedicated to the Memory of Gene Golub* (World Scientific, Singapore, 2010), pp. 247–256.

[56] With VASP [16,17] we experienced crashes using the (rather high) k-point sampling that we employed in this paper. VASP [16,17] and S/PHI/nX [54] use a different energy reference, but that is not relevant here, since we are calculating energy differences only.

[57] For detailed explanation of this algorithm see [7,55].