Systematic Structure Datasets for Machine Learning Potentials: Application to Moment Tensor Potentials of Magnesium and its Defects

Marvin Poul, Liam Huber, Erik Bitzek and Jörg Neugebauer

Computational Materials Design, Max-Planck-Institut für Eisenforschung
Abstract

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.

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[21], dislocation arrays[6] or high-symmetry planar defects[11, 13]. 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[e.g. 5, 17] 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 methodology[32], 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.

We structure the paper as follows; first a brief review of Moment Tensor Potentials in section II A. Then follows the construction of the data set in section II C, which is the main idea we want to put forth. In sections II D & II E we discuss choosing cutoffs and the fitting of the potentials. Afterwards we verify of the potentials on defects and analysis in detail the influence of training data in sections III A & III B. This analysis supports of our main idea. We close with a brief comparison with active learning in section III C before we summarize our findings in section IV.

II. METHODS

A. Moment Tensor Potentials

Moment Tensor Potentials (MTP) are machine learning interatomic potentials originally introduced by A. Shapeev.[35] We will briefly review their formalism here, but leave the details to the original authors.[32]

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

$$E^{MTP} = \sum_{i}^{N} V(n_i) = \sum_{i}^{N} \sum_{\alpha} \xi_{\alpha} B_{\alpha}(n_i),$$

where $n_i$ is the atomic environment around atom $i$, $N$ is the total number atoms, $B_{\alpha}$ are the descriptor basis functions and $\xi_{\alpha}$ the linear expansion coefficients, which are determined during the fitting procedure.[22] The descriptor basis functions are defined as contractions of the Moment Tensors, $M_{\mu,\nu}$, defined in the single-component case as

$$M_{\mu,\nu}(n_i) = \sum_{j \in n_i} f_{\mu}(|r_{ij}|) \underbrace{r_{ij} \otimes \cdots \otimes r_{ij}}_{\nu \text{ times}}$$

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 $f_{\mu}$ are expanded in an orthogonal polynomial basis and equipped with a cutoff $R_c$, such that they are 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

\[ \text{lev } M_{\mu,\nu} = 2 + 4\mu + \nu . \]  

(3)

The level of the basis functions \( B_{\alpha} \) 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 \( l_{\text{max}} \). This implicitly defines up to what values of \( \mu \) and \( \nu \) 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.

B. Data Generation

All training data is generated using VASP\textsuperscript{[18,19]} with \( \Gamma \) centered \( k \)-meshes, \( 27 \times 27 \times 27 \) \( k \)-points and plane wave energy cutoff of 550 eV. While the structures vary in volume, we keep the \( k \) points constant to avoid discontinuities in the potential energy surface. By convergence testing we find the energies to be converged to 0.6 meV and the forces to \( 7 \times 10^{-5} \) eV/Å. These values represent the mean error of the training calculations with respect to a sample of 50 structures (350 in total) of each training set calculated at a \( 37 \times 37 \times 37 \) \( k \)-point mesh and a plane wave cutoff of 687.5 eV. The chosen \( k \) point setting corresponds to a \( k \) mesh spacing of 0.06 Å\textsuperscript{-1}.\textsuperscript{[24]} We used a fixed \( k \)-mesh for the training data to ensure smooth DFT data. Verification DFT data is generated with the same parameters except large grain boundary and surface structures where we use \( k \)-mesh spacing of 0.05 Å.

All potentials are fitted with respect to energies, forces and stresses from DFT, with weights of 1, 0.01 and 0.001 respectively. Forces are additionally weighted by \( 1/N \) where \( N \) is the number of atoms in the cell as per default in the MLIP\textsuperscript{32} package.

C. Training Set Construction

Classical potentials are often trained on a set of properties that it 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 not all relevant structures are considered. Another approach recently presented is to combine active learning and some form for structure generation (randomly, by molecular dynamics or Monte Carlo simulations)\[2, 3, 34, 38\]. 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\[author?\]\[38\] demonstrates that this worked 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.

There are also parallels to the work of Podryabinkin et al.\[34\]. 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 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 structure. We will show in this paper how to construct and utilize structures that are not energetically near the equilibrium structure.

Another approach that aims in a similar direction is the recent work from M. Karabin and D. Perez.\[15\]. Their work 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 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. Since this scheme makes to no reference to structure prototypes or crystallography it is a completely unbiased procedure in this sense. We have not attempted direct comparison between their and our approach however, since their work did not produce a final potential and restricted itself to analysis of the descriptors obtained in a training set.

As with any automated procedure, systematic biases or incompleteness may still occur; this is difficult to detect especially in a high dimensional space like atomic configuration.
space. In this paper therefore, we want to propose a third option that more systematically probes phase space and might elucidate its structure in the process.

In order to do so, 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, is random (periodic) crystal structures obtained from RandSPG\cite{1}. We generate these structures with one to ten Magnesium atoms per cell, asking for all possible space groups, 1-230, and allowing volumes $\pm 10\%$ around the equilbrium atomic volume of hcp Mg of $22.87 \text{Å}^3$/atom. We label this the RandSPG set. 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. We check the space groups obtained with SpgLib\cite{41}. Figure \ref{fig:1} shows the frequency of each crystal system in our initial data set. While not all systems are equally present, there is a sizeable number of structures available for each.

From this starting point, we then successively minimized the volume, cell shape, and the internal coordinates independently using VASP\cite{18,19}. These calculations are done at low convergence parameters, since they only serve to bring the structures near the equilibrium

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Frequency of crystal systems in the RandSPG set.}
\end{figure}
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. Nevertheless, we still represent a wide range of structurally distinct environments. 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.

As a final step in our process for creating training data, the structures from IntMin are randomly disturbed by one of triaxial (Hydro, up to 80\%) strain, shear strain (Shear, up to 80\%), or random displacements and small strains (Rattle, 0.5 Å mean displacement and up to 0.05\% strain).\[25\] For each of these modifications, every structure in IntMin experiences five different random perturbations of the appropriate nature.

?? 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 ?? are the number of structures (top number) and atoms (i.e. atomic environments, bottom number) in each structure set. We will examine the performance of potentials fitted to each of these sets compared to potentials fitted to the set of all structures, Everything.

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). To test the potentials we construct completely fresh data closer to the application domain of the potential. The results of this testing is discussed in section III A. 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 \[2\] (fig. 10 and the discussion sec. 4). Since we now rely on this 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.

D. Cutoff Radius Determination

MTPs are a local potential, 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, 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. We show that this task requires a detailed and explicit analysis. We therefore select a set of cutoffs which we thoroughly investigate later in the manuscript to determine their impact on the potential’s accuracy. Figure 2 shows the distribution of neighbor distances in the RandSPG and IntMin training sets. Also drawn are the HCP Mg shell distances at 22.87 Å³/atom (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; above 9 Å the histograms tails off since we only calculate the distances for the first 120 neighbors. While the RandSPG structures are mostly evenly distributed, the IntMin distributions shows distinct peaks. This is expected under energy minimization and gives us 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 section III A 1 and it therefore important to explicitly check what cutoffs may reasonably 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.

As the lower cutoff we pick 1.8 Å for all potentials which is the pseudo-potential cutoff used in our VASP[18][19] calculations.
FIG. 2: 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 22.87 Å³/atom.

E. Fitting Results

We fit MTP models for each of the data sets at different model complexities, choosing levels 8, 12, 16, 20 and 24. We use the MLIP[32] program, which performs energy, force, and stress matching.

We obtain energy, force and stress root mean square errors (RMSE) values after each fit. Energy RMSE are plotted in fig. 3a 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, as expected that larger, more diverse, structure sets naturally have a higher interpolation error.

In fig. 3b the same energy RMSE is plotted as a function of cutoff for three selected levels. 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.[26] 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 don’t necessarily seem to include more information. Thus,
FIG. 3: Energy training RMSE of the 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.

to optimize the numerical performance of a potential one should carefully check whether for the given descriptor level the cutoff is appropriately chosen.\cite{27}

F. 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 RSME as measure for the accuracy. To provide a measure for the computational cost we run MD on $6 \times 6 \times 6$ primitive hexagonally closed packed (hcp) Mg unit cells for 10000 steps at 50K. Additionally we calculated the RMSE for four classical potentials\cite{7,16,37,39} on the EVERYTHING set and their runtime in the same MD setup. Figure 4 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
FIG. 4: Time per force call per atom versus root mean square error of the energy. Colors symbolize the cutoff radius. Each line is constructed by using potentials of increasing level, from 6 to 26. 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.

(6.4 meV/atom, dashed line). At low-cost and low-accuracy, the Pareto front follows the potential fit with a cutoff of 5.2 Å before switching over to 6.5 Å at around 10 meV/atom error. While the 8.2 Å potential is strictly behind the Pareto front, we will later see in sections III A 4 and III A 1 that it is still useful due to superior performance when treating defects and different closed packed structures. The other classical potentials are orders of magnitude faster than our MTPs, yet fail to achieve an accuracy below 100 meV/atom. The plain computational cost of each potential as a function of level and cutoff are shown in section A for all potentials fit to EVERYTHING.
III. RESULTS

This section presents our resulting potentials in three sub sections. First their verification against explicit DFT calculations that were not in the training set; second, calculation of basic thermodynamic properties to ensure the potentials do not break during routine investigations; third, a brief look at active learning.

A. Verification

Since we do not split the fitting data into traditional train and test sets, we instead 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 section III.B. In total more 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. Static 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 5 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 for these the potential it clearly failed. The EVERYTHING set achieves ≈5 meV/atom in the 10% range and <10 meV/atom in 80% range.

Figure 6 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
FIG. 5: Energy RMSE 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. 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 section III C. This additional interpolation error for larger training sets comes at the advantage of a larger applicability as can be seen referring back to fig. 5. 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. 6 shows the benefit of using larger cutoffs, even though fig. 4 alone
FIG. 6: Energy difference to HCP per atom vs atomic volume for FCC and DHCP (columns); think 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 have trouble differentiating close 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. 14. 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 section III B

2. Phonons

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 Å³/atom where it is dynamically stable. All calculations were performed with LAMMPS\cite{40} and PHONOPY\cite{41} 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.

First as an example fig. 7 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. Troncoso et al.\cite{10} review this topic and find MEAM potentials are the best so far to study the dynamical behaviour of bcc Mg, but also report that the same potentials are deficient in their elastic properties.\cite{28} Neural Network Potentials\cite{12} 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 this review the authors concluded that these properties could be improved by specifically targeting bcc structures in the training set. 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
FIG. 7: 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.

related to the errors in the forces themselves and are therefore an interesting quantity for validation. Figure 8 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. Not shown in the graph, but clear from further inspection of the data is that the unrelaxed bcc structures dominate this error with about 0.05 eV/Å² at the highest level. The other two bulk structures are better described at 0.01 eV/Å². 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
FIG. 8: 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/Å² to 0.05 eV/Å².

are run with plane-wave cutoff as 550 eV and k-mesh spacing of 0.05 Å⁻¹. Structures are relaxed in their internal degrees of freedom and volume in DFT and MTP separately before the formation energy is calculated. Results are shown in fig. Each dot is the error in vacancy formation energy for each prototype and potential. For most training sets the error clusters around 10 meV to 100 meV. The minimized sets IntMin, CellMin and VolMin (to a lesser extent) show typical signs of overfitting at high potential levels, i.e. increase in error with level and unphysical results at highest level. RATTLE performs best on vacancies of all training sets, since it is the only set including structures with atoms shifted out of local equilibrium; We conclude that the RATTLE training set is the one that contains information relevant to the description of vacancies.

The absolute predictions of formation energy for just the potentials fit to the full data set are also shown in fig. 10a as a function of potential level. Horizontal lines are the DFT reference energies. The shaded regions on the MTP curves give the maximum spread in predictions from potentials fit with the same parameters and to the same data. Generally all three cut-off potentials manage to adequately (~0.1 eV) describe the vacancy even at moderate potential levels (~16). However, the spread in predictions is appreciably larger for the 5.2 Å potentials especially at lower levels. We therefore can conclude that at least potential cutoffs beyond 5.2 Å and potential levels beyond 16 are needed.
FIG. 9: Logarithmic Maximum prediction error in vacancy formation error. Each point gives the mean error in the vacancy formation energy for each potential with different level. The x-axis gives the logarithm of this error, the y-axis lists the training sets, while the color indicates the potential level. The y-axis is ordered by the mean error across prototypes and potential level. Subplots give data for each cutoff. The black lines connect the points of potentials with level 24.
(a) Predicted energy of vacancy formation.

(b) Maximum displacement during minimization of the vacancy super cell across atoms and dimensions.

FIG. 10: Comparisons between DFT reference and MTP prediction in formation energy and maximum displacement during minimization of vacancies. Shown here only potentials fitted to EVERYTHING. Horizontal lines are DFT references, shading the spread in prediction from multiple potentials at the same level.

Next to the vacancy formation energies we also tested the performance of the potentials at predicting structural relaxations. Figure 10b shows the maximum displacement (across all atoms and dimensions) during the minimization of the vacancy super cell. The shaded regions on the MTP curves give the maximum spread in predictions from potentials fit with the same parameters and to the same data. As in fig. 10a 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 $\Sigma 7 bA$ and a basal reflection twin with lattice parameter $a = 3.195 \text{Å}$ 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 $\Sigma 7 b$ also has a second realization, called $T$-type\cite{22}, but they are very close in energy and they are not included in the calculations below, though 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

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

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.\textsuperscript{11} for the final potentials fitted to EVERYTHING. All three cutoffs initially underestimate the surfaces energy initially 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. In contrast the two grain boundaries are much better described at all potential levels and cutoffs. For the EVERYTHING set and $R_c = 8.2 \text{Å}$ 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.

This sections shows that our potential can faithfully reproduce planar defects and surfaces without having seen them during training explicitly. We interpret this surprising 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 the planar defects that we tested.
FIG. 11: Correlation plot of planar defect excess energies; DFT vs. MTP predicted. Each point is a potential fit the given cutoff and level highlighted by the hue. The dashed 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. However note, that except for the highest cutoff, even the highest levels systematically underestimate the surface energies.
B. Discussion

We now collect the errors against DFT calculated in the previous sections for all training sets and test domains in fig. 12. 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. 3b, 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 result 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. 3a, which we can understand since this is also the volume range present in each training set. On the substantially larger 80% set only Hydro (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, Hydro and Shear perform well, but also Rattle is in between both sets. Even though rattle only experience 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 section III A 1 and figs. 5 and 6. 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 cited for the planar defects and surfaces are given normalized to the area. Surfaces however are less well described than the planar defects, since they are included in the training set only indirectly due to the VolMin set, as indicated earlier. 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.

Finally, note that the RandSPG set mostly makes decent predictions and seems therefore to be a very solid starting point, to be supplemented only with few structures specific to the application domain, e.g. from active learning.
FIG. 12: 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”, ”hydrostatic”, 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”, ”GB” it is the RMSE of the respective energy excess of formation.
C. Active Learning Comparison

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\cite{32,33} is reviewed in appendix B 1. We present calculations that show the potentials fitted in this work 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\cite{32}. 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 $4 \times 4 \times 4$ super cell at moderate temperatures and ambient pressure ($T$ in 100, 400, 700K); an HCP twin boundary and a $\Sigma 7b$ grain boundary at 600 K and ambient pressure; and finally fcc and dhcp 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 threshold, $\gamma_{\text{break}} = 5$, 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 B 1 for their precise definitions. Within the given time frame none of the potentials trained on the full training set encountered structures where $\gamma > \gamma_{\text{select}}$, i.e. there was no evidence of extrapolation. Appendix B 1 also shows exemplary calculations of the extrapolation grade over simulation time for potentials fitted on less than the full data set.

We can now verify our hypothesis explaining the success of the potential in section III A 4; if the accuracy were due to good extrapolation, the active learning would still flag the structures as outside of the active set. Since that did not happen, there must be structures 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 section III A 1, we can now clearly state that this is an interpolation error, since dhcp and fcc structures were not identified as extrapolation by MLIP\cite{32}. This is in contrast to work performed by \cite{43}. There, they fit Atomic Cluster Expansion (ACE)\cite{8} 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[9, 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 [43], shows that the choice of the structure set is critical. Their underlying assumption seemed to have been that structures from liquid must necessarily be around the underlying equilibrium ice structure and that hence a convex hull around the high temperature structure includes the equilibrium ones. But that doesn’t need to be case, since the topology of descriptor space is completely unclear at this stage of the field.[30]

On the other hand the original question they set out to answer and that has been discussed in the literature before is how 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 of planar defects at least with respect to the applications tested here. Still, users that wish to use the potentials on defects we have not verified here, i.e. on isolated dislocation may need to do additional testing, but given that the $\Sigma 7b$ interface is comprised of dislocation cores [36] we do not expect significant failure. In any case we expect the potentials to be excellent starting points for additional active learning where necessary.

D. Potential Properties

As a final test of the stability of the fitted potentials in a wide range 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 $4 \times 4 \times 4$ unit cells for $1 \times 10^6$ MD steps.

Figure 13a shows the change of internal energy, volume and lattice parameters of hcp Mg with temperature at zero pressure. 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
FIG. 13: Results from $NpT$ MD simulations potentials fit to EVERYTHING and cutoff 8.2 Å. Simulations for the other two cutoffs were run, but do not seem to appreciatively differ from the ones shown. (a) Average energy as a function of temperature. Melting is indicated by the abrupt change the temperature range 800 K to 900 K depending on potential level. (b) Average energy as a function of hydrostatic pressure.

We investigate a pressure window from $-3 \text{ GPa}$ to 12 GPa. Instabilities occur at large tensile pressures and elevated temperatures. The lowest pressure that leads to unstable simulation boxes for hcp is $-0.5 \text{ GPa}$ at 800 K, well above the tensile strength of pure magnesium, so that the potentials could still be used in a simulated tensile test setup, though we have not verified the quality of the predictions there explicitly. We note that higher level potentials seem to be able to bear more tensile pressure than lower level ones before becoming unstable.

IV. CONCLUSION

We demonstrate that a physically motivated construction of the training set, which systematically covers all bulk crystal symmetries instead of just low energy structures, allows
the successful construction of ML-potentials that transfer well to bulk defects. Importantly, the training database does not include any explicit defect structures or additional data from active learning, see figs. 10a and 11.

In figs. 3b and 11 we show that the cutoff radius warrants more care than is sometimes paid in literature: It is crucial for the transferability of the potential to unseen structures (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 further understand this effect.

In section III C 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 non-optimal 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[4, 32] 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 the cutoff 8.2 Å, to avoid wrong (zero) stacking energies predicted for low level potentials[31]. Lower levels and cutoffs may be used to save computational resources, though 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 systematic physical data sets by systematically sampling bulk structures over all bulk symmetries can be straightforwardly applied to other materials systems. Tools like RANDSPG[1] facilitate the construction of such data sets in a systematic and automated fashion. This opens the route towards a largely automatized generation of transferable and accurate potentials.
DATA AVAILABILITY

The full DFT training set is uploaded to [https://github.com/eisenforschung/magnesium-mtp-training-data](https://github.com/eisenforschung/magnesium-mtp-training-data). This repository also contains the final potential files necessary to run simulations with LAMMPS[40]. We have also added example jupyter notebooks and a PYIRON[14] project that shows how to access the data and run simple simulations.

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. We also thank Ralf Drautz, whose potential fitting experience benefited us during this study.

[1] Patrick Avery and Eva Zurek. Randspg: An open-source program for generating atomistic crystal structures with specific spacegroups. Computer Physics Communications, 213:208–216, 2017.
[2] Jörg Behler. First principles neural network potentials for reactive simulations of large molecular and condensed systems. Angewandte Chemie International Edition, 56(42):12828–12840, 2017.
[3] Noam Bernstein, Gábor Csányi, and Volker L. Deringer. De novo exploration and self-guided learning of potential-energy surfaces. npj Computational Materials, 5(1):99, Oct 2019.
[4] Anton Bochkarev, Yury Lysogorskiy, Sarath Menon, Minaam Qamar, Matous Mrovec, and Ralf Drautz. Efficient parametrization of the atomic cluster expansion. Phys. Rev. Materials, 6:013804, Jan 2022.
[5] Patrick R. Cantwell, Timofey Frolov, Timothy J. Rupert, Amanda R. Krause, Christopher J. Marvel, Gregory S. Rohrer, Jeffrey M. Rickman, and Martin P. Harmer. Grain boundary complexion transitions. Annual Review of Materials Research, 50(1):465–492, 2020.
[6] Emmanuel Clouet, Lisa Ventelon, and F. Willaime. Dislocation core energies and core fields from first principles. Phys. Rev. Lett., 102:055502, Feb 2009.
[7] Doyl E Dickel, Michael I Baskes, Imran Aslam, and Christopher D Barrett. New interatomic potential for mg–al–zn alloys with specific application to dilute mg-based alloys. *Modelling and simulation in materials science and engineering*, 26(4):045010, 2018.

[8] Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019.

[9] Geneviève Dusson, Markus Bachmayr, Gábor Csányi, Ralf Drautz, Simon Etter, Cas van der Oord, and Christoph Ortner. Atomic cluster expansion: Completeness, efficiency and stability. *Journal of Computational Physics*, 454:110946, 2022.

[10] Javier Fernandez Troncoso and Vladyslav Turlo. Evaluating applicability of classical and neural network interatomic potentials for modeling body centered cubic polymorph of magnesium. *Modelling and Simulation in Materials Science and Engineering*, 2022.

[11] Daniel 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, Jan 2010.

[12] Sergei A Goreinov, Ivan V Oseledets, Dimitry V Savostyanov, Eugene E Tyrtyshnikov, and Nikolay L Zamarashkin. How to find a good submatrix. In *Matrix Methods: Theory, Algorithms And Applications: Dedicated to the Memory of Gene Golub*, pages 247–256. World Scientific, 2010.

[13] Liam Huber, Jörg Rottler, and Matthias Militzer. Atomistic simulations of the interaction of alloying elements with grain boundaries in mg. *Acta Materialia*, 80:194–204, 2014.

[14] Jan Janssen, Sudarsan Surendralal, Yury Lysogorskiy, Mira Todorova, Tilmann Hickel, Ralf Drautz, and Jörg Neugebauer. pyiron: An integrated development environment for computational materials science. *Computational Materials Science*, 163:24–36, 2019.

[15] Mariia Karabin and Danny Perez. An entropy-maximization approach to automated training set generation for interatomic potentials. *The Journal of Chemical Physics*, 153(9):094110, 2020.

[16] Young-Min Kim, Nack J Kim, and Byeong-Joo Lee. Atomistic modeling of pure mg and mg–al systems. *Calphad*, 33(4):650–657, 2009.

[17] Sandra Korte-Kerzel, Tilmann Hickel, Liam Huber, Dierk Raabe, Stefanie Sandlöbes-Haut, Mira Todorova, and Jörg Neugebauer. Defect phases – thermodynamics and impact on material properties. *International Materials Reviews*, 67(1):89–117, 2022.
[18] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54:11169–11186, Oct 1996.

[19] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15–50, 1996.

[20] AA Nayeb-Hashemi and JB Clark. Phase diagrams of binary magnesium alloys (1988). *ASM International, Metals Park, Ohio*.

[21] 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, Dec 2010.

[22] 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 usually machine learning formalism, this factor is absorbed into the basis coefficients $\xi_\alpha$.

[23] For detailed explanation of this algorithm see [12, 52].

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

[25] For the HYDRO and SHEAR sets the respective entries in the strain tensor are chosen from uniform random distribution in range ±80%.

[26] 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 work.

[27] Comparisons of the runtime cost as a function of cutoff radius are explicitly given in fig. 14.

[28] 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 section III A 1.

[29] This corresponds to Monkhorst-Pack meshes of $15 \times 15 \times 8$ for hcp defect supercell, $15 \times 15 \times 4$ for dhcp and $16 \times 16 \times 16$ for fcc and bcc.

[30] One may think of a three dimensional cone where the tip represents the equilibrium structure and the cone surface represents liquid structures with increasing temperatures close to the flat base of the cone. In this case, one may be fooled by looking through a lower dimensional projection of the cone into thinking that this assumption holds. But when constructing the convex hull in the original 3D space around the liquid points at the base, the tip of the cone is not included. This is only to show that the observations from Zeni et al. [43] do not strictly imply that MLIP always extrapolate in production, it may also be that their training sets
were insufficiently diverse.

[31] See fig. 6.

[32] Ivan S Novikov, Konstantin Gubaev, Evgeny V Podryabinkin, and Alexander V Shapeev. The MLIP package: moment tensor potentials with MPI and active learning. *Machine Learning: Science and Technology*, 2(2):025002, Jan 2021.

[33] Evgeny V. Podryabinkin and Alexander V. Shapeev. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017.

[34] Evgeny V. Podryabinkin, Evgeny V. Tikhonov, Alexander V. Shapeev, and Artem R. Oganov. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Phys. Rev. B*, 99:064114, Feb 2019.

[35] Alexander V Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3):1153–1173, 2016.

[36] Ilgyou Shin and Emily A. Carter. Simulations of dislocation mobility in magnesium from first principles. *International Journal of Plasticity*, 60:58–70, 2014.

[37] DE Smirnova, SV Starikov, and AM Vlasova. New interatomic potential for simulation of pure magnesium and magnesium hydrides. *Computational Materials Science*, 154:295–302, 2018.

[38] Justin S Smith, Benjamin Nebgen, Nithin Mathew, Jie Chen, Nicholas Lubbers, Leonid Burakovsky, Sergei Tretiak, Hai Ah Nam, Timothy Germann, Saryu Fensin, et al. Automated discovery of a robust interatomic potential for aluminum. *Nature communications*, 12(1):1–13, 2021.

[39] D. Y. Sun, M. I. Mendelev, C. A. Becker, K. Kudin, Tomorr 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, Jan 2006.

[40] 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, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.*, 271:108171, 2022.

[41] A Togo and I Tanaka. First principles phonon calculations in materials science. *Scr. Mater.*, 108:1–5, Nov 2015.
FIG. 14: 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.

[42] J Wang and I J Beyerlein. Atomic structures of symmetric tilt grain boundaries in hexagonal close packed (hcp) crystals. *Modelling and Simulation in Materials Science and Engineering*, 20(2):024002, feb 2012.

[43] Claudio Zeni, Andrea Anelli, Aldo Glielmo, and Kevin Rossi. Exploring the robust extrapolation of high-dimensional machine learning potentials. *Phys. Rev. B*, 105:165141, Apr 2022.

Appendix A: Cost-Accuracy Tradeoff

Figure 14 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 versus level as expected, since the number of free parameters in the potential increases exponentially with the level. Increments in radial cutoff only change the prefactor of the scaling, which is also expected of local models.
Appendix B: 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 whether this configuration can be safely approximated by the potential or not. \[^32\] 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. \[^23\] Figure \[15\] 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

\[
c = A^{-1}b. \tag{B1}
\]

Now if all \{\(c_i\)\} are smaller than unity, \[^32\] define this configuration to be within the interpolative region where small fitting errors can be assumed. If however one of the 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

\[
\gamma(\text{cfg}) = \max_i |c_i| = \max_i |A^{-1}_{ij}b_j(\text{cfg})| \tag{B2}
\]

where \(\gamma \leq 0\) in the first case discussed above and \(\gamma > 1\) in the second case. The active learning regime implemented in MLIP\[^32\] then consists of running any desired simulation protocol but keeping tracking of \(\gamma\) for all neighborhoods in the considered structure. They define two thresholds, which we draw also in fig. \[15\] with yellow and red lines. Exceeding the first (yellow), $\gamma_{\text{select}}$, causes the whole structure to be written out for later consideration. Exceeding the second (red), $\gamma_{\text{break}}$, 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 re-run the simulation. This is repeated until the full simulation runs without exceeding $\gamma_{\text{break}}$. 

33
FIG. 15: 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 that inside the yellow borders (green) are assumed to be approximated well by the potential; between the yellow and the red border (yellow) are selected are further training and outside the red border (red) cause simulations to terminate.
FIG. 16: Extrapolation grade as a function of time during three simulations of liquid Mg at 2500 K and 5 GPa. Potentials can be seen interpolating initially, then exceed the selection threshold and even the breaking threshold in one case.

2. Example Simulation for Extrapolation Grade

To illustrate the concept of the extrapolation grade we run three simulations of liquid Mg at 2500 K and 5 GPa and track the extrapolation grade over time. The potential used here was fitted only to the RANDSPG set and is therefore not as transferable as the one fitted to EVERYTHING described in the main text.