ARTICLE   OPEN

Check for updates

Training data selection for accuracy and transferability of interatomic potentials

David Montes de Oca Zapiain [1 ✉], Mitchell A. Wood[1], Nicholas Lubbers [2], Carlos Z. Pereyra [3], Aidan P. Thompson [1] and Danny Perez [4 ✉]

Advances in machine learning (ML) have enabled the development of interatomic potentials that promise the accuracy of first principles methods and the low-cost, parallel efficiency of empirical potentials. However, ML-based potentials struggle to achieve transferability, i.e., provide consistent accuracy across configurations that differ from those used during training. In order to realize the promise of ML-based potentials, systematic and scalable approaches to generate diverse training sets need to be developed. This work creates a diverse training set for tungsten in an automated manner using an entropy optimization approach. Subsequently, multiple polynomial and neural network potentials are trained on the entropy-optimized dataset. A corresponding set of potentials are trained on an expert-curated dataset for tungsten for comparison. The models trained to the entropy-optimized data exhibited superior transferability compared to the expert-curated models. Furthermore, the models trained to the expert-curated set exhibited a significant decrease in performance when evaluated on out-of-sample configurations.

INTRODUCTION

The rapid adoption of machine learning (ML) methods in virtually all domains of physical science has caused a disruptive shift in the expectations of accuracy versus computational expense of data-driven models. The diversity of applications arising from this swell of attention has brought about data-driven models that have accelerated pharmaceutical design[1–3], material design[4,5], the processing of observations of celestial objects[6], and enabled accurate surrogate models of traditional physical simulations[7–9]. While many of these models have proved extremely powerful, new questions and challenges have arisen due to the uncertainty in model predictions coined as extrapolations[10], i.e., when prediction occurs on input that are found outside of the support of the training data. Moreover, the accuracy of a machine-learned model can only be quantified using the training itself, or on a subset thereof, held out as validation. For that reason, it is often extremely difficult to predict "real-world" performance where unfamiliar inputs are likely to be encountered.

In this manuscript, we focus on application to classical molecular dynamics (MD), which is a powerful technique for exploring and understanding the behavior of materials at the atomic scale. However, performing accurate and robust large-scale MD simulations is not a trivial task because this requires the integration of multiple components, as Fig. 1 schematically shows. A key component is the interatomic potential (IAP)[11–15], i.e., the model form that maps local atomic environments to energies and forces needed to carry out a finite time integration step. An accurate IAP is critical because large-scale MD simulations using traditional quantum ab initio calculations such as Density Functional Theory (DFT) are prohibitively expensive beyond a few hundred atoms. Given a local (i.e., short-ranged) IAP, MD simulations can leverage large parallel computers since the calculation of forces can efficiently be decomposed into independent computations that can be concurrently executed[16],

thus enabling extremely large simulations[17,18] that would be impossible with direct quantum simulations. However, a critical limitation of empirical IAPs is that they are approximate models of true physics/chemistry and as such have to be fitted to reproduce reference data from experimental or quantum calculations. ML techniques have recently enabled the development of IAPs that are capable of maintaining an accuracy close to that of quantum calculations while retaining evaluation times that scale linearly, presenting significant computational savings over quantum mechanics. This is due to their inherent ability to learn complex, non-linear, functional mappings that link the inputs to the desired outputs[19–21]. Nevertheless, despite significant advances in the complexity of the behavior that can be captured with ML-based models, machine learning interatomic potentials (MLIAPs) often struggle to achieve broad transferability[22–24].

Indeed, increasing the complexity of the ML models, while useful to improve accuracy, is not sufficient to achieve transferability. In fact, it could even be detrimental, as more data could be required to adequately constrain a more complex model. In other words, the more flexible the IAP model form, the more critical the choice of the training data becomes. Generating a proper training set is not a trivial task given the fact that the feature space the models need to adequately sample and characterize, which is the space that describes local atomic environments, is extremely high dimensional. Consequently, MLIAPs are typically trained on a set of configurations that are deemed to be most physically relevant for a given material and/or to a given application area[25], as determined by domain experts. Numerous examples now demonstrate that high-accuracy predictions are often achieved for configurations similar to those found in the training set[23,24,26–29]. However, the generic nature of the descriptors used to characterize local environment[26,29–31], coupled with the inherent challenge in extrapolating with ML methods, often lead to poor transferability. This challenge can be addressed in two

[1]Sandia National Laboratories, Albuquerque, NM 87185, USA. [2]Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [3]Mechanical and Aerospace Engineering, University of California Davis, Davis, CA 95616, USA. [4]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. ✉email: dmonte@sandia.gov; danny_perez@lanl.gov
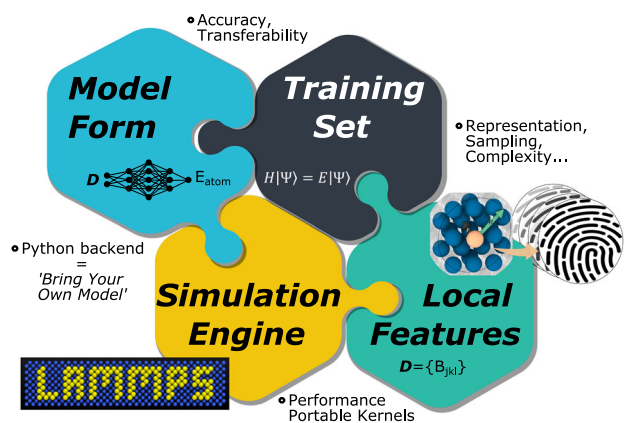
Fig. 1 Schematic for the component parts needed for a scalable and accurate Molecular Dynamics simulation utilizing a machine learned interatomic potential. Advances made herein correspond to the *Model Form - Training Set* pair whereas the computational aspects of the remaining pairs have been detailed elsewhere[16].

ways: i) by injecting more physics into the ML architectures so as to constrain predictions or reproduce known limits[9,32,33], or ii) by using larger and more diverse training sets to train the MLIAP so that a majority of the atomic environments encountered during MD simulations are found within the support of the training data[34]. In this work, we explore the second option, which is generally applicable to all materials and ML architectures. While conceptually straightforward, a challenge presented to this approach is that relying on domain expertise to guide the generation of very large training sets is not scalable, and runs the risk of introducing anthropogenic biases[35]. Therefore, the development of scalable, user-agnostic, and data-driven protocols for creating very large and diverse training sets is much preferable. Additionally it is worth stressing that this problem, and solutions herein, are not confined to the development of IAP, but apply to nearly all supervised ML training problems.

The key objective of this paper is to demonstrate the validity of a diverse-by-construction approach for the curation of training sets for general-purpose ML potentials whose transferability greatly exceeds what can be expected of potentials trained to human-crafted datasets. We believe that this approach fills an important niche for the many application domains where it is extremely difficult for users to a priori identify/enumerate all relevant atomic configuration (e.g., when simulating radiation damage, shock loading, complex microstructure/defect effects, etc.).

In the following, we demonstrate improvements to the training set generation process based on the concept of entropy optimization of the descriptor distribution[36]. We leverage this framework to generate a very large ($>2 \cdot 10^5$ configurations, $>7 \cdot 10^6$ atomic environments) and diverse dataset for tungsten (hereby referred to as the *entropy maximized*, EM) set in a completely automated manner. This dataset was used to train MLIAP models of atomic energy of various complexity, including neural-network-based potentials, as well as linear[28] and quadratic SNAP potentials[37]. The performance of EM-training was compared to equivalent models trained on a human-curated training set used for developing an MLIAP for W/Be (referred to as the *domain-expertise, DE* set below)[22,27].

While Tungsten is chosen as the material of interest, none of the results are specific to the choice of material or elemental species. The results show that EM-trained models are able to consistently and accurately capture the vast training set. We demonstrate a very favorable accuracy/transferability trade-off with the EM training set, where extremely robust transferability can be obtained at the cost of a relatively small accuracy decrease on configurations deemed important by experts. In contrast,

potentials trained on conventional DE datasets are shown to be mildly more accurate on configurations similar to those contained in the training set, but suffer catastrophic increase in errors on configurations unlike those found in the training set. While it warrants further study, a hybridization of these training data types was shown to offer an improved accuracy on DE configurations, see Supplemental Note 11.

Furthermore, the DE-trained MLIAPs exhibited significant sensitivity to the classes (expert-defined groupings) of configurations that were used for training and validation. Specifically, large errors were observed when the training and validation sets contained subjectively different manually-labeled classes of configurations instead of a random split (cf. Supplemental Note 8). These results highlight the perils of relying on physical intuition and manual enumeration to construct training sets. In contrast, the EM-based approach is inherently scalable and is fully automatable, since it does not rely on human input. Therefore, making the generation of very large diverse-by-design training sets, and hence the development of accurate and transferable MLIAPs, possible.

RESULTS AND DISCUSSION

Representation and sampling

In order to avoid over fitting complex ML models, many popular techniques such as compressive sensing, principal components, drop-out testing, etc., are employed to test if the model complexity exceeds that of the the underlying data.

While we only have tested a pair of training sets, the objective is to understand how complex of a model is needed to capture the ground truth (quantum accuracy) embedded in either training set. It is important to note that even though the discussion will be focused on the specific application to interatomic potentials, the complexity of the training set is an important factor in any supervised machine learning task.

We first characterize and contrast the characteristics of training sets generated by the two aforementioned methods. The domain expertise(DE) constructed dataset, containing about $1 \cdot 10^4$ configurations and $3 \cdot 10^5$ atomic environments, was created in order to parameterize a potential for W/Be[27], building on a previous dataset developed for W[22]. The data was manually generated and labeled into twelve groups (elastic deformations of BCC crystals, liquids, dislocation cores, vacancies, etc.) so as to cover a range of properties commonly thought to be important[25,38]. The second set was generated using an automated method that aims at optimizing the entropy of the descriptor distribution[36], as described in *Maximization of the Descriptor Entropy*.

Figure 2 compares the DE set to the EM set in terms of the distribution of configuration energy (left panel) and of three low-rank bispectrum components[28,39] (right panel), which we use to represent the atomic environment (cf. *Methods*)[26,28]. Observed in the energy distribution, the EM set is very broad in comparison with the DE set which is strongly peaked at energies close to the known ground-state BCC energy of -8.9 eV/atom, extending to about -7.5 eV/atom. In contrast, the energy distribution of the EM set spans energies between -8.5 and -5.0 eV/atom, peaking at around -8 eV/atom. Due to the large number of configurations in the EM set, a sizeable number of configurations are also located in the tails of the distribution; 442 (0.19%) below -8.5 eV/atom and 656 (0.29%) above -5.0 eV/atom. While the overlap between the energy distributions of the two sets is limited, it will be shown below that MLIAPs trained on the EM set can accurately capture the energetics of these low-energy DE configurations.

The three different probability density plots (Fig. 2 right panel) show that the descriptor distribution of the EM set is also much broader and more uniform than the DE set (note the logarithmic probability scale). To quantitatively compare the bispectrum
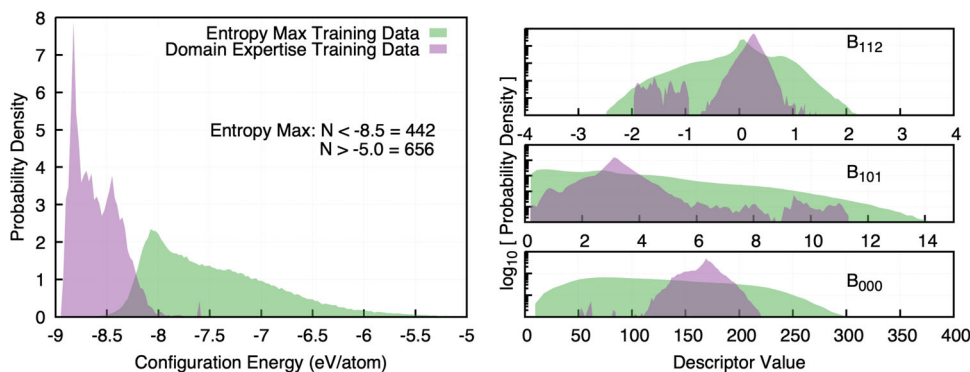
Fig. 2 Distributions of energy and bispectrum components of the configurations generated using the entropy maximization (EM) framework and using domain expertise (DE). (Left) Distribution of potential energy predicted by density functional theory. (Right) Distribution of three different low-rank bispectrum components, see *Methods* for details on descriptor values.
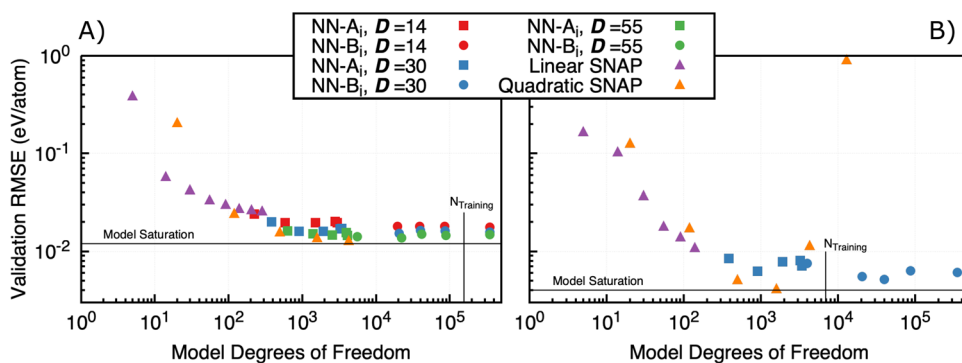

Fig. 3 RMSE validation errors for as a function of the number of degrees of freedom for different types of models trained and tested on distinct random samples from their respective dataset. A EM set; **B** DE set. Each model form and complexity shows a saturation of the validation errors indicated by the solid horizontal line. Slight improvement in NN predictions is seen with increasing D.

distributions, we compared to covariance of both datasets in the frame of reference where the DE data has mean zero and unit covariance in the space spanned by the first 55 bispectrum components (as defined by a ZCA whitening transformation). In that same frame, the EM descriptor distribution is broader in 51 dimension and at least 5× broader in 27 dimensions, strongly suggesting that Fig. 2 provides a representative view of the actual behavior in high dimension. In the four dimensions where the entropy set is slightly narrower, it is so by a factor between 0.5 and 0.9. This is consistent with the relative performance of the MLIAPs trained and tested on the different datasets, as will be shown in *Transferability*.

In order to further demonstrate the increased diversity and uniformity of the EM dataset relative to that of the DE set, we performed Principal Component Analysis (PCA) in order to represent the training data of each dataset with a reduced number of dimensions[40,41]. Supplemental Note 1 shows that the PCA representation of the DE set isolates into multiple clusters significantly in the first two PC dimensions. On the other hand the PC representation of the EM set is a single cluster that is dense and compact in the first two PC dimensions. Therefore, further demonstrating the overall more uniform and diverse sampling of the descriptor space in EM training. A projection of the DE training into the principal components of EM data is given Supplementary Fig. 14 resulting in the same conclusion.

It is important to note that the choice of bispectrum components as descriptors of the local atomic environment is not the only possible representation. Entropy maximization in another descriptor space (e.g., moment tensors, atomic cluster expansion) will result in a somewhat different distribution of configurations. However, we expect the qualitative features

highlighted here using bispectrum descriptors, including the fact that the DE set is largely contained within the support of the EM set, to be robust. Supplemental Note 13 further demonstrates this as it shows atomic environments generated using the EM method with bispectrum components still encompass DE when using Smooth Overlap of Atomic positions as descriptors[26,30,39,42].

Accuracy limits

While a very diverse training set is a priori preferable with regards to transferability, accurately capturing the energetics of such a diverse set of configurations could prove challenging. It is therefore important to assess how the choice of model form/ complexity affects the relative performance of MLIAPs trained and tested on the EM and DE sets. To do so, we consider a broad range of models, ranging from linear, quadratic, and neural network forms, in increasing levels of complexity. In all cases, the input features are the bispectrum descriptors that were used to generate and characterize the training sets. The models used can be seen as generalizations of the original SNAP approach[28,37,43]. Details of the model forms tested, as well as information on the fitting process can be found in the *Methods* section and in the examples provided in the Supplemental Information.

The accuracy of the trained models was assessed by quantifying the error on a validation set of configurations randomly held-out of the training process. We first evaluated the performance of each one of the different models for predicting the energy of configurations that were generated with the matching framework (i.e., EM or DE) on which the model was trained. Figure 3 reports the accuracy, quantified as the Root Mean Squared Error (RMSE), of the models on their respective validation set.
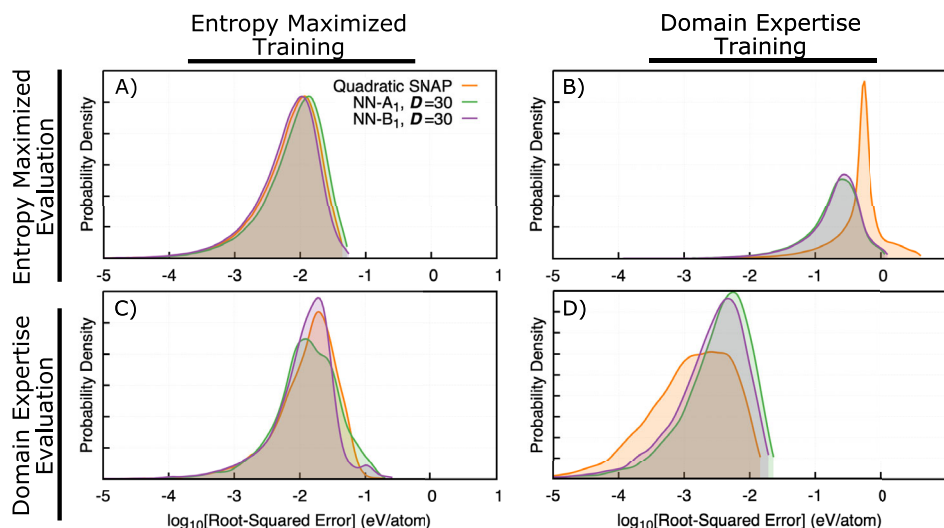
Fig. 4 Distribution of RSE errors for different combination of training and testing data. A Trained on EM and validated on EM; **B** trained on DE and validated on EM; **C** Trained on EM and validated on DE; **D** trained on DE and validated on DE. Only errors within three standard deviations from the observed mean value, about ~99% of the data, are reported to clearly convey the shape of the distribution.

In order to compare models of variable complexity, Fig. 3 displays validation RMSE against the number of free parameters N_{DoF} that are optimized in the regression step of the different ML models. For the simplified ML models (linear and quadratic SNAP) N_{DoF} is determined uniquely by the number of descriptors, $D = \{B_{jkl}\}$, where D denotes the number of bispectrum component used to characterized the atomic environments and is determined by the level of truncation used. However, for NN models N_{DoF} also depends on the number of hidden layers and the number of nodes per layer (cf. *Methods* section for more details). Figure 3A and B shows that in spite of the different nature of the models evaluated, the performance of all model classes remarkably asymptotes to roughly the same error, including the deep NNs where $N_{DoF} \geq N_{Training}$. This asymptote occurs after about 10^3 DoF, which is much less than the training size of ~10^5 for the EM set. The value of the limiting error is observed to slightly decrease with increasing number of descriptors. Finally, the errors are observed to saturate at a lower value when training and validating on the DE set (~$4 \cdot 10^{-3}$ eV/atom) than on the EM set (~10^{-2} eV/atom). This is perhaps unsurprising given the comparatively more compact and less diverse nature of the DE set, which makes it more likely that the validation set contains configurations that are relatively similar to configurations in the training set, a point that we will expand on the following paragraphs.

Figure 3B also shows that NN models surprisingly do not overfit to the training data even when $N_{DoF} \geq N_{Training}$, where one could expect the RMSE value of the validation set to increase. This is presumably caused by the non-convex nature loss function which makes it more difficult to access very low loss minima that would lead to overfitting. In addition, it is important to mention that our training protocol also reduced the learning rate when the validation loss plateaued (cf. *Methods* for the detailed protocol). In contrast, quadratic SNAP models, cf. Fig. 3B—for which the loss function is convex—show clear signs of overfitting to the DE set when $N_{DoF} \simeq N_{Training}$, an observation that was previously reported[25,37].

These results demonstrate that the details of the MLIAP's architectures appear to have limited impact on the ultimate accuracy when training and validation data are sampled from similar distributions, so long as the model is sufficiently flexible. The following section shows that the choice of the training data, not the model form, is the key factor that determines the transferability of the models.

Transferability

While the accuracy/transferability trade-off has been evident for many years for traditional IAPs that rely on simple functional forms, the development of general ML approaches with very large numbers of DoF in principle opens the door to MLIAPs that would be both accurate and transferable. However, as discussed above, the introduction of flexible and generic model forms can in fact be expected to make the selection of the training set one of the most critical factors in the development of robust MLIAPs.

In order to assess the relative transferability of the models trained on the different datasets, we select three models for further analysis: the NN-A_1, NN-B_1, and a quadratic SNAP model, all using an angular momentum limit of $J_{max} = 3$ which corresponds to $D = 30$. These three models have 383, 3965, and 495 degrees of freedom, respectively. Also, all of these models are on the saturation regime where the model performance asymptotes to roughly the same error. Figure 4 reports the distribution of the root squared error (RSE) from the different models trained and validated on the four possible combinations of DE and EM. Specifically, Fig. 4A, D report the RSE distributions predicted for configurations selected from the same set as the one used to train the model (e.g., when both training and validation are done on DE data), while panels B and C reports the error distributions when validation configurations are chosen from a different set (e.g., when training is done on DE data but validation is done on EM data and vice versa). The detailed errors corresponding to these four different possibilities, as well as to additional MLIAP models, are reported in the Supplementary Notes 4, 5, 6, and 7.

Figure 4A, D shows that the both sets of models exhibit low errors when predicting the energy of randomly held-out configurations sampled from the set the model was trained on (DE or EM), consistent with the results shown in the previous section. In both cases, the distribution peaks around 10^{-2} eV/atom and rapidly decays for larger errors, with long tails toward smaller error values. The models trained and validated on the DE data (panel D)) show a slightly heavier tail at low errors than the model trained and validated on the EM data, in agreement with the slightly lower asymptotic errors reported in Fig. 3. Otherwise, the behavior of the different models is similar, except for the quadratic-SNAP model from panel D) which shows slightly lower errors.

Transferability of a model is quantified by predicting on configurations sampled from a different dataset than the one used

for training, Fig. 4B, C. The performance of the two sets of models now show dramatic differences. Figure 4B shows a very large increase in errors, by almost two orders magnitude, when predicting the energy of configurations sampled from the EM set using models trained to DE data. Supplemental Note 9 addresses how these prediction errors are concentrated with respect to the high energy configurations that are clear extrapolations of the model.

In contrast, Fig. 4C shows only a modest increase in error when predicting the energy of configurations sampled from the DE set using models trained to EM data. In other words, models trained on a compact dataset that is concentrated in a small region of descriptor space (such as the DE set) can be very accurate, but only for predictions that are similar to the training data because excursions that force the trained MLAIPs to extrapolate out of the support of their training data leads to extremely large errors. Conversely, models trained to a very broad and diverse dataset might have comparatively slightly larger errors when validating over the same diverse dataset, but, perhaps unsurprisingly, show very good and consistent performance when tested on a dataset that is contained within the support of its training data, where testing points can be readily accessed by interpolation. This numerical experiment clearly demonstrates that transferability of a given model is critically influenced by the choice of data that is used to trained the model and that very large and diverse datasets ensure both high accuracy and high transferability.

The presented results also suggests that the fact that the distribution of energies of the EM set decays very quickly as it approaches the BCC ground state (cf., Fig. 2), did not affect the performance of the models when testing on configurations from the DE set, whose distribution is strongly peaked at low energies, see also Supplemental Note 9.

In order to determine whether the large errors observed when DE-trained models are validated on EM data are attributed to high-energy configurations, that could be argued to be irrelevant in most conditions of practical interest, different partitions of the DE set into training and validation data were also investigated. Therefore, instead of partitioning using a random split of the data, the training/validation partitions were instead guided by the manual labeling of the DE set into distinct configuration groups. In this case, *entire groups* were either assigned to the training *or* to the validation set in a random fashion, so that configurations from one group can only be found in either the training or the validation set, but not in both. This way, we limit (but not rigorously exclude) the possibility that very similar configurations are found in both the training and validation sets. The validation errors measured with this new scheme dramatically increase by one to two orders of magnitude for quadratic-SNAP potentials, as compared to the random hold-out approach. Therefore, clearly showing that even well-behaved, expert-selected, configurations can be poorly captured by MLIAPs when no similar configurations are present in the training set. Consequently, further demonstrating that MLIAP should not be used to extrapolate to new classes of configurations, even if these configurations are not dramatically different from those found in the training set (e.g., different classes of defects in the same crystalline environment). A more detailed analysis can be found in Supplemental Note 8. Therefore, the results shown in this work instead suggest that in order to ensure robust transferability one requires the generation of very large and diverse training sets that fully encompass the physically-relevant region for applications, so that extrapolation is never, or at least very rarely, required.

Finally, we demonstrate that the resultant NN models are numerically and physical stable by testing them in production MD simulations using the LAMMPS code[44,45]. Figure 5 shows the deviation in energy as a function of MD timestep δt for the Type A NN models trained using the EM data set. In all cases, the degree of energy conservation is comparable to that of the SNAP linear model and exhibits asymptotic second order accuracy in the
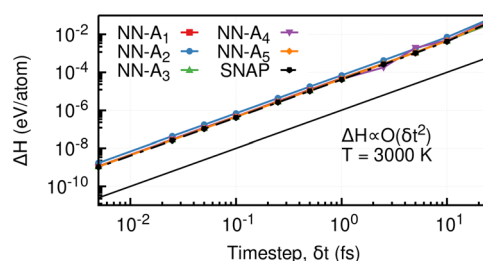

Fig. 5 Energy conservation as a function of MD timestep size for the Type A NN models and the linear SNAP model with EM training. In all cases, a tungsten BCC supercell was simulated under NVE dynamics at 3000 K for 7.5 ps. The energy deviation was calculated by Eq. (2). All the models exhibit asymptotic second order accuracy in δt, characteristic of the Störmer-Verlet time discretization. Higher order deviations emerge only at $\delta t > 10$ fs, close to the stability limit. This demonstrates that the NN models yield energy and force predictions that are consistent, smooth, and bounded.

timestep, as expected for the Störmer-Verlet time discretization used by LAMMPS. Higher order deviations emerge only at $\delta t > 10$ fs, close to the stability limit determined by the curvature of the underlying potential energy surface of tungsten. All of the models were integrated into the LAMMPS software suite using the recently developed *ML-IAP* package, completing the link between *Model Form* and *Simulation Engine* in Fig. 1. The *ML-IAP* package enables the integration of arbitrary neural network potentials into the atomistic software suite of LAMMPS, regardless of the training method used. As a result, all the models developed in this work can be used to perform simulations with the accuracy of quantum methods (i.e., direct transcription of the energy surface defined by suitably diverse DFT training data). To conclude, coupling this code package with the universal training set generation outlined in this work enables a seamless integration of ML models into LAMMPS and thus will enable a breadth of research hitherto unmatched.

CONCLUSION

The present work demonstrates a needed change in the characterization of ML models that is motivated by the goal of transferable interatomic potentials thereby avoiding known pitfalls of extrapolation. Counter-intuitive results presented here showed that model accuracy saturates even when the model flexibility increases and thereby re-directs attention to what is included as training data when assessing the overall quality of an interatomic potential. Machine learned interatomic potentials differ from traditional empirical potentials (simple functional forms derived from physics/chemistry of bonding) in this assessment of the trained space wherein the accuracy of an empirical potential is quantified on the ability to capture domain expertise selected materials properties. Transferring these practices to the training of MLIAP demonstrated that a physically motivated, user expertise, approach for defining training configurations fails to yield MLIAP capable of having desired transferability. Nevertheless, this work also demonstrated that transferability can be achieved by producing a training set that maximizes the volume of descriptor space such that the model rarely extrapolates, where high errors are expected, even when this results in high-energy, far from equilibrium states of the material. Initial efforts to generate MLIAP from hybridized training sets is promising as hybrid training sets produce lower DE validation errors than either DE or EM trained models alone, see Supplemental Note 11. In addition, and as a point that is important for the community of molecular dynamics users, the entropy optimized training sets used to generate the transferable MLIAPs are descriptor agnostic, material independent, and automated with little to no user input tuning. Ongoing work aims to incorporate uncertainty of individual DFT

training calculations (i.e., variable convergence criterion) to further address the sources of model saturation and to establish optimized hybrid training sets. Also, novel software advances in LAMMPS now allow for any ML model form to be used as an interatomic potential in a MD simulation. This is an important scientific advancement because it allows for subsequent research that utilizes these highly accurate and transferable ML models to be used within a code package that is actively utilized by tens of thousands of researchers.

Beyond the use case of IAP, the protocol presented in this work for training set generation is something that many data-sparse ML applications can take advantage of when characterizing the accuracy of a generated model. Therefore, it should be expected that ML practitioners report regression errors, but also now to quantify the complexity of the training data in order for end users to understand where to expect interpolation versus extrapolations in a more quantitative fashion.

METHODS

Maximization of the descriptor entropy

The data-driven EM set was generated using an entropy-optimization approach introduced in previous work[36]. This framework aims at generating a training set that is i) diverse, so as to cover a space of configurations that encompasses most configurations that could potentially be observed in actual MD simulation, and ii) non-redundant, in order to avoid spending computational resources characterizing many instances of the same local atomic environments. To do so, we introduce the so-called *descriptor entropy* as an objective function that can be systematically optimized. In what follows, the local environment around each atom i is described by a vector of descriptors $\boldsymbol{D}_i$ of length m. These descriptors can be arbitrary differentiable functions of the atomic positions around the target atom. In this work, the $\boldsymbol{D}_i$ are taken to be the bispectrum components which were introduced in the development of the GAP potentials[26,30,39], and then adopted in the SNAP approach[28,37]. To avoid excessive roughness on the entropy surface in high dimension, we used the five lowest-order bispectrum components ($m = 5$) in the optimization procedure. As reported above, we nonetheless observe very significant broadening of the descriptor distribution compared to the DE set in almost all directions of the 55-dimensional space induced by the lowest-order bispectrum components. This behavior, which will be studied in detail in an upcoming publication, was also observed in the original publication[36].

The diversity of local environments within a given configuration of the system can be quantified by the entropy of the $m-$dimensional descriptor distribution $S(\{\boldsymbol{D}\})$. High entropy reflects a high diversity of atomic environments within a given configuration, while low entropy corresponds to high similarity between atomic environments. The descriptor entropy is therefore an ideal objective function in order to create a diverse dataset. The creation of high entropy structures can be equivalently recast as the sampling of low-energy configurations on an effective potential energy surface given by minus the descriptor entropy. This enables the training set generation procedure to be implemented in the same molecular dynamics code that will be used to carry out the simulations.

The effective potential is of the form:

$$V_{\text{entropy}} = V_{\text{repulsive}} - KS(\{\boldsymbol{D}\}) \tag{1}$$

where $V_{\text{repulsive}}$ is a simple pairwise repulsive potential that mimics a hard-core exclusion volume, thereby prohibiting close approach between atoms, and $S(\{\boldsymbol{D}\})$ is a nonparametric estimator of the descriptor entropy based on first-neighbor distances in descriptor space[46]. The addition of $V_{\text{repulsive}}$ is essential to avoid generating nonphysical configurations that would prevent convergence of the DFT calculations. As a result, this effective potential can be used to carry out either molecular-dynamics-based annealing or direct minimization, as discussed in the original publication[36].

A possible limitation of this approach is that regions of descriptor space corresponding to crystalline configurations, which are key to many materials science applications, might be under-sampled as entropy maximization promotes configurations where each atomic environment differs from others. In order to avoid this possibility, the dataset also contains "entropy-minimized" configurations, which are obtained using the same procedure as the entropy-maximized ones, except that that sign of K in Eq. (1) is reversed, leading to configurations where order is
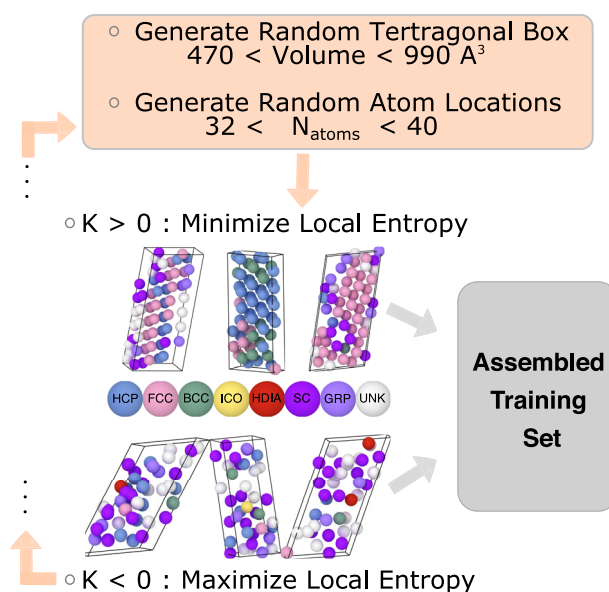

Fig. 6 Automated workflow that generates both entropy maximized and minimized atomic structures given Eq. (1). Representative structures of either type are colored by local crystal structure. Entropy minimized structures are mostly crystalline but with defects, while entropy maximized structures are largely amorphous and/or show non-closed packed structure types.

promoted instead of suppressed. It is important to acknowledge that the type of local order (e.g., FCC or BCC) that is promoted through entropy minimization is not pre-specified by this approach, the data generation procedure is captured in Fig. 6.

This loop was repeated $N/2$ times, generating a total of $N = 223,660$ configurations, half of which are entropy-minimized, and half entropy-maximized. The range of atom counts and box volumes is chosen with respect to ambient density of Tungsten ($15.9°\text{A}^3/atom$). Note that except for very small systems, this sampling procedure typically does not converge to the global minimum of V_{entropy}, but instead remains trapped in local minima of the effective potential energy surface. Different initial random starting points (w.r.t. to initial atomic positions and cell sizes and shapes) will converge to different final states. As shown in Fig. 2, this allows for the generation of a wide range of different configurations, instead of repeatedly generating structures that are internally diverse but very similar to each other. The difficulty of converging to the global entropy optimum is therefore a positive feature in this case. The same effect also occurs in the case of entropy minimization, where this procedure yields crystals that contain stacking defects, grain boundaries, line and point defects, etc., with perfect crystal being only rarely generated.

The energy and forces acting on the atoms were then obtained with the VASP DFT code using the GGA exchange correlation functional with an energy cutoff was set to 600 eV, and a $2 \times 2 \times 2$ Monkhorst-Pack k-point grid used[47–50]. The calculations were converged to an SCF energy threshold of 10^{-8} eV. Imposing a constant plane wave energy cutoff and k-point spacing for all of the diverse configurations is certainly an approximation, and was done to automate the generation of these training labels. Since the entropy maximization method does not need these DFT results to generate new structures, stricter DFT settings can be applied a posteriori.

The main advantage of this method in contrast with conventional approaches that rely on domain knowledge is that it is fully automated and executed at scale because no human intervention is required. In other words, the generated training set was not curated a posteriori to manually prune or add configurations, and the weight of the different configurations in the regression was not adjusted. The training configurations are further material-independent, except for the choice of the exclusion radius of the repulsive potential and the range of densities that was explored. As a result, this means that in the case of pure materials, configurations generated using generic parameter values can simply be rescaled based on the known ground-state density of the target material. Therefore, the method naturally lends itself to high throughput data generation, as every

training configuration can be generated and characterized with DFT in parallel, up to some computational resource limit.

Note that entropy optimization differs in philosophy from some recently proposed active learning approaches[23,24,51] where training sets are iteratively enriched by using a previous generation of the MLIAP to generate new candidate configuration using MD simulations. Candidates are added to the training set whenever a measure of the uncertainty of the prediction reaches a threshold value. These methods are appealing as they also allow for the automation of the training set curation process, and because they generate configurations that can be argued to be thermodynamically relevant. Nevertheless, the diversity of the training sets generated active learning approaches ultimately relies on the efficiency of MD as a sampler of diverse configurations. However, many potential energy surfaces are extremely rough, which can make them very difficult to systematically explore using methods based on naturally evolving MD trajectories. Take for example a configuration of atoms that has a high energy barrier to a new state of interest. Unless a high temperature is set in MD, which skews the thermodynamically relevant states, these rare events will dictate the rate new training is added. Consequently, in active learning methods the selection of the initial configuration from which MD simulations are launched then becomes very important. In contrast, entropy optimization explicitly biases the dynamics so as to cover as much of the feature space as possible using an artificial effective energy landscape that maximizes the amount of diversity contained in each configuration. Note however that both approaches can easily be combined by substituting entropy-optimization for MD as the sampler used within an iterative loop. Finally, notice that some active learning approaches, specifically those based on the d-optimality criterion of Shapeev and collaborators[23], also build on the insight that extrapolation should be avoided in order to identify candidate configurations that should be added to the training set.

Neural Network Models

Neural networks are highly flexible models capable of accurately estimating the underlying function that connects a set of inputs to its corresponding output values from available observations[20,52]. Feed Forward Neural Networks (FFNN) are quite versatile models that can be tailored to predict the results for a wide variety of applications and fields[53–55]. FFNNs have been successfully used to develop IAPs[14,15,56,57]. In this work, we train different FFNNs to learn the mapping between the local atomic environment (characterized using the bispectrum components[26,28,30,39]) and the resultant energy of each atom (denoted as E_i). The total energy of the system is subsequently obtained by summing the atomic energies (i.e., $E_{tot} = \sum_i E_i$). Note that the atomic energies are not available from DFT, so training only considers total energies of entire configurations. The different neural networks are trained by minimizing the squared error that quantifies the discrepancy between the values predicted by the model (i.e., the FFNN) and a set of "ground-truth" output

values obtained for the training set. In this work we trained two different sets of FFNNs, the EM and DE sets described above, using the same training protocol for both.

The training protocol starts by selecting a subset of the training set on which to train the model. In this work we used 70% (randomly sampled) of the configurations to train the model. In addition, we used 10% (again randomly sampled) as a validation set and the remaining 20% as the test set. The validation set and the test set are not used directly for the training on the model. These sets are used to monitor for over-fitting and to critically assess the ability of the trained model to generalize to new/unseen data. Partitioning a given data set into training, validation, and testing is a common strategy in deep learning model development.

This work considered ten different neural networks architectures. Five of the architectures systematically reduced the number of features before predicting the energy in the last layer. These neural networks are denoted as Type A (step-down) and their architecture is illustrated in Fig. 7A. The activation function used between each layer is the SoftPlus activation function and is applied to all the layers (after transforming the inputs adequately) except the final output layer because it is the one that predicts the energy associated to the input bispectrum components. The other five neural networks initially increased the number of features and subsequently decreased the number of features before predicting the energy in the last layer and are denoted Type B(expand-then-contract). Similar to Type A networks, the activation function used is the SoftPlus function. Figure 7B illustrates the architectures of the Type B neural networks. Supplemental Note 2 details the procedure leveraged for selecting the optimal learning rate and batch size and Supplemental Note 3 details the procedure used for selecting the optimal featurization of the bispectrum components for training deep-learning potentials.

Each one of the ten different neural networks was trained for 800 epochs using the optimal values for the learning rate and batch-size previously identified. Furthermore, we used the ADAM optimizer[58] and incorporated a learning rate scheduler that reduced the learning rate by half if the validation loss did not change by $1 \cdot 10^{-4}$ over 50 epochs. After each network was trained we assessed its accuracy by comparing its energy prediction to the ground-truth (obtained with ab-initio calculations) using the Root Squared Error (RSE).

Simulation stability

We tested the numerical and physical stability of all of the models by running realistic molecular dynamics simulations in the LAMMPS atomistic simulation code[44,45]. The most important characteristic of any classical potential is the degree to which it conserves energy when used to model Hamiltonian or NVE dynamics. Theoretically, under these conditions, the total energy or Hamiltonian $H = T + V$ is a constant of the motion, while the kinetic energy T and potential energy V fluctuate equally and oppositely. In practice, the extent to which the total energy is conserved is strongly affected by the timestep size δt, as well as the time discretization
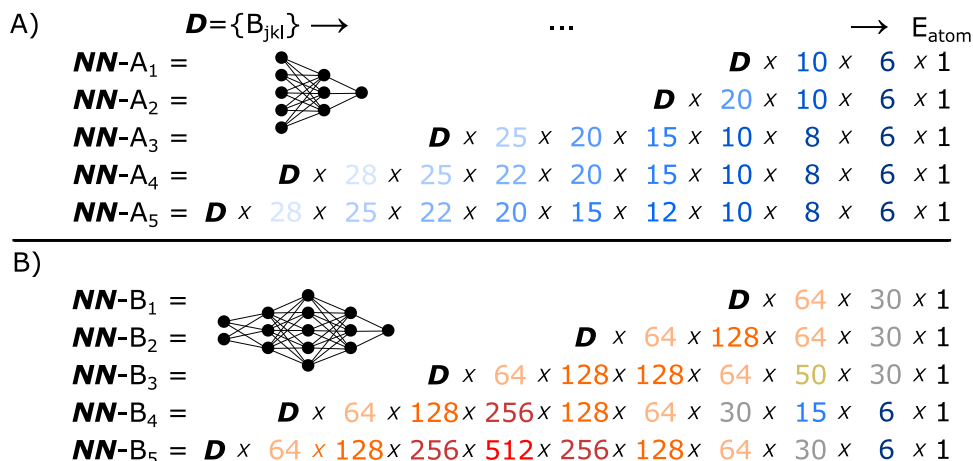

Fig. 7 Visual representation of the two neural network architecture types. A type A step-down and **B** type B expand-and-contract. The input is a vector of **D** descriptors of the atomic environment of one atom. These are passed through each layer in the network to yield the atomic energy E_i of the atom. Integer factors indicate the number of nodes in each layer. The total number of nodes defines the number of degree of freedom N_{DoF} for each model. Recall from Fig. 3 that **D** is 14, 30, or 55. Notice that **D** only affects the nodes in the input layer since the nodes on the other layers remain unchanged.

npj

scheme, and any pathologies in the potential energy surface and corresponding forces. Because LAMMPS uses Störmer-Verlet time discretization that is both time reversible and symplectic[59], a well-behaved potential should exhibit no energy drift and the small random variations in energy that do occur should have a mean amplitude that is second order in the timestep size. We characterize both of these effects by simulating a fixed-length trajectory with a range of different timestep sizes and sample the change in total energy relative to the initial state. The average energy deviaton is defined to be

$$\Delta H(\delta t) = \frac{1}{nN} \sum_{i=1}^{n} |H(t_i; \delta t) - H(t_0)|, \qquad (2)$$

where $H(t_i; \delta t)$ is the total energy sampled at time t_i from a trajectory with timestep δt, n is the total number of samples, and N is the number of atoms.

All the NVE simulations were initialized with 16 tungsten atoms in a BCC lattice with periodic boundary conditions, equilibrated at a temperature of 3000 K, close to the melting point. Each simulation was run for a total simulation time of 7.5 ps and the number of time samples n was 1000. All calculations were performed using the publicly released version of LAMMPS from November 2021. In addition to the base code, LAMMPS was compiled with the *ML-IAP*, *PYTHON*, and *ML-SNAP* packages. The *MLIAP_ENABLE_PYTHON* and *BUILD_SHARED_LIBS* compile flags were set. An example LAMMPS input script has been included in the Supplemental Information.

The dependence of energy deviation on timestep size for some representative models is shown in Fig. 5. In all cases, we observe asymptotic second order accuracy, as expected for the Störmer–Verlet time discretization used by LAMMPS. Higher order deviations emerge only at $\delta t > 10$ fs, close to the stability limit determined by the curvature of the underlying potential energy surface of tungsten.

Bispectrum components and SNAP potentials

The entropy maximization effective potential, the neural network potentials, and the SNAP potentials described in this paper all use the bispectrum components as descriptors of the local environment of each atom. These were originally proposed by Bartok et al.[26,30,39] and then adopted in the SNAP approach[28,37].

In the linear SNAP potential, the atomic energy of an atom i is expressed as a sum of the bispectrum components $\mathbf{B}_i$ for that atom, while for quadratic SNAP, the pairwise products of these descriptors are also included, weighted by regression coefficients

$$E_i(\mathbf{r}^N) = \boldsymbol{\beta} \cdot \mathbf{B}_i + \frac{1}{2} \mathbf{B}_i \cdot \boldsymbol{\alpha} \cdot \mathbf{B}_i, \qquad (3)$$

where the symmetric matrix $\boldsymbol{\alpha}$ and the vector $\boldsymbol{\beta}$ are constant linear coefficients whose values are determined in training. The bispectrum components are real, rotationally invariant triple-products of four-dimensional hyperspherical harmonics $\mathbf{U}_j$[39]

$$B_{j_1 j_2 j} = \mathbf{U}_{j_1} \otimes_{j_1, j_2}^{j} \mathbf{U}_{j_2} : \mathbf{U}_j^*, \qquad (4)$$

where symbol $\otimes_{j_1, j_2}^{j}$ indicates a Clebsch-Gordan product of two matrices of arbitrary rank, while : corresponds to an element-wise scalar product of two matrices of equal rank. For structures containing atoms of a single chemical element, the $\mathbf{U}_j$ are defined to be

$$\mathbf{U}_j = \sum_{r_{ik} < R} f_c(r_{ik}) \mathbf{u}_j(\mathbf{r}_{ik}), \qquad (5)$$

where the summation is over all neighbor atoms k within the cutoff distance R. The radial cutoff function $f_c(r)$ ensures that atomic contributions go smoothly to zero as r approaches R from below. The hyperspherical harmonics $\mathbf{u}_j$ are also known as Wigner U-matrices, each of rank $2j + 1$, and the index j can take half-integer values $\{0, \frac{1}{2}, 1, \frac{3}{2}, \dots\}$. They form a complete orthogonal basis for functions defined on S_3, the unit sphere in four dimensions[30,60]. The relative position of each neighbor atom $\mathbf{r}_{ik} = (x, y, z)$ is mapped to a point on S_3 defined by the three polar angles ψ, θ, and φ according to the transformation $\psi = \pi r/r_0$, $\cos \theta = z/r$, and $\tan \varphi = x/y$. The bispectrum components defined in this way have been shown to form a particular subset of third rank invariants arising from the atomic cluster expansion[61]. The vector of descriptors $\mathbf{B}_i$ for atom i introduced in Eq. (3) is a flattened list of elements $B_{j_1 j_2 j}$ restricted to $0 \leq j_2 \leq j_1 \leq j \leq J$, so that the number of unique bispectrum components scales as $\mathcal{O}(J^3)$. In the current work, J values of 1, 2, 3, and 4, are used, yielding descriptor vectors $\mathbf{B}_i$ of length $D = 5, 14, 30$, and 55, respectively.

The radial cutoff value used for entropy maximization, neural network and SNAP models was $R = 4.73$ Å.

CODE AVAILABILITY
The codes used to calculate the results of this study are available from https://github.com/FitSNAP/FitSNAP.

REFERENCES

1. Lounkine, E. et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361–367 (2012).
2. Ietswaart, R. et al. Machine learning guided association of adverse drug reactions with in vitro target-based pharmacology. *EBioMedicine* **57**, 102837 (2020).
3. Chua, H. E., Bhowmick, S. S. & Tucker-Kellogg, L. Synergistic target combination prediction from curated signaling networks: machine learning meets systems biology and pharmacology. *Methods* **129**, 60–80 (2017).
4. Panchal, J. H., Kalidindi, S. R. & McDowell, D. L. Key computational modeling issues in integrated computational materials engineering. *Comput. Aided Des.* **45**, 4–25 (2013).
5. Ramakrishna, S. et al. Materials informatics. *J. Intell. Manuf.* **30**, 2307–2326 (2019).
6. Davies, A., Serjeant, S. & Bromley, J. M. Using convolutional neural networks to identify gravitational lenses in astronomical images. *Mon. Not. R. Astron. Soc.* **487**, 5263–5271 (2019).
7. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl Acad. Sci. U.S.A.* **113**, 3932–3937 (2016).
8. Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R. & Kutz, J. N. Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique* **347**, 845–855 (2019).
9. Patel, R. G., Trask, N. A., Wood, M. A. & Cyr, E. C. A physics-informed operator regression framework for extracting data-driven continuum models. *Comput. Methods Appl. Mech. Eng.* **373**, 113500 (2021).
10. Fort, S., Hu, H. & Lakshminarayanan, B. Deep ensembles: a loss landscape perspective. *arXiv preprint arXiv:1912.02757* (2019).
11. Plimpton, S. J. & Thompson, A. P. Computational aspects of many-body potentials. *MRS Bull.* **37**, 513–521 (2012).
12. Becker, C. A., Tavazza, F., Trautt, Z. T. & de Macedo, R. A. B. Considerations for choosing and using force fields and interatomic potentials in materials science and engineering. *Curr. Opin. Solid. State Mater. Sci.* **17**, 277–283 (2013).
13. Hale, L. M., Trautt, Z. T. & Becker, C. A. Evaluating variability with atomistic simulations: the effect of potential and calculation methodology on the modeling of lattice and elastic constants. *Model. Simul. Mat. Sci. Eng.* **26**, 055003 (2018).
14. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
15. Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **13**, 17930–17955 (2011).
16. Nguyen-Cong, K. et al. Billion atom molecular dynamics simulations of carbon at extreme conditions and experimental time and length scales. In *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–12 (2021).
17. Zepeda-Ruiz, L. A., Stukowski, A., Oppelstrup, T. & Bulatov, V. V. Probing the limits of metal plasticity with molecular dynamics simulations. *Nature* **550**, 492–495 (2017).
18. Germann, T. C. & Kadau, K. Trillion-atom molecular dynamics becomes a reality. *Int. J. Mod. Phys. C* **19**, 1315–1319 (2008).
19. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935 (2017).
20. Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning (MIT Press).
21. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
22. Szlachta, W. J., Bartók, A. P. & Csányi, G. Accuracy and transferability of gaussian approximation potential models for tungsten. *Phys. Rev. B* **90**, 104108 (2014).
23. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171–180 (2017).

24. Smith, J. S. et al. Automated discovery of a robust interatomic potential for aluminum. *Nat. Commun.* **12**, 1–13 (2021).

25. Zuo, Y. et al. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **124**, 731–745 (2020).

26. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).

27. Wood, M. A., Cusentino, M. A., Wirth, B. D. & Thompson, A. P. Data-driven material models for atomistic simulation. *Phys. Rev. B* **99**, 184305 (2019).

28. Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).

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

30. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).

31. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).

32. Patel, R. G. et al. Thermodynamically consistent physics-informed neural networks for hyperbolic systems. *J. Comput. Phys.* **449**, 110754 (2022).

33. Mao, Z., Jagtap, A. D. & Karniadakis, G. E. Physics-informed neural networks for high-speed flows. *Comput. Methods Appl. Mech. Eng.* **360**, 112789 (2020).

34. Bernstein, N., Csányi, G. & Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *Npj Comput. Mater.* **5**, 1–9 (2019).

35. Jia, X. et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019).

36. Karabin, M. & Perez, D. An entropy-maximization approach to automated training set generation for interatomic potentials. *Chem. Phys.* **153**, 094110 (2020).

37. Wood, M. A. & Thompson, A. P. Extending the accuracy of the snap interatomic potential form. *Chem. Phys.* **148**, 241721 (2018).

38. Shapeev, A. V. Moment tensory potentials: a class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153 (2016).

39. Bartók, A. P. The Gaussian Approximation Potential: An Interatomic Potential Derived from First Principles Quantum Mechanics (Springer Science & Business Media, 2010).

40. Jolliffe, I. Principal component analysis. Encyclopedia of Statistics in Behavioral Science (2005).

41. Suh, C., Rajagopalan, A., Li, X. & Rajan, K. The application of principal component analysis to materials science data. *Data Sci. J.* **1**, 19–26 (2002).

42. Rosenbrock, C. W., Homer, E. R., Csányi, G. & Hart, G. L. Discovering the building blocks of atomic systems using machine learning: application to grain boundaries. *Npj Comput. Mater.* **3**, 1–7 (2017).

43. Cusentino, M. A., Wood, M. A. & Thompson, A. P. Explicit multielement extension of the spectral neighbor analysis potential for chemically complex systems. *J. Phys. Chem. A* **124**, 5456–5464 (2020).

44. LAMMPS website and GitHub repository. https://www.lammps.org, https://github.com/lammps/lammps (2021).

45. Thompson, A. P. et al. LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Commun.* **271**, 108171 (2022).

46. Beirlant, J. et al. Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.* **6**, 17–39 (1997).

47. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558 (1993).

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

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

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

51. Vandermause, J. et al. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *Npj Comput. Mater.* **6**, 1–11 (2020).

52. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

53. Feng, S., Zhou, H. & Dong, H. Using deep neural network with small dataset to predict material defects. *Mater. Des.* **162**, 300–310 (2019).

54. Wang, D., He, H. & Liu, D. Intelligent optimal control with critic learning for a nonlinear overhead crane system. *IEEE Trans. Ind. Inform.* **14**, 2932–2940 (2018).

55. Gao, W. & Su, C. Analysis on block chain financial transaction under artificial neural network of deep learning. *J. Comput. Appl. Math.* **380**, 112991 (2020).

56. Sosso, G. C., Miceli, G., Caravati, S., Behler, J. & Bernasconi, M. Neural network interatomic potential for the phase change material gete. *Phys. Rev. B* **85**, 174103 (2012).

57. Tang, L. et al. Development of interatomic potential for al–tb alloys using a deep neural network learning method. *Phys. Chem. Chem. Phys.* **22**, 18467–18479 (2020).

58. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. 3rd International Conference for Learning Representations*, San Diego (2015). http://arxiv.org/abs/1412.6980.

59. Hairer, E. E. Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations (Springer, 2006).

60. Varshalovich, D. A., Moskalev, A. N. & Khersonskii, V. K. *Quantum Theory of Angular Momentum*. (World Scientific, Singapore, 1988).

61. Lysogorskiy, Y. et al. Performant implementation of the atomic cluster expansion (pace) and application to copper and silicon. *Npj Comput. Mater.* **7**, 97 (2021).

AUTHOR CONTRIBUTIONS

D.M.Z., N.L., and M.A.W. implemented fitting code, fit, and evaluated NN and SNAP models. D.P. and M.A.W. generated training data. C.Z.P., N.L., and A.P.T. implemented and evaluated models in LAMMPS. All authors participated in conceiving the research, discussing the results, and writing of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00872-x.

Correspondence and requests for materials should be addressed to David Montes de Oca Zapiain or Danny Perez.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.