Overcoming the chemical complexity bottleneck in on-the-fly machine learned molecular dynamics simulations

Lucas Timmerman[a], Andrew J. Medford[a]
{ltimmerman3, ajm}@gatech.edu

Shashikant Kumar[b], Phanish Suryanarayana[b,c]
shashikant@gatech.edu, phanish.suryanarayana@ce.gatech.edu

[a]School of Chemical & Biomolecular Engineering
[b]School of Civil & Environmental Engineering
[c]School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

April 12, 2024

Abstract

We develop a framework for on-the-fly machine learned force field molecular dynamics simulations based on the multipole featurization scheme that overcomes the bottleneck with the number of chemical elements. Considering bulk systems with up to 6 elements, we demonstrate that the number of density functional theory calls remains approximately independent of the number of chemical elements, in contrast to the increase in the smooth overlap of atomic positions scheme.

Keywords Machine learned force fields; Molecular dynamics, Density functional theory; Bayesian inference; High entropy alloys

A multitude of machine learning (ML) models and algorithms have been developed in the past decade to replace and improve first principles or semi-empirical predictions of energies and forces at the atomic scale[1, 2, 3, 4, 5, 6]. The most accurate of these models rely on graph neural networks or transformer architectures trained on large and carefully curated datasets to learn a map directly from structures to energies or forces[7, 8, 9, 10, 11, 12, 6, 5]. These models are designed to act as general purpose force fields and have been used successfully to predict the properties of a wide range of materials[13, 11, 14, 15, 16]. However, due to the finite nature of the training sets of these models, edge cases often arise during inference, resulting in unpredictable performance or unstable simulations[17]. To address this shortcoming, active learning and fine tuning have been used to dynamically update models[18, 19, 20, 21]. These schemes utilize uncertainty quantification (UQ) methods, such as Bayesian estimates, to determine when a model needs to be updated. This procedure requires interfacing with electronic structure theory software packages to generate additional reference data. UQ methods have also been used to train smaller, data-adaptive models "on-the-fly" during a simulation[3, 21]. Implementations of this approach have been built into popular quantum chemistry packages such as VASP[3] and CASTEP[22], improving the reproducibility of the training procedure and portability of the resulting models. These methods rely on a combination of Bayesian error estimates and heuristics to automatically form training sets, eliminating the burden of manually generating the reference data. In principle, this approach can be used for any chemical system. Unfortunately, the feature-based regression and UQ models used in these schemes cause the cost of both training and inference to increase dramatically with the size of the training set and the number of elements present. Furthermore, the robustness of the on-the-fly training procedure is not well established for complex chemistries.

Due to the inherent trade-offs among the various approaches to machine learning force fields (MLFF), researchers must select from a diverse range of options to identify the model best suited to a given problem. The large, general

purpose pretrained models offer a straightforward option since they do not require training and are often designed to work with common scientific packages such as the Atomic Simulation Environment[23] or even have web-based interfaces[10]. Beyond accessibility, it is also crucial to evaluate the validity of model predictions with a first-principles result. This validation is straightforward in the case of large-scale computational screening, where single-point density functional theory (DFT) calculations can be used to validate a subset of candidate materials based on the results of the ML accelerated screening. However, first-principles validation quickly becomes intractable as the quantity of interest moves from DFT energies to properties based on ensemble averages, such as diffusion coefficients or free energies from enhanced sampling[24, 25, 26]. This is particularly concerning, as atomistic ML models are known to behave unreliably when making out-of-domain predictions or during extended simulations under reaction conditions[17, 27]. As a result, utilizing on-the-fly potentials that quantify uncertainty and dynamically validate and update becomes a favorable strategy for obtaining reliable estimates of properties that require long molecular dynamics (MD) or Monte Carlo simulations.

On-the-fly ML algorithms typically require a close coupling between the ML and DFT codes involved. This can be done by directly integrating the ML with the DFT code, which restricts the complexity of ML algorithms available due to the incompatibility between the lower level languages used to write DFT codes and the more interpretable languages commonly used for cutting edge ML packages[28, 29, 30]. Alternatively, it is possible to couple DFT codes to advanced ML frameworks through socket interfaces or hybrid language extensions, allowing fine-tuning of pre-trained models on-the-fly[18, 31, 32]. However, this strategy has only emerged recently and substantially increases the complexity of software installation and maintenance. The most prominent on-the-fly implementations are based on legacy smooth overlap of atomic positions (SOAP) chemical descriptors[33] and Bayesian linear regression[34]. Existing implementations can be found in the VASP[3], CASTEP[22], and SPARC[35] codes. However, the generality of these models is not well established, and the SOAP framework is known to scale poorly to systems with many unique chemical elements due to the increasing computational cost of descriptor computation and inference, as well as the inclusion of redundant information in the descriptor vector[36, 37, 38].

We aim to reduce the gap between state-of-the-art ML models that work for an arbitrary number of elements but are not always portable or transferable, and the existing SOAP-based on-the-fly force fields that are straightforward to use within DFT codes and work well for simple systems but struggle to scale to systems with many chemical elements. To achieve this goal, we introduce a modified workflow based on the normalized Gaussian multipole (GMP) descriptor[39], which shows improved efficiency without compromising performance. See the SI for details on the normalization factor computation. The GMP scheme differs from SOAP in the dependence of the size of the feature vector on the number of unique elements and the formulation of the design matrices. The size of the feature vector for the SOAP chemical descriptor scales quadratically with the number of unique chemical elements, requiring additional computational resources and sometimes causing poor conditioning of the resulting design matrices. Several studies have addressed this issue using compression schemes[40, 37]. However, the dimension of the resulting feature vector scales linearly with the number of unique elements, and the procedure to calculate the compressed feature vector increases computational cost and complexity. The GMP-based models overcome this scaling issue by implicitly embedding elemental identity through a Gaussian representation of atomic valence densities, leading to a fixed vector size independent of the number of chemical elements in the system [41, 39, 42]. This lack of explicit elemental dependence leads to denser representations of chemical environments, and allows the use of pooled design matrices that combine the chemical information from all element types via the kernel evaluation. To ensure portability and ease of use, we implemented the GMP-based on-the-fly potentials in the SPARC DFT code[43]. The SPARC code has minimal dependencies (MPI[44] and BLAS[45]/LAPACK[46]/MKL) ensuring ease of compilation and uses a real-space formalism that enables mixed boundary conditions and short wall times, allowing rapid generation of on-the-fly MLFF training data for systems of arbitrary chemical complexity.

To test the implementation, we used a series of bulk metals and alloys with up to six elements. The systems include bulk Al, Ag, Au, Ir, Pd , Pt, and Rh; and alloys of Pt, Ag, Au, Ir, Pd, and Rh, each with 32 atoms in the unit cell. Aluminum was included for benchmarking, as it is a common reference system. The remainder of the elements are of interest in heterogeneous catalysis[47, 48, 49, 50, 51]. Single-element systems provide both a benchmark for the performance of the ML algorithms and allow us to compare energies of formation of the complex alloys. We perform ab initio molecular dynamics (AIMD) for 10,000 steps (20 ps) for each system as a ground truth, and we evaluate the models by comparing the total variation distance (TVD)[52] of the pair correlation functions (PCFs). Conceptually, the TVD represents the degree of overlap between the PCFs computed using the MLFF and DFT. Additional details on the TVD calculation are provided in the SI. The use of TVD is inspired by previous work revealing that analysis of PCFs provides the best metric for the stability of machine-learned force fields used for molecular dynamics simulations[27]. We assess the average TVD over the full trajectory, as well as time-resolved TVDs that provide further insight into stability. We also provide a comparison of the free energies of formation for the alloys to demonstrate a potential application of the

GMP MLFF. We compare the common finite displacement (FD) method[53, 54] to MD methods[55, 56, 57] to obtain thermodynamic corrections to electronic energies.

Figure 1 presents the key accuracy metrics for the AIMD and MLFF models, as well as an illustrative PCF plot to contextualize the TVD metric. We computed the PCFs for all pairwise interactions in each alloy, so the number of distributions scales quadratically with the number of elements. The box and whisker plot in Figure 1 a) provides a visualization of the distribution of TVD means broken down by pairwise interactions (for alloys) or by element type (for pure metals). There is a similar increase in the spread of TVD for the GMP and SOAP models associated with an increase in the number of chemical elements, most notably as the number of elements becomes greater than two. We hypothesize that this increasing variance with chemical complexity arises because the total amount of force data per structure remains fixed while the number of unique chemical interactions that must be modeled increases quadratically. It is interesting that the magnitude of this trend is comparable for both the GMP and SOAP models, since the SOAP models explicitly differentiate between unique chemical elements, whereas the GMP models do not. Despite the increase in error for alloys, the median performance for both models appears to stabilize after 4 unique chemical elements. To provide visual context for the TVD metric, we plot the PCF for two outliers of the six-element system in Figure 1 b) which corresponds to the same system for GMP and SOAP. Visual inspection of the PCF shows that the ML models accurately reproduce the locations of both major peaks for the system, indicating that despite the relatively large TVD, neither model fails catastrophically. Finally, Figure 1 c) shows the time-resolved TVD corresponding to the same Ir-Ir outlier for the 6-element alloy. Although oscillations are clearly present, they do not increase with time and generally remain below a TVD of 0.25 for both models. These findings suggest that the models are stable even for the least accurate systems. We further confirmed the stability of the corresponding GMP models by simulating an additional 100K steps (200 ps) for all systems using the MLFFs in inference-only mode and observed that the formation energies for the alloys shifted by no more than 70 meV relative to the DFT MD values (see convergence table in SI). Overall, these results indicate that the MLFFs are similarly accurate and robust for both SOAP- and GMP-based models.
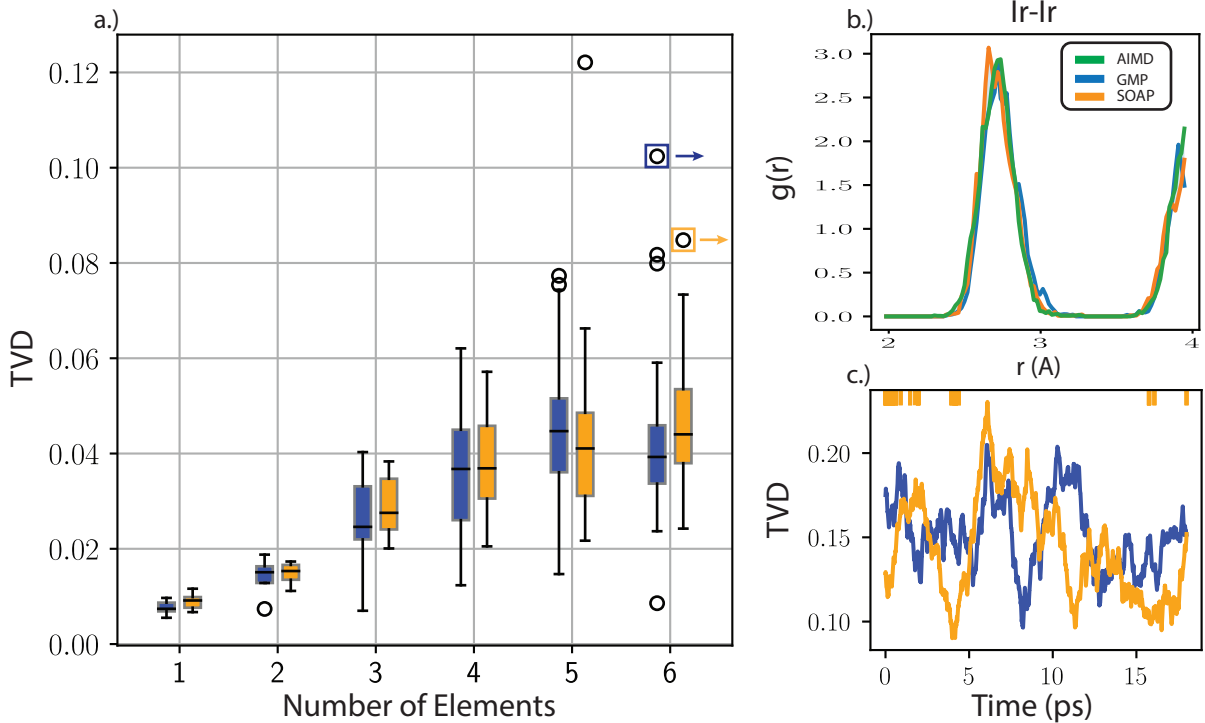

Figure 1: Performance summary for SOAP and GMP models trained on alloys of 1 to 6 elements a.) Mean TVD for all alloys broken down by pariwise interaction for GMP and SOAP in a standard box and whisker plot. The distribution in the case of single-element systems comes from the distribution of all 6 individual single element systems; in all other cases it corresponds to multi-element pair correlation functions. Open circles correspond to outliers b.) Representative partial PCF. Plots correspond to the outliers highlighted for the 6 element alloy which is the Ir-Ir distribution for both SOAP and GMP c.) Time resolved TVD for SOAP and GMP again for the highlighted outliers. Time resolution was included by computing the TVD for a 2 ps interval incremented by 2 fs across the entire trajectory.

Although the accuracy of SOAP and GMP based models are similar, their computational cost differs significantly. In on-the-fly MLFF, the computational cost is dominated by the generation of DFT training data. Figure 2 shows the relationship between the number of chemical elements and the number of DFT calls and the CPU time for both the SOAP and GMP models. Addtional efficiency metrics for both models are tabulated in the SI. For SOAP-based models, the number of DFT calls increases sharply when moving from pure metals to alloys and continues to increase as more unique elements are included. The CPU times follow a similar trend and increase steadily as a function of the number of elements. Both observations are attributed to the increase in the size of the SOAP feature vectors with the number of elements. The amount of data needed to train a reliable model generally increases with the dimension of the feature vector due to the curse of dimensionality[58, 59], and the CPU time increases because the size of the descriptor vector increases, which adds to the compute and memory requirements for both the calculation of the features and the evaluation of the kernel. In contrast, both the number of DFT calls and the CPU time required for GMP-based models are approximately constant regardless of the number of elements present. This shows that GMP-based on-the-fly models can overcome the elemental scaling bottleneck that will cause SOAP-based models to become unwieldy as the number of elements present increases.
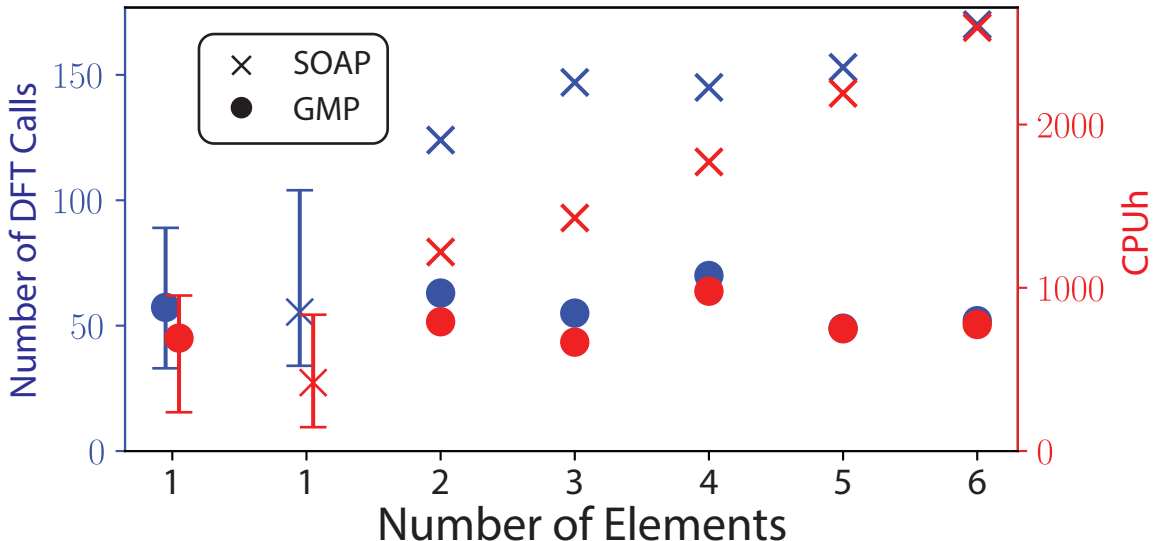

Figure 2: Number of DFT calls (blue, left axis) and cpu times (red, right axis) required to train GMP and SOAP models for all systems. The cpu times include all operations necessary to compute energies and forces using both DFT and the MLFF, as well as time spent training the MLFFs. The error bars for the single element GMP and SOAP systems correspond to the range in number of DFT calls and CPU times among the different pure systems.

To demonstrate the utility of the on-the-fly ML models, we use GMP-based models to compute the Helmholtz free energies of formation for each of the five alloys, including internal, vibrational, and configurational contributions. Here, we compare the results obtained using the FD and MD approaches to obtain the phonon density of states. The FD approach is a well-established method that relies on perturbing atoms corresponding to a primitive cell within a supercell structure to obtain an approximate Hessian of a potential energy surface. This Hessian is used to construct the dynamical matrix, from which phonon modes can be extracted[60]. The phonon density of states can also be extracted directly from an MD simulation using a Fourier transform of the velocity autocorrelation function[57, 56]. We carry out both procedures at the MLFF and DFT levels of theory. The total free energy for each material was computed from the phonon density of states using ASE's `CrystalThermo` package. Additional details on these computations as well as tabulated values can be found in the SI. The corresponding results shown in Fig. 3 show excellent agreement between the DFT and ML formation energies computed using the MD method, while the differences in the ML and DFT FD approximations is substantial ($\sim$0.5 eV) in some cases. However, the FD approximation is plagued by numerical uncertainty associated with the selection of phonon broadening and handling of low-frequency modes. The default options are used for the DFT and ML comparison which includes a broadening of 1E-3 and includes the contributions from imaginary modes by taking the negative square root of the eigenvalue from diagonalizing the Hessian. Alternatively, the black dashed lines in Fig. 3 show the results using the DFT data with slightly different options (1E-4 and and a low frequency cutoff equivalent to the largest imaginary mode). The results illustrate that the difference between the DFT and ML models is lower than the numerical uncertainty of the DFT model, and highlight the advantage of the MD approach where these numerical issues are avoided.

Furthermore, it has been shown that the MD method implicitly captures some degree of anharmonicity, which has a large impact on the vibrational contribution to entropy[61, 62]. This effect is not present in pure metal systems, but it plays a significant role in the vibrational energy of the alloys. Indeed, the discrepancy between the formation energies computed using the FD and MD methods is dominated by the vibrational term, accounting for ∼0.5-1.5 eV of the formation energy contribution. Figure 3 b) shows that the difference between the thermal corrections of the FD and MD methods can be as large as ∼2 eV. Beyond accounting for anharmonicity, the MD method also provides a straightforward and potentially more efficient procedure for treating thermal corrections in HEA. The definition of a primitive cell becomes opaque and perhaps meaningless in the context of a random HEA, making the FD method difficult to apply. Furthermore, the number of DFT calls necessary to use the FD method is a factor of 6 larger than the number of vibrational degrees of freedom for the 32-atom cells used here. We expect that the FD approach will become increasingly more expensive than training a GMP-based MLFF model as cell size or number of elements increases, since we have demonstrated that the number of DFT calls necessary to train a robust model is approximately independent of the chemical complexity.
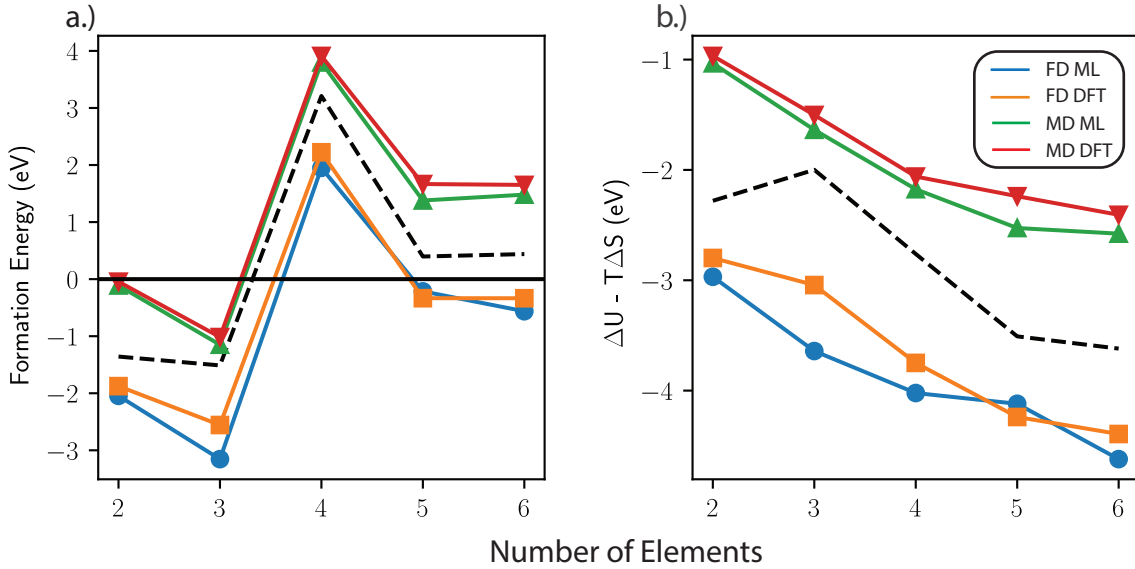

Figure 3: Thermodynamic formation energies for each of the six alloys from the pure bulk components. Each figure contains data for energies computed using phonons extracted directly from MD simulations and the FD method at both the DFT and MLFF levels of theory a) Helmholtz free energies of formation with electronic energy included b.) Thermal corrections including vibrational and configurational entropy as well as internal energy

The results presented here suggest that on-the-fly potentials with GMP features are a promising strategy for complex chemical systems with many elements. However, some challenges remain. Both SOAP- and GMP-based on-the-fly models require many hyperparameters, some related to how the features are constructed and others related to the uncertainty quantification and the active learning loop. Although it is possible to set reasonable defaults for some of these hyperparameters, others require systematic optimization [63, 64, 65] or heuristic tuning. We found that the hyperparameters affecting the ML algorithm shared by GMP and SOAP such as initial training set size and regularization strength dominated the outcomes of on-the-fly runs. We selected the optimal parameters via grid search on the six-element alloy and pure Pt systems. Tuning the descriptor hyperparameters further improved the MLFF performance, although less dramatically. We found that the GMP models improved when we increased the maximum order of the spherical harmonic angular probe and increased the density of radial probes near atom centers. There is a well-established literature regarding the SOAP descriptor hyperparameters, and here we selected parameters conservatively to ensure adequate accuracy. Interestingly, we discovered that using data taken from an AIMD simulation to tune the hyperparameters was not an effective strategy. We attempted to evaluate model hyperparameters by decoupling the DFT data generation from the training procedure. That is, we utilized the same algorithm to determine when the model needed to be updated, but we drew the first principles configuration, energy, and force data from a fixed AIMD trajectory. This allowed us to evaluate a far greater set of hyperparameters without requiring additional DFT calls for each run. However, even when the hyperparameters were optimized in this way, the models trained on-the-fly often resulted in unstable simulations as indicated in Figure 4 a) where the black curve depicts force errors for the above "offline" training procedure and the red curve corresponds to the true on-the-fly run. In both cases, the data points map to instances of model retraining. One strategy that was effective was to focus on the most complex systems, since

hyperparameters that worked for the 6-element system tended to work well for alloys with fewer elements as well. Nevertheless, hyperparameter optimization required substantial computational and human effort, so establishing more systematic approaches to identifying hyperparameters of on-the-fly ML models for complex chemical systems is an important step to make these approaches more accessible and efficient.

A related challenge is the robustness of on-the-fly models without properly selected hyperparameters. Ideally, the strength of an on-the-fly model is its ability to access the underlying DFT method to ensure reliability. Thus, a user might expect that improper selection of hyperparameters causes a model to be inefficient (i.e. call DFT more often than needed), but should not cause it to yield unphysical results. Unfortunately, this is not the case for the current class of on-the-fly models, where improper hyperparameter selection leads to exploration of highly unphysical portions of phase space resulting in uncontrolled errors, as illustrated in Figure 4 a). This failure mode is related to the uncertainty quantification and structure of the active learning loop. Typically, the error estimate from the Bayesian regression are assumed to be correlated with the actual root mean squared error[3, 66, 67, 68, 69], although the error estimates are known to be poorly calibrated. However, we observe that in some cases the error estimates are not even correlated with the true error. Figure 4 b) shows the lack of correlation between the Bayesian error estimate and the actual errors for the points where DFT was performed during an on-the-fly simulations for two different systems with poor correlation (the 5 element system for GMP and Au for SOAP). Pearson's correlation coefficients between the predicted and actual error for all systems can be found in the SI. Despite the poor correlation, these simulations did not result in catastrophic failure, highlighting the lack of direct connection between the quality of UQ estimates and simulation stability. We hypothesize that this disconnect between the quality of the UQ method and on-the-fly performance indicates that the method for sampling phase space is less important than the quality of the training set constructed on-the-fly. This also helps to explain why on-the-fly simulations that visit unphysical portions of phase space are non-recoverable.

The issue of unreliable uncertainty estimates is compounded by the dynamically updated threshold that is often employed to overcome the lack of calibration in the error estimates [70, 3]. Once a large error is observed for a training point, the threshold increases and the model is unable to recover because the DFT calculations are no longer triggered. This can be improved by using hueristics such as periodically forcing DFT calculations[22], but this requires additional hyperparameters and can still lead to catastrophic failure between checks. Avoiding catastrophic failures will require the integration of well-calibrated UQ estimates with more robust active learning loops, although the small number of datapoints and highly correlated nature of MD data make statistically rigorous UQ challenging in the case of on-the-fly MLFF.

Despite these challenges, the results presented here indicate that both SOAP and GMP featurization schemes can be used to construct accurate on-the-fly potentials for systems with up to six elements, and the GMP featurization scheme enables the development of on-the-fly potentials that are more efficient for many-element systems. We demonstrate that the MLFFs generated on-the-fly are robust enough to be used for simulations at length scales not accessible via AIMD, and demonstrate their utility in computing free energies for complex HEAs. The GMP potentials exhibit favorable scaling in terms of the number of DFT calls necessary to train robust potentials, whereas SOAP models require more DFT data as the number of unique chemical elements increases. The models are implemented in the open-source, portable, and highly parallelized real-space DFT code, SPARC[71, 72, 73], so that they can be used by the community and applied to large systems with or without periodic boundary conditions. Future work will focus on improving the efficiency of ML operations, automation of hyperparameter selection, and the development of more reliable uncertainty quantification techniques and active learning strategies. We expect that this platform will enable wider adoption and application of on-the-fly MLFF for a wide range of systems relevant to chemistry, materials science, and chemical engineering.

1 Computational Methods

We previously adapted the on-the-fly ML algorithm developed by Jinnouchi et. al.[3] and inspired by the Gaussian approximation potentials[2] to correct orbital free DFT calculations to Kohn-Sham accuracy in the SPARC electronic structure code [43, 70]. This implementation has been extended to be compatible with the full Kohn-Sham formalism with the option to use the SOAP[35] or GMP descriptors with all on-the-fly functionality implemented in a development branch. The initial training set size and regularization strength for the ML models were systematically optimized using a grid search routine on the 6 element alloy and pure Pt. Descriptor parameters were either taken from the literature or tweaked heuristically to improve accuracy. We used 2x2x2 bulk super cells with 32 atoms for all systems. We utilized the PBE exchange correlation functional with the D3 correction scheme of Grimme[74]. The k-point density and mesh spacing were adjusted until the free energy was converged to at least 10^{-3} Ha/atom. The SPMS ONCV pseudopotentials with non-linear core corrections were used to treat core electrons[75, 76]. The calculations were performed with periodic boundary conditions in all principal directions. All systems were treated as spin-paired since none of the metals considered are magnetic [77]. We validated the accuracy of our AIMD simulations using the blocking
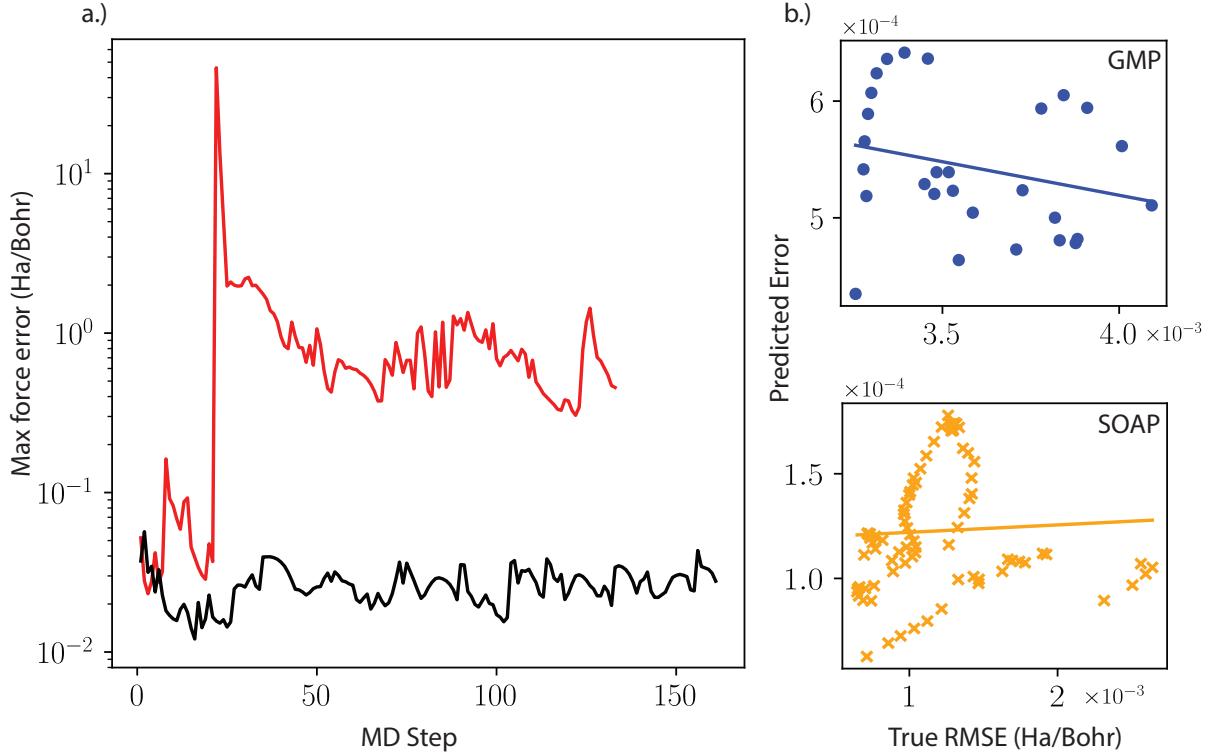
Figure 4: a.) Maximum error of an MD trajectory for the 6 element system system encountered during an on-the-fly run (red) as compared to training from a fixed AIMD trajectory. Each MD Step indicates a point where the model was updated. Identical hyperparameters were used in both cases. b.) Scatter plots showing the lack of correlation between the maximum predicted error and the actual RMSE in the on-the-fly simulation for the 5 element alloy for GMP (top) and for Au for SOAP (bottom). The lines represent linear fits between the true and predicted error.

method to quantify the blocked standard error for the distance associated with the dominant peak in each all-atom PCF[78, 79]. The standard error did not exceed 10^{-2} Bohr for any of the simulations indicating that the simulations were valid. Free energies were computed using the ASE `CrystalThermo` package for both the FD and MD methods. The phonon denisty of states were extracted using the ASE `Phonons` module for the FD method with both DFT and ML calculators for force estimations. The phonon density of states were extracted from MD simulations using the pwtools python package. Additional details are provided in the SI and all input scripts are available via Zenodo.

2 Acknowledgements

References

[1] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):1–4, 2007.

[2] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian Approximation Potentials: the accuracy of quantum mechanics, without the electrons. 10 2009.

[3] Ryosuke Jinnouchi, Ferenc Karsai, and Georg Kresse. On-the-fly machine learning force field generation: Application to melting points. *Physical Review B*, 100:14105, 2019.

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

[5] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, William J Baldwin, Noam Bernstein, Arghya Bhowmik, Samuel M Blau, Vlad Cărare, James P Darby, Sandip De, Flaviano Della Pia, Volker L Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Andrea C Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P Grey, Shuang Han, Will Handley, Hendrik H Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S Jakob, Hyunwook Jung, Venkat Kapil, Aaron D Kaplan, Nima Karimitari, Namu Kroupa, Jolla Kullgren, Matthew C Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J Harry Moore, Aakash A Naik, Samuel P Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A Persson, Karsten Reuter, Andrew S Rosen, Lars L Schaaf, Christoph Schran, Eric Sivonxay, Tamás K Stenczel, Viktor Svahn, Christopher Sutton, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *arXiv e-prints*, page arXiv:2401.00096, 12 2023.

[6] Saro Passaro and C Lawrence Zitnick. Reducing SO(3) Convolutions to SO(2) for Efficient Equivariant GNNs. *arXiv e-prints*, page arXiv:2302.03655, 2 2023.

[7] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials*, 31(9):3564–3572, 5 2019.

[8] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catalysis*, 11(10):6059–6072, 2021.

[9] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

[10] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Félix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H. Sargent, Zachary Ulissi, and C. Lawrence Zitnick. The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 3 2023.

[11] Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. *arXiv e-prints*, page arXiv:2206.07697, 6 2022.

[12] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. *arXiv e-prints*, page arXiv:2306.12059, 6 2023.

[13] Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C. Lawrence Zitnick, and Abhishek Das. GemNet-OC: Developing Graph Neural Networks for Large and Diverse Molecular Simulation Datasets. pages 1–21, 4 2021.

[14] Janice Lan, Aini Palizhati, Muhammed Shuaibi, Brandon M. Wood, Brook Wander, Abhishek Das, Matt Uyttendaele, C. Lawrence Zitnick, and Zachary W. Ulissi. AdsorbML: Accelerating Adsorption Energy Calculations with Machine Learning. (1):1–20, 2022.

[15] Ethan M Sunshine, Muhammed Shuaibi, Zachary W Ulissi, and John R Kitchin. Chemical Properties from Graph Neural Network-Predicted Electron Densities. *The Journal of Physical Chemistry C*, 127(48):23459–23466, 12 2023.

[16] Dávid Péter Kovács, J Harry Moore, Nicholas J Browning, Ilyes Batatia, Joshua T Horton, Venkat Kapil, William C Witt, Ioan-Bogdan Magdău, Daniel J Cole, and Gábor Csányi. MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules. *arXiv e-prints*, page arXiv:2312.15211, 12 2023.

[17] Sina Stocker, Johannes Gasteiger, Florian Becker, Stephan Gunnemann, and Johannes T Margraf. How Robust are Modern Graph Neural Network Potentials in Long and Hot Molecular Dynamics Simulations? *ChemRxiv*, pages 1–7, 2022.

[18] Joseph Musielewicz, Xiaoxiao Wang, Tian Tian, and Zachary Ulissi. FINETUNA: fine-tuning accelerated molecular simulations. *Machine Learning: Science and Technology*, 3(3):03LT01, 9 2022.

[19] Jonathan Vandermause, Boris Kozinsky, and Cameron J Owen. Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of H / Pt. 2022.

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

[21] Yury Lysogorskiy, Anton Bochkarev, Matous Mrovec, and Ralf Drautz. Active learning strategies for atomic cluster expansion models. *PHYSICAL REVIEW MATERIALS*, 7:43801, 2023.

[22] Tamás K Stenczel, Zakariya El-Machachi, Guoda Liepuoniute, Joe D Morrow, Albert P Bartók, Matt I J Probert, Gábor Csányi, and Volker L Deringer. Machine-learned acceleration for molecular dynamics in CASTEP. *The Journal of Chemical Physics*, 159(4):44803, 2023.

[23] Ask Hjorth Larsen, Jens JØrgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, BjØrk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob SchiØtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment - A Python library for working with atoms. *Journal of Physics Condensed Matter*, 29(27), 6 2017.

[24] Noel Jakse and Alain Pasturel. Liquid aluminum: Atomic diffusion and viscosity from ab initio molecular dynamics. *Scientific Reports*, 3:1–8, 2013.

[25] Simon Bailleul, Karen Dedecker, Pieter Cnudde, Louis Vanduyfhuys, Michel Waroquier, and Veronique Van Speybroeck. Ab initio enhanced sampling kinetic study on MTO ethene methylation reaction. *Journal of Catalysis*, 388:38–51, 8 2020.

[26] Giovannimaria Piccini, Mal Soon Lee, Simuck F. Yuk, Difan Zhang, Greg Collinge, Loukas Kollias, Manh Thuong Nguyen, Vassiliki Alexandra Glezakou, and Roger Rousseau. Ab initio molecular dynamics with enhanced sampling in heterogeneous catalysis. *Catalysis Science & Technology*, 12(1):12–37, 1 2022.

[27] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations. pages 1–25, 2022.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, 12 2019.

[29] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A Fresh Approach to Numerical Computing, 11 2014.

[30] William C Witt, Cas van der Oord, Elena Gelžinytė, Teemu Järvinen, Andres Ross, James P Darby, Cheuk Hin Ho, William J Baldwin, Matthias Sachs, James Kermode, Noam Bernstein, Gábor Csányi, and Christoph Ortner. ACEpotentials.jl: A Julia implementation of the atomic cluster expansion. *Journal of Chemical Physics*, 159:164101, 10 2023.

[31] Kevin Tran and Zachary W Ulissi. Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution. *Nature Catalysis*, 1(9):696–703, 2018.

[32] Muhammed Shuaibi, Saurabh Sivakumar, Rui Qi Chen, and Zachary W Ulissi. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. *Machine Learning: Science and Technology*, 2(2):25007, 2021.

[33] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. 9 2012.

[34] Ryosuke Jinnouchi and Ryoji Asahi. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *Journal of Physical Chemistry Letters*, 8(17):4279–4283, 9 2017.

[35] Shashikant Kumar, Xin Jing, John E Pask, and Phanish Suryanarayana. On-the-fly machine learned force fields for the study of warm dense matter: application to diffusion and viscosity of CH, 2 2024.

[36] J. Byggmästar, K. Nordlund, and F. Djurabekova. Simple machine-learned interatomic potentials for complex alloys. *Physical Review Materials*, 6(8):1–11, 2022.

[37] James P. Darby, James R. Kermode, and Gábor Csányi. Compressing local atomic neighbourhood descriptors. *npj Computational Materials*, 8(1):166, 8 2022.

[38] J Byggmästar, K Nordlund, and F Djurabekova. Modeling refractory high-entropy alloys with efficient machine-learned interatomic potentials: Defects and segregation. *Physical Review B*, 104(10):104101, 9 2021.

[39] Xiangyun Lei and Andrew J. Medford. A Universal Framework for Featurization of Atomistic Systems. *The Journal of Physical Chemistry Letters*, 13(34):7911–7919, 9 2021.

9

[40] James P. Darby, Dávid P. Kovács, Ilyes Batatia, Miguel A. Caro, Gus L. W. Hart, Christoph Ortner, and Gábor Csányi. Tensor-reduced atomic density representations. 10 2022.

[41] Mingjie Liu and John R. Kitchin. SingleNN: Modified Behler-Parrinello Neural Network with Shared Weights for Atomistic Simulations with Transferability. *Journal of Physical Chemistry C*, 124(32):17811–17818, 2020.

[42] Muhammed Shuaibi, Yuge Hu, Xiangyun Lei, Benjamin M Comer, Matt Adams, Jacob Paras, Rui Qi Chen, Eric Musa, Joseph Musielewicz, Andrew A Peterson, Andrew J Medford, and Zachary Ulissi. AmpTorch: A Python package for scalable fingerprint-based neural network training on multi-element systems with integrated uncertainty quantification. *The Journal of Open Source Software*, 8(87):5035, 7 2023.

[43] Boqin Zhang, Xin Jing, Qimen Xu, Shashikant Kumar, Abhiraj Sharma, Lucas Erlandson, Sushree Jagriti Sahoo, Edmond Chow, Andrew J. Medford, John E. Pask, and Phanish Suryanarayana. Version 2.0.0 – SPARC: Simulation Package for Ab-initio Real-space Calculations. pages 1–10, 2023.

[44] Message P Forum. MPI: A message-passing interface standard, 1994.

[45] An updated set of basic linear algebra subprograms (BLAS). *ACM Trans. Math. Softw.*, 28(2):135–151, 6 2002.

[46] E Anderson, Z Bai, C Bischof, L S Blackford, J Demmel, J Dongarra, J Du Croz, A Greenbaum, S Hammarling, A McKenney, and D Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, third edition, 1999.

[47] J. Greeley, I. E.L. Stephens, A. S. Bondarenko, T. P. Johansson, H. A. Hansen, T. F. Jaramillo, J. Rossmeisl, I. Chorkendorff, and J. K. Nørskov. Alloys of platinum and early transition metals as oxygen reduction electrocatalysts. *Nature Chemistry*, 1(7):552–556, 2009.

[48] Thomas A.A. Batchelor, Jack K. Pedersen, Simon H. Winther, Ivano E. Castelli, Karsten W. Jacobsen, and Jan Rossmeisl. High-Entropy Alloys as a Discovery Platform for Electrocatalysis. *Joule*, 3(3):834–845, 3 2019.

[49] Jack K Pedersen, Thomas A A Batchelor, Alexander Bagger, and Jan Rossmeisl. High-Entropy Alloys as Catalysts for the CO2 and CO Reduction Reactions. *ACS Catalysis*, 10(3):2169–2176, 2 2020.

[50] Jeff Greeley and Manos Mavrikakis. Alloy catalysts designed from first principles. *Nature Materials*, 3(11):810–815, 2004.

[51] Ifan E L Stephens, Alexander S Bondarenko, Ulrik Grønbjerg, Jan Rossmeisl, and Ib Chorkendorff. Understanding the electrocatalysis of oxygen reduction on platinum and its alloys. *Energy & Environmental Science*, 5(5):6744–6762, 2012.

[52] Alison L Gibbs and Francis Edward Su. On Choosing and Bounding Probability Metrics. *International Statistical Review*, 70(3):419–435, 12 2002.

[53] K Parlinski, Z Q Li, and Y Kawazoe. First-principles determination of the soft mode in cubic ZrO2. *Physical Review Letters*, 78(21):4063–4066, 1997.

[54] G. Kresse, J. Furthmüller, and J. Hafner. Ab initio Force Constant Approach to Phonon Dispersion Relations of Diamond and Graphite. *Europhysics Letters*, 32(9):729, 1995.

[55] Steve Schmerler. elcorto/pwtools:, 11 2021.

[56] Shiang-Tai Lin, Mario Blanco, and William A Goddard III. The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of Lennard-Jones fluids. *The Journal of Chemical Physics*, 119(22):11792–11805, 12 2003.

[57] D V Minakov, P R Levashov, and V B Fokin. Vibrational spectrum and entropy in simulation of melting. *Computational Materials Science*, 127:42–47, 2017.

[58] Jeremy Watt, Reza Borhani, and Aggelos K. Katsaggelos. *Machine learning refined: Foundations, algorithms, and applications*. 2016.

[59] Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[60] Dario Alfè. PHON: A program to calculate phonons using the small displacement method. *Computer Physics Communications*, 180(12):2622–2633, 2009.

[61] Duancheng Ma, Blazej Grabowski, Fritz Körmann, Jörg Neugebauer, and Dierk Raabe. Ab initio thermodynamics of the CoCrFeMnNi high entropy alloy: Importance of entropy contributions beyond the configurational one. *Acta Materialia*, 100:90–97, 2015.

[62] Blazej Grabowski, Yuji Ikeda, Prashanth Srinivasan, Fritz Körmann, Christoph Freysoldt, Andrew Ian Duff, Alexander Shapeev, and Jörg Neugebauer. Ab initio vibrational free energies including anharmonicity for multicomponent alloys. *npj Computational Materials*, 5(1):80, 2019.

[63] Michael A Lones. How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv e-prints*, page arXiv:2108.02497, 8 2021.

[64] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[65] Bernd Bischl, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, Theresa Ullmann, Marc Becker, Anne-Laure Boulesteix, Difan Deng, and Marius Lindauer. Hyperparameter Optimization: Foundations, Algorithms, Best Practices and Open Challenges. *arXiv e-prints*, page arXiv:2107.05847, 7 2021.

[66] Guicheng Zhang, Wenting Liu, Tao Hu, Sansan Shuai, Chaoyue Chen, Songzhe Xu, Wei Ren, Jiang Wang, and Zhongming Ren. On-the-fly machine learning force field study of liquid-Al/α-Al2O3 interface. *Applied Surface Science*, 638:158141, 2023.

[67] Xiangyue Liu, Weiqi Wang, and Jesús Pérez-Ríos. Molecular dynamics-driven global potential energy surfaces: Application to the AlF dimer. *The Journal of Chemical Physics*, 159(14):144103, 10 2023.

[68] Namrata Jaykhedkar, Roman Bystrický, Milan Sýkora, and Tomáš Bučko. How the Temperature and Composition Govern the Structure and Band Gap of Zr-Based Chalcogenide Perovskites: Insights from ML Accelerated AIMD. *Inorganic Chemistry*, 62(31):12480–12492, 8 2023.

[69] Bhubnesh Lama, Alevtina L Smirnova, and Tula R Paudel. Enhanced Li-Ion Diffusivity of LiFePO4 by Ru Doping: Ab Initio and Machine Learning Force Field Results. *ACS Applied Energy Materials*, 6(20):10424–10431, 10 2023.

[70] Shashikant Kumar, Xin Jing, John E Pask, Andrew J Medford, and Phanish Suryanarayana. Kohn–Sham accuracy from orbital-free density functional theory via Δ-machine learning. *The Journal of Chemical Physics*, 159(24):244106, 2023.

[71] Qimen Xu, Abhiraj Sharma, Benjamin Comer, Hua Huang, Edmond Chow, Andrew J. Medford, John E. Pask, and Phanish Suryanarayana. SPARC: Simulation Package for Ab-initio Real-space Calculations. *SoftwareX*, 15:1–17, 2021.

[72] Swarnava Ghosh and Phanish Suryanarayana. SPARC: Accurate and efficient finite-difference formulation and parallel implementation of Density Functional Theory: Isolated clusters. *Computer Physics Communications*, 212:189–204, 2017.

[73] Swarnava Ghosh and Phanish Suryanarayana. SPARC: Accurate and efficient finite-difference formulation and parallel implementation of Density Functional Theory: Extended systems. *Computer Physics Communications*, 216:109–125, 2017.

[74] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *Journal of Chemical Physics*, 132(15), 2010.

[75] Mostafa Faghih Shojaei, John E. Pask, Andrew J. Medford, and Phanish Suryanarayana. Soft and transferable pseudopotentials from multi-objective optimization. *Computer Physics Communications*, 283:108594, 2 2023.

[76] D R Hamann. Optimized norm-conserving Vanderbilt pseudopotentials. *Physical Review B*, 88(8):85117, 8 2013.

[77] Nataliya Lopanitsyna, Guillaume Fraux, Maximilian A. Springer, Sandip De, and Michele Ceriotti. Modeling high-entropy transition-metal alloys with alchemical compression. 045802:1–15, 2022.

[78] H Flyvbjerg and H G Petersen. Error estimates on averages of correlated data. *The Journal of Chemical Physics*, 91(1):461–466, 7 1989.

[79] Alan Grossfield and Daniel M Zuckerman. Chapter 2 Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. In Ralph A Wheeler, editor, *Annual Reports in Computational Chemistry*, volume 5, pages 23–48. Elsevier, 2009.

3 Supporting Information

3.1 Normalized GMP Descriptor

The GMP descriptor, originally published by Lei. et. al., is a universal mapping from Cartesian space to chemical descriptor space that embeds elemental identity via Gaussian approximations to pseudopotential valence electron densities. The original formulation of the descriptor was not suitable for on-the-fly kernel based ML as the units of the descriptors and hence magnitudes varied wildly. To overcome this issue, we apply a simple normalization factor computed at the beginning of every simulation based on the selected radial $g_n(r)$ and angular Y_m probe functions:

$$A_{n,m} = \frac{1}{\sqrt{\iiint g_n^2(r) Y_m^2(\theta, \phi) dr d\theta d\phi}}$$

where $g_n(r)$ represents a Gaussian with a specified standard deviation of n and Y_m represents the collection of Maxwell Cartesian Spherical Harmonics of order m. This scaling factor ensures uniformity in units across different angular channels and forces that magnitude of higher order descriptors to 0 as in traditional multipole expansions. This step allowed us to use traditional kernel methods to evaluate the similarity between chemical environments represented by the GMP descriptors.

3.2 Additional Computational Details

Primitive `fcc` unit cells were constructed for Ag, Al, Au, Ir, Pd, Pt, and Rh with initial guesses for lattice parameters taken from the Materials Project. The lattice constants were determined utilizing the `fmin` function from `sklearn` to minimize the energy as a function of lattice constant. Super cells were created from relaxed primitives by repeating cells. The alloys were formed by randomly replacing atoms in a Pt super cell with ~equimolar amounts of the alloyed elements. The mixing was done element-wise proceeding alphabetically, resulting in five random bulk alloys (PtAg, PtAgAu, PtAgAuIr, PtAgAuIrPd, PtAgAuIrPdRh). Initial guesses for the lattice parameters of the super cells were formed via composition-weighted averages of the lattice constants for the constituent elements. Relaxations on the cell and positions were carried out iteratively until both the maximum pressure in the cell fell below 1E-02 GPa and the maximum force component fell below 6E-04 Ha/Bohr. Relaxed cells were used as starting points for all MD simulations, both AIMD and on-the-fly. We used the isokinetic ensemble with a Gaussian thermostat to maintain the simulation temperature at 500K. A time step of 2 fs was used for integration. The simulations were run until 10,000 integration steps were completed, resulting in a simulation length of 20 ps. This provided an adequate balance between a desire to minimize computational cost and the need to robustly characterize the stability of the simulations. We ran all calculations with a k-point density of 32 Å which corresponds to a k-point grid of (8,8,8) for the primitive fcc cells and (4,4,4) for the supercells. We employed mesh spacing of 0.25 Bohr which is equivalent to a plane-wave cutoff of ~1340 eV. The SCF tolerance was set to achieve at least 1E-4 Ha/Bohr accuracy in atomic forces for all relaxations and MD simulations.

We defined the TVD for integral normalized PCFs as:

$$TVD = \frac{1}{2} \int_0^{R_{max}} |g_{DFT}(r) - g_{ML}(r)| dr$$

where $g(r)$ represents the integral normalized PCF for a given interaction and the integration domain is radial distance. Normalizing the PCFs by their integral area allows us to treat them as probability density functions that describe the probability of finding a given pair of atoms at a specified radial distance from each other. This allows for a straightforward application of the TVD metric.

We define the formation energy of the alloys as the energy of the alloy minus the composition weighted sum of the constituent components:

$$E_f = E_{alloy} - \sum_i^{N_{types}} n_i E_i$$

where E_i corresponds to the per atom energy of species i in its reference state and n_i corresponds to the number of atoms of species i in the alloy. We include electronic, internal, vibrational, and configurational contributions in our calculation. The configurational entropy was computed assuming an ideal solid solution:

$$S_{conf} = -k_B \sum_i^{N_{types}} x_i ln x_i$$

This term is 0 in the case of pure bulk components thereby acting as a strong driving force for the formation of the HEAs. The computation of electronic, internal, and vibrational contributions were computed differently for the FD and MD approaches. For both methods, the electronic contribution was taken as the fully relaxed or ground state energy computed using DFT. The phonon density of the states was calculated using both DFT and trained ML models to assemble and diagonalize the dynamical matrix from a numerical approximation of the PES Hessian using the built-in ASE `Phonons` module. The phonon density of states was extracted directly from the simulation via the Fourier transform of the velocity autocorrelation function using the `pwtools` python package. In both the FD and MD method, the internal and vibrational energies were calculated using the ASE `CrystalThermo` package. Note that the ASE module automatically normalizes the phonon density of states to match the expected vibrational degrees of freedom in the system. The density of states extracted from the MD simulations was manually normalized.

3.3 Convergence of Formation Energy

Table S1 contains the convergence information for the extended ML-only runs. These values represent the deviation of the formation energy of each alloy from the DFT MD value computed after 20 ps.

Formation Energy Convergence	50.543	68.808	4.506	41.976	28.507	meV

Table S1: Difference between reference MD DFT formation energies and MD ML formation energies computed running the simulation in predict only mode for an additional 200 ps.

3.4 Efficiency Data

Table S2 contains a summary of efficiency data for the SOAP and GMP models during on-the-fly training on the alloys. In every case, the efficiency of the on-the-fly procedure favors the GMP models. The GMP models require anywhere from 2-3x fewer DFT calls, 4-5x fewer training atoms, and less overall walltime. The simulation time here is dominated by DFT calls. The amount of walltime spent in ML operations is occasionally larger for GMP, but this inverts as the number of elements in the systems increases due to an increase in the walltime for SOAP models and no change in the GMP walltimes.

Material System	Method	No. of KS Steps Performed	WTime DFT (h)	WTime MLFF (h)	Total WTime (h)	Columns
PtAg	GMP	63	2.31	0.99	3.30	47
	SOAP	124	4.67	0.41	5.08	203
AuPtAg	GMP	55	1.78	1.00	2.78	39
	SOAP	147	5.13	0.82	5.95	209
IrPtAgAu	GMP	70	3.02	1.06	4.08	55
	SOAP	145	6.31	1.07	7.38	227
PdPtAgAuIr	GMP	49	2.11	1.01	3.12	49
	SOAP	153	6.55	2.58	9.13	230
RhPtAgAuIrPd	GMP	52	2.24	0.99	3.23	52
	SOAP	170	8.45	2.36	10.81	276

Table S2: Selected metrics for assessing model efficiency for GMP and SOAP simulations on alloys. The timings are approximate and correspond to wall times. CPU hours may be computed by multiplying the wall times by 240. The "Columns" metric refers to the number of columns in the design matrix. The number of rows corresponds direcly to the number of KS steps.

Table S3 contains the efficiency data for the single element systems. These results differ from the alloy systems in a couple of ways. First, the number of DFT calls is similar during the on-the-fly simulation for GMP and SOAP models. Second, the wall time spent on MLFF operations heavily favors the SOAP based models in all but one case. Interestingly, the number of columns in the GMP single element models is on par with the the number of columns in the alloy models which suggests that the GMP descriptor may be more information dense than SOAP. The SOAP models, on the other hand, contain nearly 2x fewer columns for the single element systems than the alloys. The total wall times for the GMP models are comparable to the alloy systems, whereas the wall times for the SOAP runs are greatly reduced.

13

Material System	Method	No. of KS Steps Performed	WTime DFT (h)	WTime MLFF (h)	Total WTime (h)	Columns
Al	GMP	33	0.25	0.74	0.99	33
	SOAP	52	0.48	0.13	0.61	129
Ag	GMP	60	1.80	0.72	2.52	31
	SOAP	59	1.64	0.09	1.73	81
Au	GMP	89	2.90	1.07	3.97	29
	SOAP	104	3.40	0.08	3.48	81
Ir	GMP	50	1.81	1.35	3.16	29
	SOAP	34	1.07	0.07	1.14	62
Pd	GMP	62	2.46	0.86	3.32	36
	SOAP	53	1.98	0.09	2.07	87
Pt	GMP	64	2.35	1.24	3.59	28
	SOAP	46	1.59	0.07	1.66	64
Rh	GMP	43	1.75	0.88	2.63	34
	SOAP	40	1.45	0.09	1.54	82

Table S3: Selected metrics for assessing model efficiency for GMP and SOAP simulations on single element bulk structures. The metrics are the same as in Table S2

3.5 Tabulated Values of Energies

Here, we include tabulated values of internal energies and entropies (Table S4) and formation energies (Table S5) corresponding to the figures of the main text for reference, as well as tabulated values of the contributions due to configurational entropy (S6.

Number of Elements	2	3	4	5	6	eV
FD ML	-2.968360	-3.641422	-4.021960	-4.119515	-4.621012	
FD DFT	-2.797997	-3.043473	-3.748951	-4.239140	-4.393300	
FD Uncertainty Bound	-2.279862	-1.998956	-2.761698	-3.508418	-3.619274	
MD ML	-1.034114	-1.636513	-2.174891	-2.526585	-2.577619	
MD DFT	-0.966875	-1.501839	-2.060034	-2.238274	-2.406698	

Table S4: Differences in internal energy and entropy at 500 K for the formation of the alloy from its constituent elements. Includes the configurational entropy of the alloy as an ideal solid solution

Number of Elements	2	3	4	5	6	eV
FD ML	-2.045798	-3.154633	1.950198	-0.214359	-0.562661	
FD DFT	-1.875435	-2.556684	2.223207	-0.333984	-0.334949	
FD Uncertainty Bound	-1.952282	-2.580561	2.174884	-0.450343	-0.371092	
MD ML	-0.111552	-1.149723	3.797267	1.378571	1.480732	
MD DFT	-0.044313	-1.015050	3.912124	1.666882	1.651653	

Table S5: Formation Energies. Includes electronic energy as well. Equivalent to a Helmholtz free energy of formation.

Config S	0.95569	1.51337	1.91138	2.21504	2.46514	eV

Table S6: Configurational entropy corresponding to an ideal solid solution for each of the alloys considered here. The numbers were computed by assuming that 1 mol of alloy corresponds to N_{Av} atoms in the ratio equivalent to that represented in our 32 atom supercells.

3.6 Pearson Correlation Coefficients for UQ

In this work, we utilized the maximum diagonal component of the Bayesian covariance matrix as the uncertainty quantification metric. Figure S1 presents the Pearson's correlation coefficient between the maximum Bayesian error estimate for the forces and the actual RMS force errors for all on-the-fly simulations. It is clear that there is great disparity in the quality of the Bayesian estimate as a probe for the actual error in the system. Several features are of special interest. First, there do not appear to be any clear trends for the quality of the estimate. There is variation both within the single-element systems and as new species are introduced via the alloys. Second, the absolute values of the correlations coefficients range from nearly 0 to 1 for SOAP, whereas the range for GMP is -0.5 to 0.7. It is rare that the correlation exceeds 0.6 regardless of the model type used.
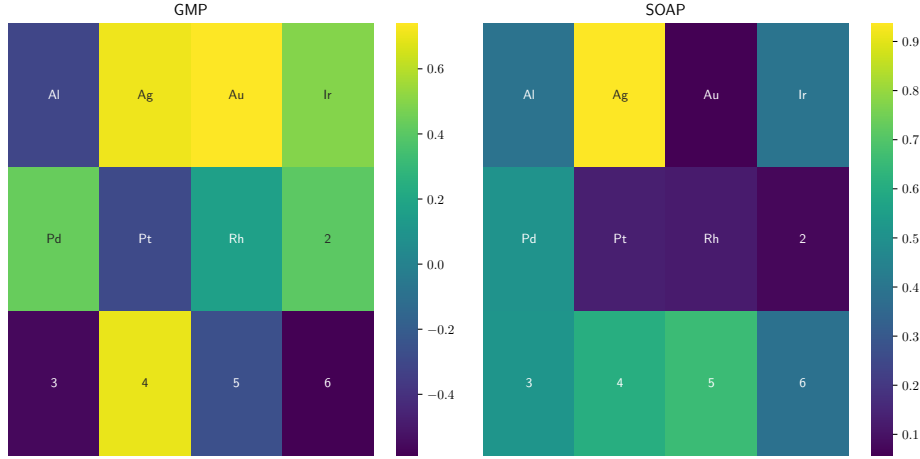

Figure S1: Heatmap of correlation coefficients for the maximum predicted error and the actual RMS error in the system.