Accurate force field for molybdenum by machine learning large materials data

Chi Chen, Zhi Deng, Richard Tran, Hanmei Tang, Iek-Heng Chu, and Shyue Ping Ong

Accurate Force Field for Molybdenum by Machine Learning Large Materials Data

Chi Chen, Zhi Deng, Richard Tran, Hanmei Tang, Iek-Heng Chu, and Shyue Ping Ong[*]

Department of NanoEngineering, University of California San Diego,
9500 Gilman Dr, Mail Code 0448, La Jolla, CA 92093-0448, United States

In this work, we present a highly accurate spectral neighbor analysis potential (SNAP) model for molybdenum (Mo) developed through the rigorous application of machine learning techniques on large materials data sets. Despite Mo's importance as a structural metal, existing force fields for Mo based on the embedded atom and modified embedded atom methods do not provide satisfactory accuracy on many properties. We will show that by fitting to the energies, forces and stress tensors of a large density functional theory (DFT)-computed dataset on a diverse set of Mo structures, a Mo SNAP model can be developed that achieves close to DFT accuracy in the prediction of a broad range of properties, including elastic constants, melting point, phonon spectra, surface energies, grain boundary energies, etc. We will outline a systematic model development process, which includes a rigorous approach to structural selection based on principal component analysis, as well as a differential evolution algorithm for optimizing the hyperparameters in the model fitting so that both the model error and the property prediction error can be simultaneously lowered. We expect that this newly developed Mo SNAP model will find broad applications in large-scale, long-time scale simulations.

I. INTRODUCTION

One of the fundamental challenges in computational materials science is the trade-off between accuracy and scale (length/time). *Ab initio* or first principles methods, such as those based on the density functional theory (DFT) formalism[1,2], have become the method of choice in problems where good chemical accuracy is required. However, due to the high computational cost of solving the Schrödinger equation, most *ab initio* calculations are limited to < 1000 atoms, and *ab initio* molecular dynamics (AIMD) simulations are limited to a time scale of hundreds of picoseconds. Alternatively, simulations of materials can be carried out using empirical force fields[3,4], which assume an explicit functional form for the relationship between atomic configurations and local energies. Force field calculations are usually orders of magnitude faster than *ab initio* calculations, allowing for simulations of systems that contain thousands or even millions of atoms, and over time scales of nanoseconds to microseconds. However, most empirical force fields lack transferability and fail in simulations of complex chemistry with significant variations in chemical bonding and local environments.

In recent years, there have been an upsurge in the application of machine learning (ML) methods to physics problems and material simulations[5–13]. ML models are often trained on DFT calculations, aiming at reproducing material properties with DFT-level accuracy. Notable successes have been demonstrated in predicting atomization energies[14], highest occupied molecular orbital and lowest unoccupied molecular orbital eigenvalues,[15] dielectric constants,[16] energies[17], etc.[18] Similarly, there have been attempts to construct force field models using ML. The models include high-dimensional neural networks[19], gaussian approximation potentials[20], kernel ridge regression[21], and moment tensor potentials[22]. A typical ML model works by first converting structures into numerical values, called features or descriptors, and then the model directs the features into a training procedure along with DFT-calculated quantities, such as energy, force and stress, as the model targets.

A critical component in the ML model development is the choice of feature functions. Ideally, the feature conversion should be invariant to transformations that preserve material properties, for example, permutation of equivalent atoms, rotations and translations. Some examples of proposed feature functions that satisfy or partially satisfy these invariant properties include the Coulomb matrix and its derivatives[23,24], bag-of-bonds[14], symmetry functions[25], bispectrum coefficients[20], among others.[26,27] In principle, ML models can be systematically improved with more training data, if the feature functions are complete and can distinguish unique local environments. Atomic distance-based features such as Weyl matrices[28] and the histogram of pair distance distributions[29] cannot fulfill these requirements because they fail to provide a unique representation of the local environment[26].

Recently, Bartók *et al.*[20] introduced the bispectrum coefficients as a means of mapping the local atomic density function into invariant representations. The bispectrum coefficients have the advantage of providing an almost one-to-one representation of the atomic neighborhood. Subsequently, Thompson et al.[30,31] demonstrated that a spectral neighbor analysis potential, or SNAP, that expresses energies, forces and stress tensors as a linear model of the bispectrum coefficients and their first derivatives can produce quantum-accurate property predictions for Ta and W. One of the key advantages of the SNAP formalism is that the DFT energies, forces and stress tensors can be trained in the same framework. Furthermore, due to its simple formalism, the model is less likely to experience overfitting compared to conventional force fields such as the embedded atom model (EAM) or modified embedded atom model (MEAM), which usually require the optimization of nested nonlinear functions.

In this work, we will present a systematic ML approach to build a SNAP model for Mo. Mo is one of the most important structural metals, valued for its ability to withstand high temperatures, high corrosion resistance, and excellent strength-to-weight ratio. Despite its importance, currently available force fields for Mo based on the EAM[32] and MEAM[33] still do not provide satisfactory accuracy on many properties. This work builds on the excellent work of Bartók *et al.*[20] and Thomson et al.[30,31], but improves on the training procedure in two ways. First, we outline a principal component analysis approach to the selection of training structures, which are obtained from large diverse DFT datasets that we have accumulated via our previous work on Mo grain boundaries[34] and surfaces[35], supplemented with additional data obtained via high-throughput DFT calculations of liquid and solid structures. Second, we propose the use of a differential evolution algorithm to simultaneously optimize the hyperparameters, such as the cutoff radius and weights of the training structures, and the model parameters. We demonstrate that this machine-learned Mo SNAP model can achieve near-DFT accuracy across a wide range of properties, including energies, forces and stress tensors, elastic properties, melting point, surface and grain boundary (GB) energies, outperforming currently available potentials for Mo.

II. BISPECTRUM AND SNAP FORMALISM

The bispectrum and SNAP formalism has been covered extensively in previous works.[20,30] We will only provide a brief summary of the key concepts here, and refer interested readers to those excellent works.

The basic idea of the bispectrum formalism is to map a 3D local atomic neighbor density into a set of coefficients that satisfy the invariant properties. The atomic neighbor density around atom i at location $\mathbf{r}$ is expressed as:

$$\rho_i(\mathbf{r}) = \delta(\mathbf{r}) + \sum_{r_{ii'} < R_c} f_c(r_{ii'}) w_{i'} \delta(\mathbf{r} - \mathbf{r}_{ii'}) \tag{1}$$

where i' denotes a neighbor atom, and $w_{i'}$ is the dimensionless weight to distinguish atom types. The weight is set as 1 in this work as only one element is present. The cutoff function $f_c(r)$ ensures that the neighbor atomic density goes smoothly to zero when the distance $r_{ii'}$ is greater than the cutoff radius R_c.

The angular information in the 3D local density function can be projected onto spherical harmonic functions $Y_m^l(\theta, \phi)$. In the bispectrum approach, the radial component is converted into a third polar angle defined by $\theta_0 = \theta_0^{\max} \frac{r}{R_c}$. Thus the density function can be represented in the 3-sphere (θ, ϕ, θ_0) coordinates instead of (θ, ϕ, r). The density function defined on the 3-sphere can then be expanded using 4D hyperspherical harmonics as follows:

$$\rho(\mathbf{r}) = \sum_{j=0,\frac{1}{2},\ldots}^{\infty} \sum_{m=-j}^{j} \sum_{m'=-j}^{j} u_{m,m'}^j U_{m,m'}^j(\theta, \phi, \theta_0) \tag{2}$$

where the coefficients $u_{m,m'}^j$ are obtained as the inner products between the density function and the basis, given by the following:

$$u_{m,m'}^j = U_{m,m'}^j(0,0,0) + \sum_{r_{ii'} < R_c} f_c(r_{ii'}) w_{i'} U_{m,m'}^j(\theta, \phi, \theta_0) \tag{3}$$

The bispectrum coefficients $B_{j_1,j_2,j}$ can then be obtained via the following equation:

$$B_{j_1,j_2,j} = \sum_{m_1,m_1'=-j_1}^{j_1} \sum_{m_2,m_2'=-j_2}^{j_2} \sum_{m,m'=-j}^{j} \left(u_{m,m'}^j\right)^* H_{j_2 m_2 m_2'}^{j_1 m_1 m_1'} u_{m_1,m_1'}^{j_1} u_{m_2,m_2'}^{j_2} \tag{4}$$

where the constants $H_{j_2 m_2 m_2'}^{j_1 m_1 m_1'}{}^{jmm'}$ are coupling coefficients and $||j_1 - j_2|| \le j \le ||j_1 + j_2||$.

In the SNAP formalism, the energy E_{SNAP}, force $\boldsymbol{F}_{\mathrm{SNAP}}^j$ and stress $\boldsymbol{\sigma}_{\mathrm{SNAP}}^j$ are related to the bispectrum coefficients $\boldsymbol{B}$ by the following

$$E_{\mathrm{SNAP}} = \beta_0 N + \boldsymbol{\beta} \cdot \sum_{i=1}^N \boldsymbol{B}^i \tag{5a}$$

$$\boldsymbol{F}_{\mathrm{SNAP}}^j = -\boldsymbol{\beta} \cdot \sum_{i=1}^N \frac{\partial \boldsymbol{B}^i}{\partial \boldsymbol{r}_j} \tag{5b}$$

$$\boldsymbol{\sigma}_{\mathrm{SNAP}}^j = -\boldsymbol{\beta} \cdot \sum_{j=1}^N \boldsymbol{r}_j \otimes \sum_{i=1}^N \frac{\partial \boldsymbol{B}^i}{\partial \boldsymbol{r}_j} \tag{5c}$$

where β_0 and the vector $\boldsymbol{\beta}$ are the coefficients in the linear models and are fitted from the DFT data to relate E_{SNAP}, $\boldsymbol{F}_{\mathrm{SNAP}}$ and $\boldsymbol{\sigma}_{\mathrm{SNAP}}$, to the structural bispectrum coefficients $\boldsymbol{B}$ and their derivatives $\frac{\partial \boldsymbol{B}}{\partial \boldsymbol{r}}$.

III. POTENTIAL DEVELOPMENT

Figure 1 provides an overview of the potential fitting workflow, which comprises three key steps. First, a set of training structures were generated using structural transformation functions in the Python Materials Genomics (pymatgen) library,[36] as well as AIMD simulations. Second, we propose a structure selection process using principal component analysis (PCA) to identify a reasonable set of structures that provide a good coverage of the feature space for training. Third, the selected training structures were converted into bispectrum coefficients (the feature set), and DFT calculations were performed using these structures. The features and DFT results were fed into the inner
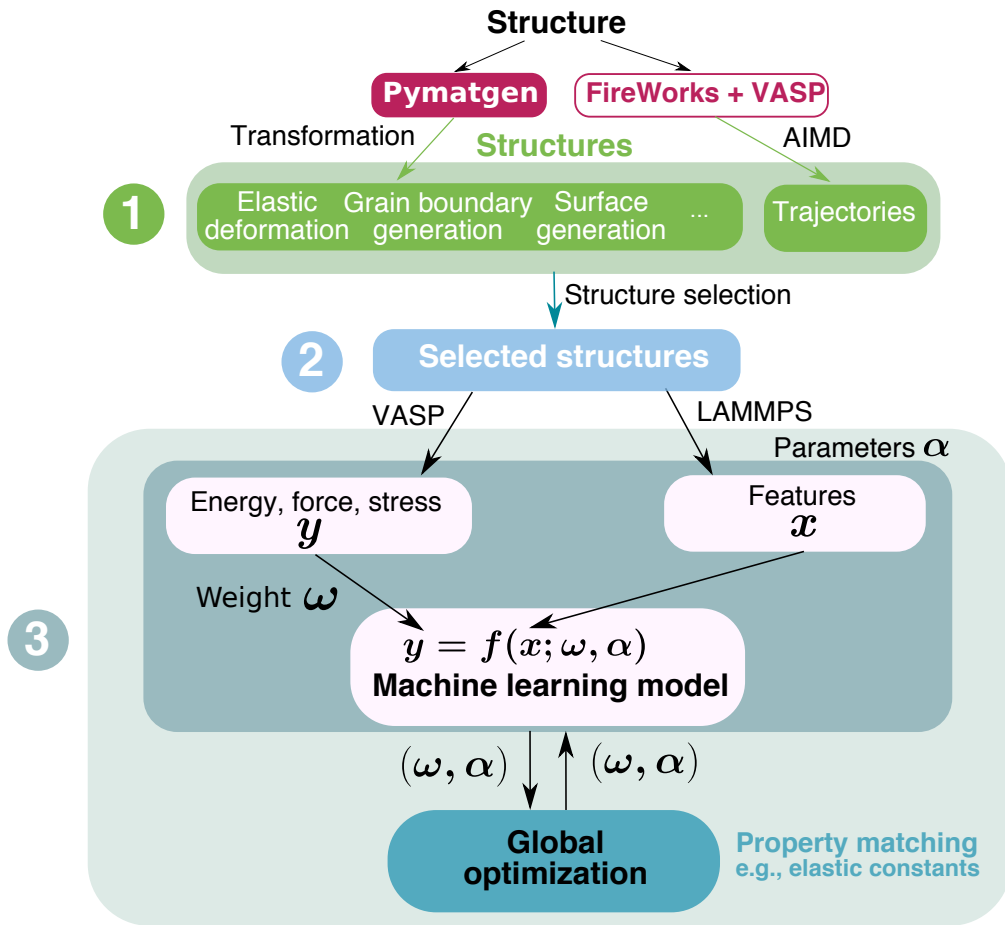
FIG. 1: Model fitting workflow.

loop ML model. On top of the ML model, a differential evolution global optimization algorithm[37] was used to tune the weights from different data groups and the parameters used in feature calculations so that the final model can provide good predictions on material properties as well as basic quantities such as energies, forces, and stress tensors. The overall fitting can be seen as an alternating two-step process. In the inner loop, fitting of the ML model was performed. In the outer loop, the ML model generated in each iteration was then used to compute material properties such as the elastic tensors, and the differences between the predicted and reference values were then used to optimize the hyperparameters. This iterative process was continued until satisfactory accuracy was achieved for both material properties and basic quantities.

A. Training data generation

To develop an effective and robust potential, it is critical that the training data is diverse. Here, we exploit three sets of pre-computed data from our previous work:

1. Ground state structures and energies for Mo from the Materials Project database.[38]

2. Surface slab structures from the Crystalium database,[35,39] which contains the pre-computed surface energies and Wulff shapes of most elements in the Periodic Table. For Mo, the data on all 13 distinct surfaces up to a maximum Miller index of three - (100), (110), (111), (210), (211), (221), (310), (311), (320), (321), (322), (331), and (332) - were included.

3. GB structures from our previous study of the effect of dopants on Mo GBs[34]. Specifically, DFT data from the

relaxation of the (100) $\Sigma5$ twist and (310) $\Sigma5$ tilt, and static calculations of (110) $\Sigma3$ twist, (111) $\Sigma3$ tilt and (110) $\Sigma11$ twist boundaries were included.

We then further augmented this dataset with additional structures incorporating elastic, defect, dynamics and phase transformation information, as follows:

1. Strains of -10% to 10% at 1% intervals were applied to a $3 \times 3 \times 3$ supercell containing 54 Mo atoms in six different modes, as described in the work by De Jong et al.[40]

2. NVT AIMD simulations of a 54-atom supercell were performed at 300 K, 3000 K and 6000 K. 40 snapshots were obtained from each AIMD simulation at intervals of 0.1 ps.

3. An NpT simulation at 6000 K was performed using a 54-atom supercell to obtain the liquid phase of Mo. Nine snapshots were extracted for DFT calculations and further deformations were also carried out on one liquid supercell structure to obtain an additional 40 training structures.

4. AIMD simulations were performed for vacancy-containing structures at 300 K, 3000 K and 6000 K, and 40 snapshots were extracted from each simulation.

DFT calculations on all structures were carried out using the Vienna *Ab initio* Simulation Package (VASP)[41] within the projector augmented wave approach.[42] The Perdew-Burke-Ernzerhof (PBE)[43] generalized gradient approximation (GGA) was adopted for the exchange-correlation functional, and the pseudopotential used was Mo_pv 04Feb2005 with $4p$, $5s$ and $4d$ electrons. The kinetic energy cutoff was set to 520 eV and k-point density was at least 3000 per reciprocal atom. The electronic energy and atomic force components were converged to within 10^{-5} eV and 0.02 eV/Å, respectively. We found that the energy error converged to less than 1 meV/atom using this scheme. For previously relaxed structures, static calculations with the same settings as current work were performed to ensure consistency. As the existing structures in the open databases already provide a good starting point, the additional computational effort associated with performing these static calculations are relatively minimal. The AIMD simulations were performed with a single Γ k-point and were non-spin-polarized. However, the energy, force and stress computations carried out on the snapshots were performed using the same parameters are the rest of the data. All structure manipulations and analysis of DFT computations were performed using pymatgen and automation of calculations were carried out using the FireWorks software.[44] The structures and the corresponding DFT computed data are provided at https://github.com/materialsvirtuallab/snap.

The training structures were converted into bispectrum coefficients (the features) using the implementation in the LAMMPS software[45] by Thompson *et al.*[30] Based on extensive benchmarks carried out over our dataset, we found that an order of three for the bispectrum ($j_{\max} = 3$) is sufficient, in line with previous works.[20,30] We have also kept the angle conversion factor θ_0^{max} at the default value of 0.99363π because we do not expect it to have a significant impact on the model performance. The cutoff radius R_c was further optimized during training (see later section).

B. Data selection

Prior to model training, we performed an exploratory data analysis to examine the distribution of features in the feature space. The aim of this step, which was absent in prior works, is (a) to ensure that we have a good coverage of the feature space of interest, and (b) to minimize the number of structures in (relatively) expensive DFT computations. Data reduction is also particularly important for non-parametric models, e.g. kernel ridge regression, which tends to scale poorly (usually $O(n^3)$ or more) with training data size. It was shown previously that a rational selection of training data improves the ML model performance with fewer data[46]. Figure 2 shows the results of PCA carried out to project the bispectrum features of the entire dataset and their first derivatives onto a two-dimensional plane formed by the first two dominating principal components (PCs). We may observe that the the bispectrum coefficients of atoms in the AIMD structures cover a wide swath of feature space. While the elastic, surface and GB data contribute additional features at the edges of the space, the features from the vacancy data sets lie within the AIMD data group. We surmise that the reason is because the AIMD structures already include a rich variety of local environments from both liquid and solid structures, some of which resemble the vacancy structures. Based on the feature distribution analysis, we have excluded the vacancy dataset from the model training. We will discuss the effect of this exclusion in the Discussion section.
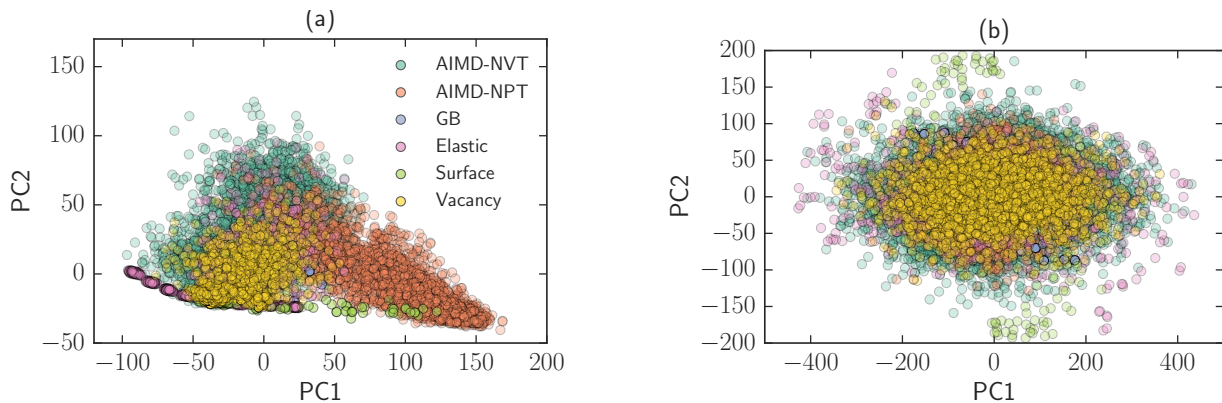
FIG. 2: Two-dimensional projection of the principal components of (a) the atomic bispectrum coefficients and (b) their first derivatives.

C. Model training and optimization

The Mo SNAP model was trained using the bispectrum coefficients as the features, and the DFT energies, forces and stress components as the outputs. Having excluded the vacancy data based on the PCA analysis, the features and DFT data were obtained from the AIMD, elastic, surface and GB structures. In line with the previous work by Thompson *et al.*[30], the DFT energies were normalized by the number of atoms in the corresponding structures, while the model-predicted virial stresses according to Equation 5c, having a unit of energy, were normalized using the structural volume. The inner loop fitting of the model coefficients was done with the ordinary least squares algorithm implemented in the scikit-learn package[47].

The data weights, along with the cutoff radius for R_c for the calculation of features, were treated as hyperparameters for optimization. Whereas Thompson *et al.*[30] relied on the DAKOTA toolkit[48], this work utilizes the differential evolution algorithm[37] implemented in the widely available SciPy[49] package to optimize the hyperparameters, with the target being to minimize the error between the DFT and SNAP predicted elastic constants. The lower and upper bounds for the energy weights and force weights were set as (0.5, 3000) and (0.001, 100), respectively for the AIMD, surface and GB datasets. Virial stresses were not used for these three training data groups. For the elastic data group, the bounds for energy weights and stress weights were (0.05, 10000) and (0.001, 10), respectively, while the forces were not used. The higher upper bounds for the energy and force weights for the elastic data group were chosen to ensure that the predicted energies and forces are more accurate for these groups relative to other groups. Smaller weights were chosen for the stress components, which have much large absolute magnitudes than the forces. The bounds for R_c was set to (4 Å, 5 Å), which is up to the third nearest neighbor distance in pristine bcc Mo. Internal testing shows that both higher and lower R_c lead to higher energy and force errors. We believe this is due to the fact that interactions in metals are relatively short range. For example, conventional MEAM force fields for similar systems typically do not go beyond second nearest neighbors[50], and previous SNAP models on others metals (Ta and W) use similar cutoffs[30,31].

The final SNAP model coefficients (β_0 and $\boldsymbol{\beta}$ in equation 5) are provided in Table I below, while the optimized weights of the training data are provided in Table S1. The optimized cutoff radius R_c is 4.615858 Å, which is slightly larger than the third nearest neighbor distance in pristine bcc Mo.

IV. PERFORMANCE OF MO SNAP MODEL

In this section, we will compare the performance of the optimized Mo SNAP model in predicting many properties of interest.

TABLE I: SNAP coefficients for Mo.

k	$2j_1$	$2j_2$	$2j$	β_k
0				-17.2757958404
1	0	0	0	0.00431015861472
2	1	0	1	0.0657685117891
3	1	1	2	0.477733335702
4	2	0	2	0.0152688837211
5	2	1	3	0.77559888196
6	2	2	2	0.284846429566
7	2	2	4	0.148804982644
8	3	0	3	0.0573702179736
9	3	1	4	0.19281989434
10	3	2	3	0.323441703578
11	3	2	5	0.101324335724
12	3	3	4	0.0139639846514
13	3	3	6	-0.0324444749083
14	4	0	4	0.0349797952779
15	4	1	5	0.0613023441282
16	4	2	4	0.0881078513046
17	4	2	6	0.118716074611
18	4	3	5	0.0069662975532
19	4	4	4	-0.0174658914685
20	4	4	6	-0.0178902177779
21	5	0	5	0.0195993756659
22	5	1	6	0.0719238160707
23	5	2	5	0.0344832661036
24	5	3	6	-0.0358616891662
25	5	4	5	-0.0292380783172
26	5	5	6	-0.0334933909866
27	6	0	6	0.00595462520243
28	6	2	6	0.0754556638328
29	6	4	6	-0.000972545258845
30	6	6	6	-0.0100170422751

A. Energies, forces and stresses

Figure 3 shows the comparison between the DFT and SNAP predicted energies, forces and stress components using the training dataset. It should be noted that this comparison based on training data only validates whether the SNAP formalism can capture the complex relationship between variations in local environments and the energies and forces. Actual generalizability of the model beyond the training data are tested using separate sets of dat in following sections. For all three quantities, the SNAP model predictions are in line with the DFT results with a unity slope. The mean absolute error (MAE) between the DFT and SNAP predictions are 8.9 meV/atom, 0.30 eV/Å, and 0.89 GPa for the energies, forces and stress components, respectively. In comparison, the corresponding MAEs in the energies, forces, and stress components for the Mo EAM potential of Zhou *et al.*[32] are 122 meV/atom, 0.41 eV/Å and 3.48 GPa, respectively, while those for the Mo MEAM potential of Park *et al.*[33] are 70 meV/atom, 0.22 eV/Å, and 1.25 GPa. The Mo SNAP model in this work provides almost an order of magnitude better accuracy in the energy predictions, and good improvement in the accuracy of the stress components. For the prediction of forces, the Mo SNAP model clearly outperforms the EAM potential, but performs slightly worse than the MEAM potential. Nevertheless, it should be noted that the range of the DFT forces spans from around -23 eV/Å to 25 eV/Å , and the error of SNAP and MEAM compared to this range is small. As shown in Figure S1b, the force predictions of MEAM start to deviate substantially when the DFT forces are larger than 10 eV/Å, while both the SNAP and EAM models maintain a reasonably linear correlation with the DFT calculated forces.

To further validate our model, we performed DFT calculations on the previously excluded vacancy structures and used this dataset of 120 structures as a test set. We need to stress that such validation of the model on previous unseen data is necessary, since it gives indications of how the model will be generalized. The predicted MAEs for the energies, forces and stress components are 5.9 meV/atom, 0.27 eV/Å and 0.85 GPa respectively, comparable to
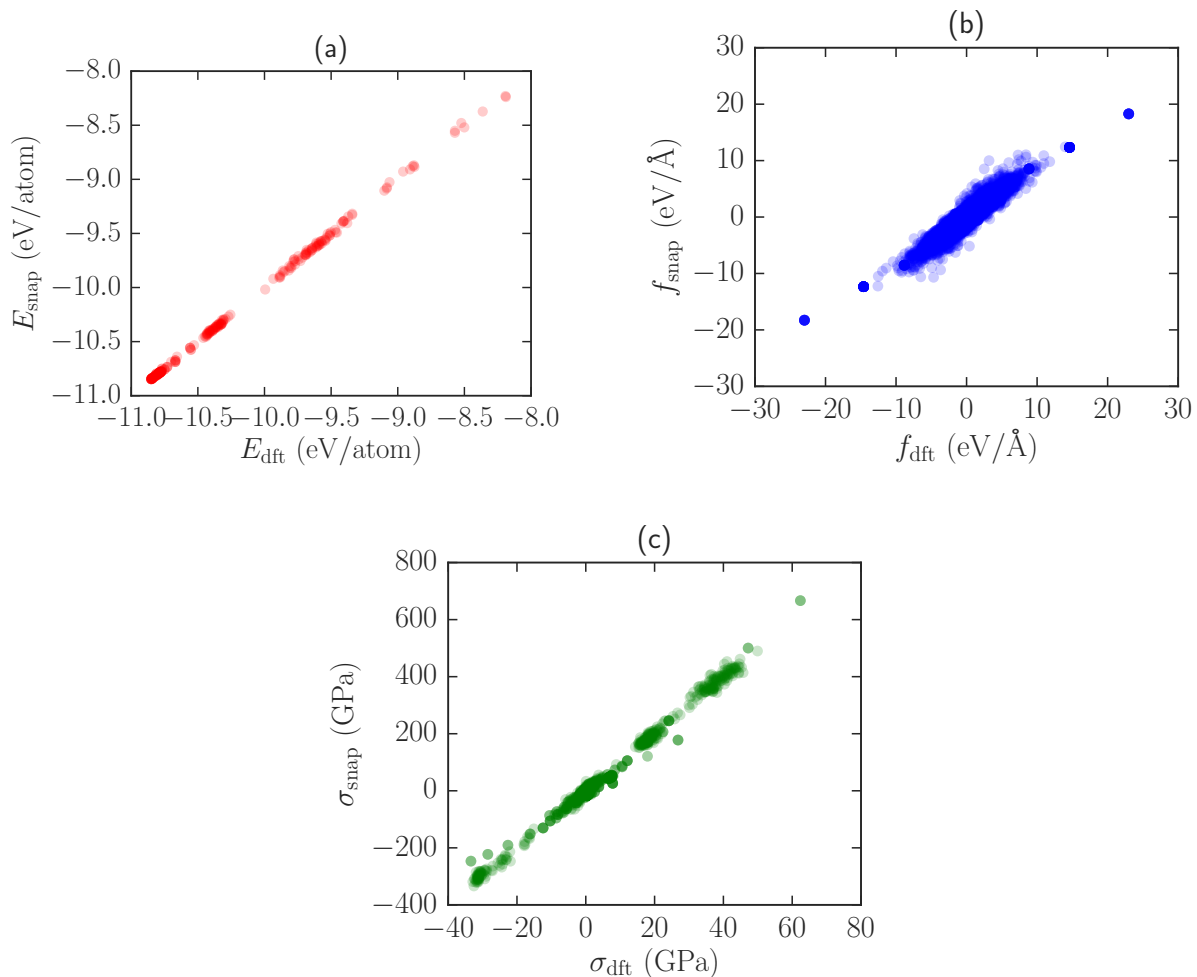
FIG. 3: SNAP predictions compared with DFT for (a) energy (b) force and (c) virial stress component.

the model performance on the training datasets. The results confirm that the bispectrum coefficient features are remarkably effective at distinguishing variations in the local environment, and our feature distribution PCA data selection strategy is valid.

B.   Lattice constant, elastic constants and equation of state

Table II provides a comparison of the Mo SNAP model predictions of the cubic lattice constant and elastic properties of bcc Mo with other force fields and experiments[51–53]. We find that the calculated cubic lattice constant and elastic properties of the SNAP model are in extremely good agreement with the DFT and experimental values. For example, the SNAP predictions of c_{11}, c_{12} and c_{44} are 473 GPa, 158 GPa and 106 GPa respectively, with errors of 0.2%, 3.8% and 0% compared to DFT, while the errors of the EAM (3.4-8.5%) and MEAM ($\sim 10\%$) potentials are significantly higher. The bulk modulus estimated using the Voigt-Reuss-Hill approximation[54] (B_{VRH}) with DFT, SNAP and EAM are in good agreement and slightly lower than the experimental value, but that from the MEAM potential is significantly underestimated.

We have also constructed the energy-versus-volume equation of state curves using DFT, SNAP, EAM and MEAM potentials in Figure 4. It should be noted that this set of data was not included in the training data and works as test data for model evaluation. We observe that the SNAP curve overlaps with DFT for volume changes in the range of -15% to 19% from the equilibrium volume, but begins to deviate slightly when a larger volume compression of magnitude $> 15\%$ is applied. The EAM potential deviates significantly from the DFT curve at both tensile and

TABLE II: Calculated cubic lattice parameter a, elastic constants (c_{ij}), Voigt-Reuss-Hill bulk modulus (B_{VRH})[54]., vacancy formation energy (E_v) and migration energy (E_m) with the DFT, SNAP, EAM and MEAM. Error percentages with respect to DFT values are shown in parentheses.

	DFT	SNAP	EAM[32]	MEAM[33]	Exp.
a (Å)	3.168	3.160 (-0.3%)	3.150 (-0.6%)	3.167 (0%)	3.147[52]
c_{11} (GPa)	472	473 (0.2%)	456 (-3.4%)	423 (-10%)	479[51]
c_{12} (GPa)	158	152 (-3.8%)	167 (5.7%)	143 (-9.5%)	165[51]
c_{44} (GPa)	106	107 (0.9%)	115 (8.5%)	95 (-10.4%)	108[51]
B_{VRH} (GPa)	263	259 (-1.5%)	264 (0.4%)	236 (-10.3%)	270[51]
E_v (eV)	2.87	2.61 (-9.1%)	3.02 (5.2%)	2.99 (4.2%)	-
E_m (eV)	1.12	1.39 (24.1%)	1.54 (37.5%)	1.64(46.4%)	-
$E_a = E_v + E_m$ (eV)	3.99	4.00 (-0.1%)	4.56 (14.3%)	4.63 (16.0%)	4.00 (1850-2350 °C)[53]

compressive strains, while for the MEAM potential, the agreement with DFT is slightly better than the SNAP model. By fitting the Murnaghan equation of state, the estimated bulk moduli from Figure 4 are 259, 261, 254, and 261 GPa for DFT, SNAP, EAM and MEAM respectively. We note that this estimate of the bulk modulus for the MEAM potential deviates significantly from that estimated using the Voigt-Reuss-Hill approximation (Table II), and is in much better agreement with the experimental value.
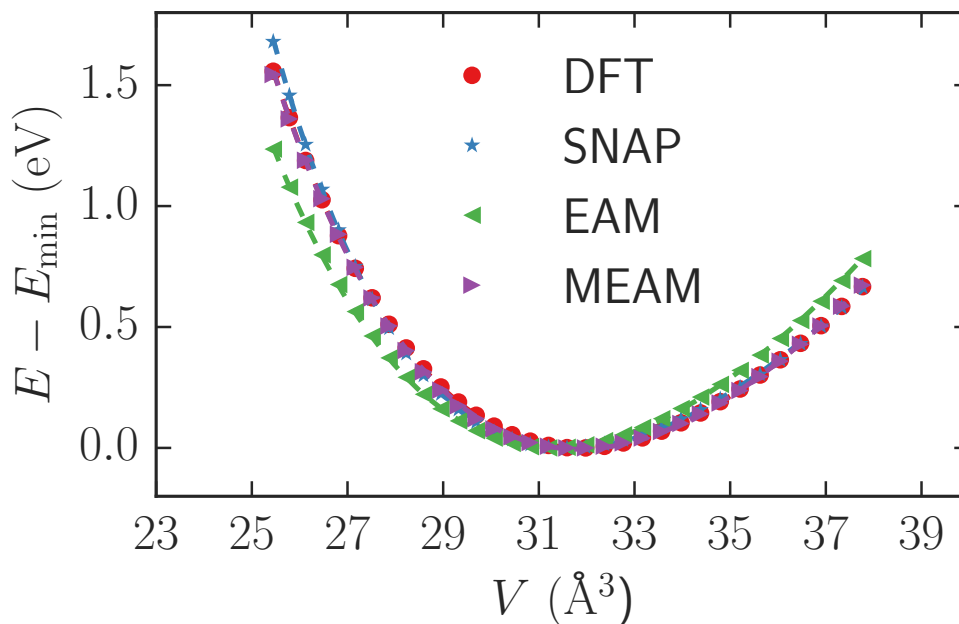

FIG. 4: Energy versus volume curves of a conventional bcc Mo cell for the DFT, SNAP, EAM and MEAM models. The dash lines show the fitted Murnaghan equation of state. The energy at the equilibrium volume has been set as the zero reference.

C. Lattice dynamics

To further investigate the prediction of forces and lattice dynamics using the Mo SNAP, the phonon dispersion curve of a Mo $5 \times 5 \times 5$ supercell containing 250 atoms was calculated by feeding the force predictions into the phonopy[55] package and is shown in Figure 5. The predicted phonon dispersion curves are in good agreement with

the DFT calculated phonon dispersion curves, though a systematic slight overestimation (relative to DFT) of the calculated frequencies are seen with the SNAP. No imaginary frequencies are observed, and the lowest frequency lies on the Γ point. The phonon dispersion curves of both EAM and MEAM are provided in Figure S2. Both EAM and MEAM show the same overestimation. We further calculated the thermal properties of Mo by employing different models. Figure 6 shows the Helmholtz free energy (A), entropy (S) and constant volume molar thermal capacity (C_v) of Mo calculated by DFT and SNAP. Both approaches show almost identical curves for all the three quantities. The estimated value of the heat capacity C_v at 300 K is 23.307 J/K/mol for DFT and 23.244 J/K/mol for the SNAP model, an discrepancy of only 0.27%. The EAM and MEAM potentials also similarly give relatively good predictions of the thermal properties, as shown in Figure S3.
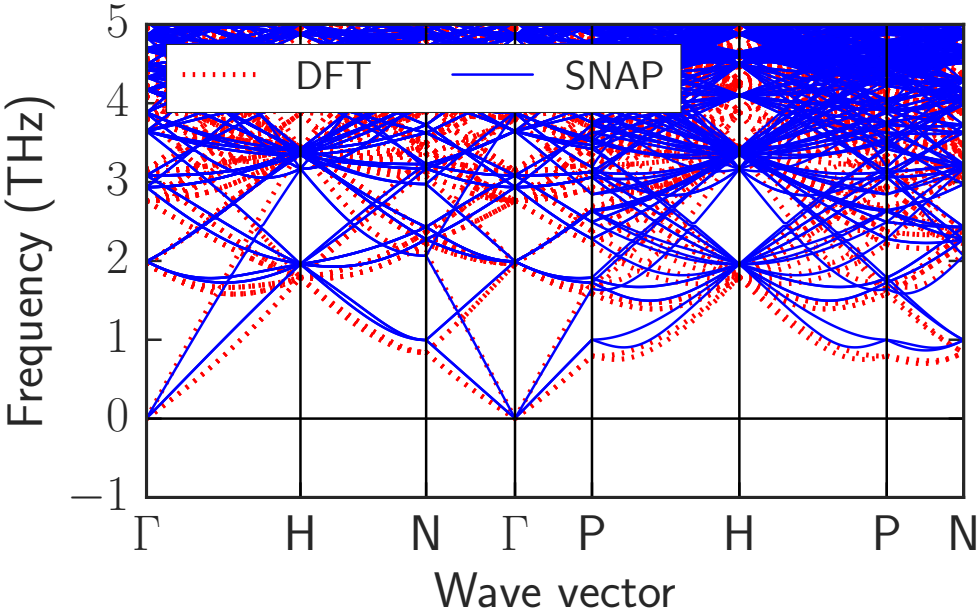

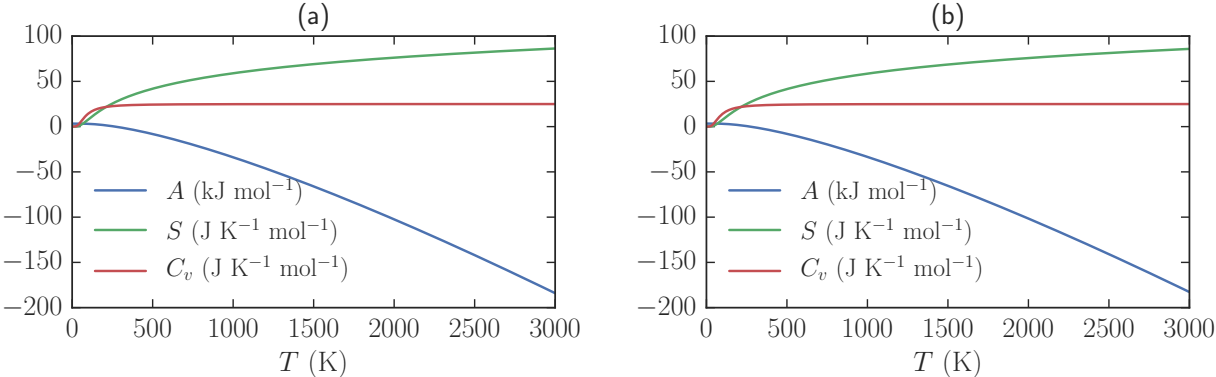FIG. 5: Phonon dispersion curves of Mo SNAP model compared to DFT.

FIG. 6: Helmholtz free energy (A), entropy (S) and constant volume molar thermal capacity (C_v) of Mo calculated with DFT (a) and SNAP (b).

D.   Vacancy formation and migration energies

The formation and migration of defects such as vacancies are of immense interest in practical applications of Mo, especially at high temperatures. Here, we estimate the vacancy formation and migration energies with the climbing-image nudged elastic band (CI-NEB) method[56] as well as molecular dynamics (MD) using the Mo SNAP model.

The DFT, SNAP, EAM and MEAM calculated Mo vacancy formation energy E_v and CI-NEB migration energy E_m are given in Table II. The SNAP model underestimates E_v by $\sim 11\%$, while EAM and MEAM overestimate it by 4-5%. For all force fields, E_m is predicted to be higher than the DFT value, though that predicted by the SNAP model is the closest. The overall predicted activation barrier for vacancy diffusion ($E_a = E_v + E_m$) are 3.99, 4.00, 4.56 and 4.37 eV for the DFT, SNAP, EAM and MEAM models, respectively. Both the DFT and SNAP predicted E_a are close to the experimental activation energy of 4.00 eV (measured on single crystal Mo at the temperature range from 1850-2350 °C[53]).

We have also performed MD simulations of a $10 \times 10 \times 10$ Mo bcc cell containing one vacancy (1999 atoms with the vacancy concentration of 0.05%) over 500 ps at seven temperatures (1500-2900 K) using the Mo SNAP. Simulations with this system size and running time are near the upper limit of AIMD simulations. NpT simulations were carried out using LAMMPS with a time step 1 fs. For each simulation, an equilibration run was carried out over 10 ps, and data was collected during the production run of 500 ps. The mean-squared displacements (MSDs) with respect to simulation time are shown in Figure 7(a). The diffusion of Mo is extremely slow below the melting point, and the calculated self-diffusivity is $2.57 \times 10^{-8} \mathrm{cm^2/s}$ at 2500 K. This is two orders of magnitude higher than the measured experimental diffusivity of about $8.22 \times 10^{-10} \mathrm{cm^2/s}$ at 2513 K[53]. We attribute this discrepancy to the fact that the experimental vacancy concentration may be much lower given the high vacancy formation energy, and is temperature dependent. In the MD simulation, a vacancy was artificially introduced, and the vacancy concentration was fixed at all temperatures. From the Arrhenius plot in Figure 7(b), we estimate the vacancy migration barrier for Mo to be 1.46 eV, which is consistent with the CI-NEB calculated value E_m in Table II.
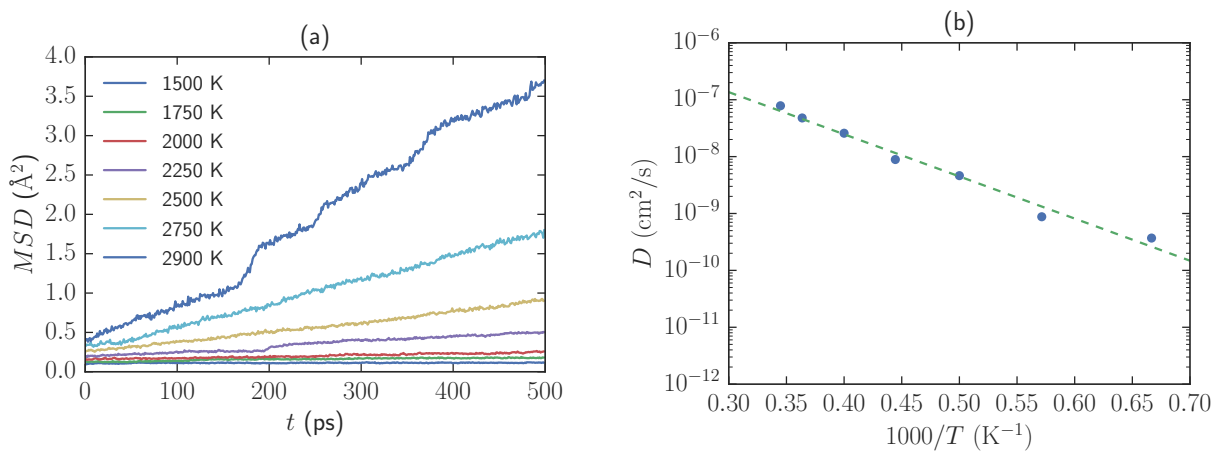

FIG. 7:   (a) MSDs of a $10 \times 10 \times 10$ Mo bcc cell with 1 vacancy (0.05% concentration) at various temperatures and (b) the corresponding Arrhenius plot.

E.   Melting point

We investigated the ability of the Mo SNAP to reproduce the melting point, which is one of the most challenging properties for conventional force fields to predict. The challenge arises because the phase transition results in a sharp change in the interatomic forces, which are difficult to describe in simple interaction terms. In this work, we performed an NpT MD simulation of a $6 \times 6 \times 6$ conventional Mo bcc cell using the SNAP, EAM and MEAM models. The simulation time step was set to 1 fs. The simulations were started at 300 K, and the temperature was ramped up to the desired temperature over 1 ps, followed by equilibration over 10 ps before data collection. Figure 8 plots the cell volume against the temperature. We find that the SNAP model predicts a melting point of 3000 K, which is in good agreement with the experimental melting point of 2890.15 K. In contrast, both the EAM and MEAM potentials

significantly overestimate the melting point of Mo by more than 700 K. We attribute the much better prediction of the SNAP model to the effectiveness of the bispectrum coefficients as a local environment descriptor (whether in the solid or liquid state).
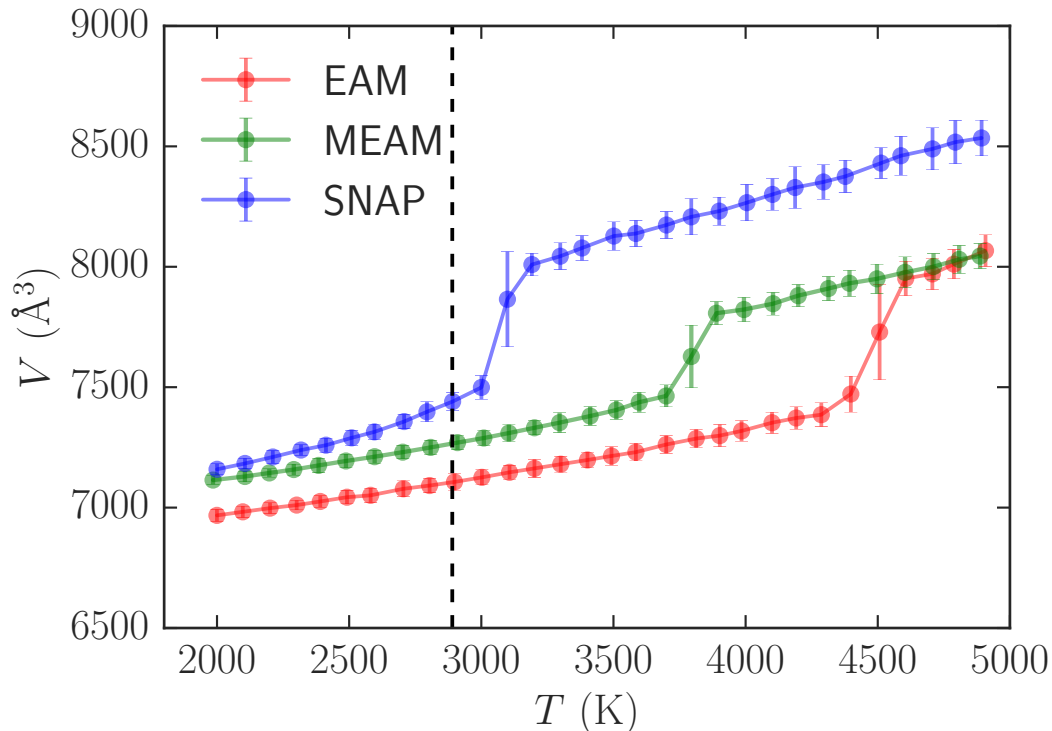

FIG. 8: Heating simulations of a $6 \times 6 \times 6$ Mo cell. The vertical dash line indicates the experimental melting point at 2890.15 K. Error bars of the volume at each temperature are shown.

F. Surfaces and grain boundaries

Finally, the performance of the Mo SNAP was assessed with regards to its ability to predict surface and GB energies. As can be seen from Figure 9, the surfaces energies computed by the Mo SNAP using the DFT-relaxed slab structures are in excellent agreement with the DFT calculations for both low and high Miller index surfaces. If the slab structures are first relaxed with the SNAP model, the predicted surface energies are somewhat lower than those from DFT calculations. Regardless of whether the slabs are relaxed with the SNAP model, the qualitative trends are reproduced well, as shown in Figure S4. For example, the lowest energy surface is predicted to be (110), with the (111) surface only slightly higher in energy. The errors of EAM and MEAM potentials in predicting the surface energies are much higher. More importantly, the predicted surface energies by the EAM and MEAM potentials are qualitatively different from those of DFT. For example, the predicted surface energies of (111), (322) and (332) are higher than that of (100), which is the opposite from that of DFT calculations.

From the previous work by the authors[34], the calculated DFT energies of the (100) Σ5 twist and (310) Σ5 tilt GBs are 2.46 J/m^2 and 1.81 J/m^2, respectively. The optimized SNAP model predicts GB energies of 2.52 J/m^2 and 1.94 J/m^2 for the (100) Σ5 twist and (310) Σ5 tilt GBs, which are in good agreement with the DFT values.

V. DISCUSSION

A. Limitations of SNAP model

In this work, we have developed a SNAP model for Mo that significantly outperforms existing EAM and MEAM potentials across a broad range of properties, including energies, structural stresses, elastic constants, thermochemical
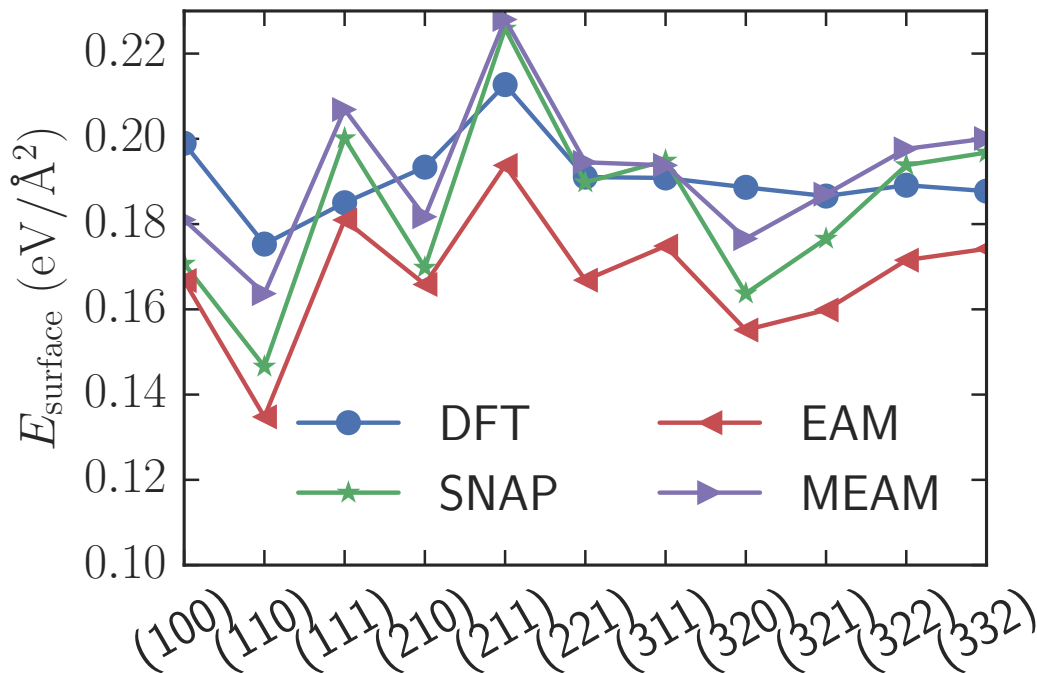
FIG. 9: Comparison of calculated surface energies for surfaces with Miller indices up to a maximum of 3 using DFT, SNAP, EAM and MEAM.

properties, melting point, surface and grain boundary energies. We attribute this overall better performance to the effectiveness of the bispectrum coefficients as a unique descriptor for the local environment that is invariant to transformations that preserve material properties.

The most notable "failure" of the optimized SNAP model is in the prediction of the vacancy formation energy, where the EAM and MEAM potentials perform significantly better (see Table II). We speculate that this could be due to the fact that the SNAP models the energy as a simple linear relation to the bispectrum descriptors, which undergo a large, discontinuous change when a vacancy is introduced. This is a deficiency that can potentially be addressed by relaxing the linear constraint. On the other hand, the SNAP model provides significantly more accurate vacancy migration energies compared to the EAM and MEAM models. We believe this is because the pairwise interactions in EAM and MEAM models are not able to capture the transition state with sufficient accuracy. The overall errors in the vacancy formation and migration energies cancel out, resulting in overall activation energies that are remarkably close to experimental values.

Another disadvantage of the SNAP model is its higher ($\sim 2 - 3$ orders of magnitude) computational cost relative to MEAM. Nevertheless, the SNAP model is still orders of magnitude cheaper than DFT calculations and have the advantage of approximately linear scaling with respect to number of atoms[30]. We have performed static calculations using both DFT and the SNAP model for $2 \times 2 \times 2$ and $3 \times 3 \times 3$ supercells of Mo. On a 16-core CPU, the DFT calculations took 303 and 11,869 seconds for the $2 \times 2 \times 2$ and $3 \times 3 \times 3$ supercells, respectively. On the other hand, the SNAP computations on a single core took 0.0052 and 0.0168 seconds, respectively. These results demonstrate a $>$ 5 orders of magnitude difference in computational cost between DFT and SNAP, which is further amplified for large systems due to the almost linear scaling of SNAP. Furthermore, using a single computing node of 24 cores, we were able to carry out a MD simulation of ~ 2000 atoms for hundreds of ps ($> 100,000$ timesteps of 1 fs) over 24 hours. This scale of simulations would be otherwise be impossible, or at the very least, can only be conducted with great difficulty and cost, with first principles methods. Simulations of tens of thousands of atoms with SNAP should be well within the capabilities of modern supercomputing clusters. Access to such length and time scales with close to DFT chemical accuracy would significantly enhance our ability to probe interesting science in lower symmetry systems, for example, crack propagation and effect of dopants in high Σ or general grain boundaries, which, unlike the low Σ, high-symmetry idealized GBs that have been studied so far, requires large models with hundreds or even thousands of atoms. Finally, we note that the SNAP formalism is not limited to single component systems. The number of parameters in SNAP models scale linearly with the species number, compared to the quadratic scaling for classical pair potentials. The extension of SNAP models to multi-component systems such as Mo alloys and other systems is

currently ongoing, and the Mo elemental SNAP model in this work serves as an important starting point.

B. Model development

We have also demonstrated several enhancements that we believe would be of relevance to future efforts in the development of machine-learned potentials.

First, we have leveraged on a combination of existing data from previous works[34,35,38,39], as well as newly generated structures from high-throughput DFT calculations for model training and testing. The use of well-validated, pre-existing data led to a significantly more streamlined and efficient model development process. The pre-existing data in this work are mostly from previous works by the current authors, some of which (e.g., surface energies) are in large, open access online databases.[35,39,57] We foresee that the continued proliferation of such central materials data repositories would provide future potential developers with a wealth of data for model training and testing. We note that a key gap is that existing open databases such as the Materials Project[38] are focused mainly on ground state structures, energies and properties (e.g., elastic constants). To our knowledge, there are currently no open databases for trajectory data from AIMD simulations, which has been shown to be an extremely rich data set for model training. Together with this work, we have published the trajectory data for the Mo AIMD simulations (as well as all other data used in model development and the final data weights and optimized SNAP model parameters) in an open repository hosted on Github (https://github.com/materialsvirtuallab/snap). We hope to address this gap through a more systematic large-scale effort encompassing more diverse chemistries in future.

Second, we have shown that an exploratory data analysis performed prior to expensive DFT calculations can result in better quality training data with lower computational effort. In this work, we utilized a PCA approach to identify the datasets that provide distinct local environment information. This data selection process avoids biasing the training with duplicate data that share common features and also improves the accuracy of the eventual model. For instance, we have previously excluded the vacancy data group based on its overlap with the AIMD groups.

Third, similar to previous works, we find that the hyperparameters, i.e., the cut-off radius R_c for the bispectrum calculations and weights for different data group, can influence the model performance and have to be optimized together with the model training. The choice of the cut-off radius R_c has a substantial effect on model accuracy (e.g., in energy and force predictions), while the choice of data weights tend to impact derived material properties. For example, the c_{12} elastic constant can vary from 20 GPa to 400 GPa with changes in data weights. A global optimization approach, such as the differential evolution algorithm used in this work, can improve the likelihood of obtaining good solutions, if the bounds are properly set. We note that Bartók *et al.*[26] subsequently further enhanced the bispectrum coefficients by incorporating a Gaussian atomic density contribution, i.e., the so-called Smooth Overlap of Atomic Positions (SOAP) approach. Slightly improved performances were obtained with a Gaussian Approximation Potential using SOAP compared to the bispectrum coefficients on Si. Our decision to use the SNAP formalism is motivated by both by the simplicity and extensibility of the functional form (a topic we will explore in future works), as well as the practical advantage of SNAP being already implemented in the popular LAMMPS package, which makes it accessible to a broad range of users.

VI. CONCLUSION

To conclude, we have developed a SNAP model for Mo by applying a systematic data selection and global optimization approach on a combination of existing as well as newly generated data from DFT calculations. The optimized SNAP model outperforms existing EAM and MEAM potentials for Mo, achieving close to DFT accuracy in the prediction of a spectrum of properties of immense fundamental and technological relevant, including energies, forces, elastic constants, melting point, surface and grain boundary energies. We believe this new chemically-accurate Mo SNAP potential open ups our ability to probe properties of Mo in simulations of large-scale models over longer time scales.

VII. ACKNOWLEDGEMENT

* ongsp@eng.ucsd.edu

[1] W. Kohn and L. J. Sham, Physical Review **140**, A1133 (1965).

[2] L. J. Sham and M. Schlüter, Physical Review Letters **51**, 1888 (1983).

[3] T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, and H. M. Aktulga, npj Computational Materials **2**, 15011 (2016).

[4] A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard Iii, and W. M. Skiff, Journal of the American Chemical Society **114**, 10024 (1992).

[5] L. Zdeborová, Nature Physics **13**, 420 (2017).

[6] G. Carleo and M. Troyer, Science **355**, 602 (2017).

[7] J. Carrasquilla and R. G. Melko, Nature Physics **13**, 431 (2017).

[8] Z. W. Ulissi, A. J. Medford, T. Bligaard, and J. K. Nørskov, Nature Communications **8** (2017).

[9] A. Jain, G. Hautier, S. P. Ong, and K. Persson, Journal of Materials Research (2016), 10.1557/jmr.2016.80.

[10] C. Chen, D. Chen, and F. Ciucci, Physical Chemistry Chemical Physics **17**, 7831 (2015).

[11] C. Chen, Z. Lu, and F. Ciucci, Scientific Reports **7**, 40769 (2017).

[12] Z. Li, J. R. Kermode, and A. De Vita, Physical Review Letters **114**, 96405 (2015).

[13] A. Glielmo, P. Sollich, and A. De Vita, Physical Review B **95**, 214302 (2017).

[14] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, The Journal of Physical Chemistry Letters **6**, 2326 (2015).

[15] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, New Journal of Physics **15**, 95003 (2013).

[16] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Scientific Reports **3** (2013).

[17] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Physical Review B **89**, 94104 (2014).

[18] R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sanchez-Carrera, L. Vogt, and A. Aspuru-Guzik, Energy & Environmental Science **4**, 4849 (2011).

[19] J. Behler and M. Parrinello, Physical Review Letters **98**, 146401 (2007).

[20] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Physical Review Letters **104**, 136403 (2010).

[21] V. Botu and R. Ramprasad, International Journal of Quantum Chemistry **115**, 1074 (2015).

[22] A. V. Shapeev, Multiscale Modeling & Simulation **14**, 1153 (2016).

[23] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Physical Review Letters **108**, 58301 (2012).

[24] J. E. Moussa, Physical Review Letters **109**, 59801 (2012).

[25] J. Behler, The Journal of Chemical Physics **134**, 74106 (2011).

[26] A. P. Bartók, R. Kondor, and G. Csányi, Physical Review B **87**, 184115 (2013).

[27] T. L. Pham, H. Kino, K. Terakura, T. Miyake, I. Takigawa, K. Tsuda, and H. C. Dam, arXiv preprint arXiv:1705.01043 (2017).

[28] H. Weyl, *The classical groups: their invariants and representations* (Princeton university press, 2016).

[29] S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi, Bioinformatics **21**, i359 (2005).

[30] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, Journal of Computational Physics **285**, 316 (2015).

[31] M. A. Wood and A. P. Thompson, arXiv preprint arXiv:1702.07042 , 1 (2017).

[32] X. W. Zhou, R. A. Johnson, and H. N. G. Wadley, Physical Review B **69**, 144113 (2004).

[33] H. Park, M. R. Fellinger, T. J. Lenosky, W. W. Tipton, D. R. Trinkle, S. P. Rudin, C. Woodward, J. W. Wilkins, and R. G. Hennig, Physical Review B **85**, 214121 (2012).

[34] R. Tran, Z. Xu, N. Zhou, B. Radhakrishnan, J. Luo, and S. P. Ong, Acta Materialia **117**, 91 (2016).

[35] R. Tran, Z. Xu, D. W. Balachandran Radhakrishnan, W. Sun, K. A. Persson, and S. P. Ong, Scientific Data **3** (2016).

[36] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. a. Persson, and G. Ceder, Computational Materials Science **68**, 314 (2013).

[37] F. Neri and V. Tirronen, Artificial Intelligence Review **33**, 61 (2010).

[38] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Materials **1**, 11002 (2013).

[39] R. Tran, Z. Xu, B. Radhakrishnan, and S. P. Ong, "Crystalium," (2016).

[40] M. De Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. K. Ande, S. Van Der Zwaag, and J. J. Plata, Scientific Data **2** (2015).

[41] G. Kresse and J. Furthmüller, Physical Review. B, Condensed Matter **54**, 11169 (1996).

[42] P. E. Blöchl, Physical Review B **50**, 17953 (1994).

[43] J. P. J. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **77**, 3865 (1996).

[44] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, and K. A. Persson, Concurrency Computation Practice and Experience **27**, 5037 (2015).

[45] S. Plimpton, Journal of Computational Physics **117**, 1 (1995).

[46] N. J. Browning, R. Ramakrishnan, O. A. Von Lilienfeld, and U. Roethlisberger, The Journal of Physical Chemistry Letters **8**, 1351 (2017).

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, Journal of Machine Learning Research **12**, 2825 (2011).

[48] B. M. Adams, W. E. Hart, M. S. Eldred, D. M. Dunlavy, P. D. Hough, A. A. Giunta, J. D. Griffin, M. L. Martinez-Canales, J.-P. Watson, and T. G. Kolda, *DAKOTA, a Multilevel Parellel Object-oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 4.0 Uers's Manual* (United States. Department of Energy, 2006).

[49] E. Jones, T. Oliphant, and P. Peterson, "SciPy: open source scientific tools for Python," (2014).

[50] B.-J. Lee, M. I. Baskes, H. Kim, and Y. K. Cho, Physical Review B **64**, 184102 (2001).

[51] G. Simmons and H. Wang, *Single crystal elastic constants and calculated aggregate properties* (Mass., MIt Press, 1971).

[52] W. B. Pearson, *A handbook of lattice spacings and structures of metals and alloys: International series of monographs on metal physics and physical metallurgy*, Vol. 4 (Elsevier, 2013).

[53] J. Askill and D. H. Tomlin, Philosophical Magazine **8**, 997 (1963).

[54] R. Hill, Proceedings of the Physical Society. Section A **65**, 349 (1952).

[55] A. Togo, F. Oba, and I. Tanaka, Physical Review B **78**, 134106 (2008).

[56] G. Henkelman, B. P. Uberuaga, and H. Jónsson, The Journal of Chemical Physics **113**, 9901 (2000).

[57] "Materials Project," (2011).