Data-driven material models for atomistic simulation

M. A. Wood, M. A. Cusentino, B. D. Wirth, and A. P. Thompson

Phys. Rev. B **99**, 184305 — Published 30 May 2019

DOI: 10.1103/PhysRevB.99.184305

Data-driven Material Models for Atomistic Simulation

M. A. Wood,[1] M. A. Cusentino,[1] B. D. Wirth,[2] and A. P. Thompson[1]

[1] *Center for Computing Research, Sandia National Laboratories, Albuquerque, New Mexico 87185, USA*
[2] *University of Tennessee, Knoxville, Tennessee 37996, USA*

(Dated: May 15, 2019)

The central approximation made in classical molecular dynamics simulation of materials is the interatomic potential used to calculate the forces on the atoms. Great effort and ingenuity is required to construct viable functional forms and find accurate parameterizations for potentials using traditional approaches. Machine-learning has emerged as an effective alternative approach to develop accurate and robust interatomic potentials. Starting with a very general model form, the potential is learned directly from a database of electronic structure calculations and therefore can be viewed as a multiscale link between quantum and classical atomistic simulations. Risk of inaccurate extrapolation exists outside the narrow range of time- and length-scales where the two methods can be directly compared. In this work, we use the Spectral Neighbor Analysis Potential (SNAP) and show how a fit can be produced with minimal interpolation errors which is also robust in extrapolating beyond training. To demonstrate the method, we have developed a new tungsten-beryllium potential suitable for the full range of binary compositions. Subsequently, large-scale molecular dynamics simulations were performed of high energy Be atom implantation onto the (001) surface of solid tungsten. The new machine learned W-Be potential generates a population of implantation structures consistent with quantum calculations of defect formation energies. A very shallow ($<$ 2nm) average Be implantation depth is predicted which may explain ITER diverter degradation in the presence of beryllium.

I. INTRODUCTION

Over the past few decades the rapid advancements and availability of computing technologies has changed the way research is conducted in many areas of science and engineering. Modern supercomputing systems enable researchers to perform hundreds or thousands of virtual experiments before setting foot in a traditional laboratory. One of the main advances with these computational efforts has been the curation of results into extensive open source databases, enabling the data to be used to drive materials discovery and model development, often in ways never intended by the originators.[1–5] A recent trend in material science is the adoption of data science techniques to derive new understanding of material properties from modeling and simulation. In the present work we use machine learning to bridge between quantum and classical atomistic simulation methods, which can be viewed as a particular case of data-driven materials modeling.

For materials behavior that originates at the atomic scale, molecular dynamics (MD) is a powerful and popular computational tool. This work highlights a computational multiscale approach where a database of electronic structure calculations is translated into a classical interatomic potential(IAP) for MD. Calculation of forces using an IAP is many orders of magnitude more computationally efficient than using quantum electronic structure methods such as density functional theory (DFT), while capturing the same essential physics. It is important to realize that the key approximation made in MD simulations is the interatomic potential. For this reason, great care needs to be taken in the IAP construction, as well as in interpretation of simulation results that are computed from a given potential. There are many different mathematical forms that can be used to construct an interatomic potential, many of these use physics and chemistry as a model[6–11] to determine the forces on the atoms. However, there is a recent trend of relying on machine-learning approaches[12–15] to construct an IAP that can significantly decrease the time investment needed while simultaneously improving the accuracy with respect to electronic structure predictions[16–18]. Additionally, this data-science approach to an IAP can be applied to materials with complex bonding characteristics which are challenging for traditional potentials[19–21].

An example case where traditional IAP have trouble representing atomic interactions is the W-Be material system which is of relevance to modeling plasma material interactions for fusion devices. In the International Thermonuclear Experimental Reactor (ITER), beryllium and tungsten have been chosen as the first wall and diverter materials, respectively. They have already been used in experimental fusion reactors[22,23]. Due to the low atomic number of beryllium and its favorable thermal conductivity, it is a suitable material for the first wall where impurity transport into the plasma is a concern[24]. On the other hand, the divertor region receives the highest ion and heat fluxes, on the order of 10^{24} $\mathrm{m^{-2}s^{-1}}$ and 10 $\mathrm{MWm^{-2}}$ respectively[24]. Tungsten has been chosen for these extreme conditions, because of its high melting point, good thermal conductivity, and low sputtering yield[24]. While the divertor region is expected to receive the highest ion and heat fluxes, some beryllium will be eroded from the first wall and deposited onto the divertor material[25,26]. This deposition of beryllium into the tungsten surface could lead to the formation of stable W-Be intermetallic compounds with much lower melting points than pure tungsten[27]. Any reduction in the melting point of the divertor material could lead to a drastic increase in sputtering yield and deterioration of the divertor performance. For this reason it is important to understand in detail how beryllium implants into tungsten and what types of mixed phases are formed near the surface.

Multiple experiments at PISCES-B (Plasma Interaction with Surface and Components Experimental Simulator)[28–30] of beryllium seeded deuterium plasma exposure of tungsten have been conducted to assess mixed material effects on deuterium retention and intermetallic formation. For plasmas containing as little as 0.1% beryllium electron microscopy images of the tungsten targets show both layers and deposits of various W-Be intermetallics including WBe_{12}[30]. Additional XPS measurements indicate the formation of WBe_2 during the annealing process from 300 K up to 970 K[31]. These experiments indicate that W-Be intermetallic formation in the diverter of ITER can occur and correspondingly, additional experimental and modeling efforts are needed to understand the underlying physical processes and mechanisms leading to intermetallic formation.

Molecular dynamics is well suited to modeling these effects. However, there are not many IAP developed for tungsten and beryllium and their accuracy is limited for this particular application. While many potentials exist for tungsten[32], only one exists for modeling W and Be[33], which is a Tersoff style bond order potential (BOP)[10]. This potential has been used to study both beryllium implantation in tungsten[34] and mixed beryllium-deuterium implantation in tungsten[35]. However, this potential form is not robust enough to capture the complex interactions between tungsten, beryllium, and their intermetallic structures. In this article we show how the Spectral Neighbor Analysis Potential (SNAP) machine learning technique can be used to derive an IAP for W-Be that is capable of studying in detail these mixed material interactions.

II. POTENTIAL ENERGY MODEL

An interatomic potential should accurately represent the many-body potential energy surface as a function of the local environment around an atom. By only considering neighbors within a distance of approximately 1 nm, classical MD simulations using parallel algorithms can be scaled far beyond what is possible for electronic structure codes. This remains true for the data-science inspired potentials.[36] Machine learned interatomic potentials (ML-IAP) can be distinguished from one another based on three key factors; regression technique, choice of descriptors and energy model form. Many of the recently developed machine learned interatomic potentials can be placed on a continuous scale of being more physical- or data-science based.[37] Deep neural networks (NN) with simple descriptors and activation functions[38–41] directly exploit the recent advances in the field of data-science. The key advantage of NN-based potentials is the immense flexibility of the model to capture even the most subtle features of the training data. A limiting factor of these ML-IAP is the uncertainty in extrapolating beyond the training data to predict energies and forces in previously unseen atomic environments. Nonparametric regression methods like Gaussian process[42] or kernel ridge regression[43] use physically motivated kernels like local atom densities or bond topology and are toward the center of this scale[44]. The Spectral Neighborhood Analysis Potential (SNAP)[45], which is used in this work, is more strongly physics-based, due to its use of the bispectrum as descriptors, which are closely related to invariants of the radial and angular basis functions of the atomic cluster expansion that is the natural description of the bonding environment around an atom[46,47]. Additionally, for simplicity and computational efficiency, SNAP uses linear regression in order to decouple the computational cost at MD runtime from the details of the training set used.

A. Spectral Neighborhood Analysis Potential

We outline here the structure of the SNAP ML-IAP in terms of the underlying descriptor space.[45] The total potential energy of a configuration of atoms is first written as the sum of SNAP energy contributions associated with each atom, combined with a reference potential

$$E(\mathbf{r}^N) = E_{ref}(\mathbf{r}^N) + \sum_{i=1}^{N} E^i_{SNAP}, \tag{1}$$

where $\mathbf{r}^N$ is the vector of N atom positions in the configuration. E and E_{ref} are the total and reference potential energies, respectively. E^i_{SNAP} is the SNAP potential energy associated with a particular atom i, and depends only on the relative positions of its neighbor atoms. Including a reference potential is advantageous because it can correctly represent known limiting cases of atomic interactions, leaving the SNAP contribution to capture many-body effects. The ZBL pair potential[48] is a convenient choice, because it captures the known short-range repulsive interactions between atomic cores that are not well represented by quantum calculations.

The construction of the SNAP component of the potential energy in terms of the bispectrum components follows the same approach described in Ref. [45], which we briefly summarize here. The SNAP formulation begins with a very general characterization of the neighborhood of an atom. The density of neighbor atoms at location **r** relative to a central atom i located at the origin can be considered as a sum of δ-functions located in a three-dimensional space:

$$\rho_i(\mathbf{r}) = \delta(\mathbf{r}) + \sum_{r_{i'} < R_{ii'}} f_c(r_{i'}) w_{i'} \delta(\mathbf{r} - \mathbf{r}_{i'}) \tag{2}$$

where $\mathbf{r}_{i'}$ is the position of neighbor atom i' relative to central atom i. The $w_{i'}$ coefficients are dimensionless weights that are chosen to distinguish atoms of different types, while the central atom is arbitrarily assigned a unit weight. This sum is over all atoms i' within the cutoff distance $R_{ii'}$ that is defined in terms of the effective radii of the two atoms

$$R_{ii'} = \alpha(R_i + R_{i'}), \tag{3}$$

where α is a universal scale factor and R_i and $R_{i'}$ are the effective radii of atom i and i' respectively. The switching function $f_c(r)$ ensures that the contribution of each neighbor atom goes smoothly to zero at $R_{ii'}$.

Typically, this density function is expanded in an angular basis of spherical harmonics combined with an orthonormal radial basis.[13] Instead, we use an idea originally proposed by Bartók et al.[42], in which the radial coordinate r is mapped on to a third angular coordinate $\theta_0 = \theta_0^{max} r / R_{ii'}$.

Each neighbor position (r, θ, ϕ) is mapped to (θ_0, ϕ, θ), a point on the unit 3-sphere. The natural basis for functions on the 3-sphere is formed by the 4D hyperspherical harmonics $U^j_{m,m'}(\theta_0, \theta, \phi)$, defined for $j = 0, \frac{1}{2}, 1, \ldots$ and $m, m' = -j, -j+1, \ldots, j-1, j$[49]. The neighbor density function can now be expanded in the basis of hyperspherical harmonics $U^j_{m,m'}$. Because the neighbor density is a weighted sum of δ-functions, each expansion coefficient is a sum over discrete values of the corresponding basis function evaluated at each neighbor position

$$u^j_{m,m'} = U^j_{m,m'}(0) + \sum_{r_{i'} < R_{ii'}} f_c(r_{i'}) w_{i'} U^j_{m,m'}(\theta_0, \theta, \phi) \tag{4}$$

The bispectrum components are formed as the scalar triple products of the expansion coefficients

$$B_{j_1,j_2,j} = \sum_{m,m'} u^{j*}_{m,m'} \sum_{\substack{m_1,\, m'_1 \\ m_2,\, m'_2}} H^{jmm'}_{\substack{j_1 m_1 m'_1 \\ j_2 m_2 m'_2}} u^{j_1}_{m_1,m'_1} u^{j_2}_{m_2,m'_2} \tag{5}$$

where $*$ indicates complex conjugation and the constants $H^{jmm'}_{\substack{j_1 m_1 m'_1 \\ j_2 m_2 m'_2}}$ are Clebsch-Gordan coupling coefficients for the hyperspherical harmonics. Importantly, the bispectrum components are real-valued and invariant under rotation[42]. They are also symmetric in the three indices j_1, j_2, j up to a normalization factor.[45] They characterize the strength of density correlations at three points on the 3-sphere. The lowest-order components describe the coarsest features of the density function, while higher-order components reflect finer detail. The number of distinct bispectrum components with indices j_1, j_2, j less than or equal to J increases as J^3. For a particular choice of J, we can list the K bispectrum components in some arbitrary order as $B_1, \ldots, B_K$. The SNAP energy of an atom is written as a linear function of the bispectrum components

$$E^i_{SNAP} = \beta_0 + \sum_{k=1}^{K} \beta_k (B^i_k - B^i_{k0}) \tag{6}$$

$$= \beta_0 + \boldsymbol{\beta} \cdot \mathbf{B}^i \tag{7}$$

where B^i_k is the kth bispectrum component of atom i and β_k is the associated linear coefficient, a free parameter in the SNAP model. As a computational convenience, the contribution of each bispectrum component to the SNAP energy is shifted by the contribution of an isolated atom, $\beta_k B^i_{k0}$, so that the SNAP energy of the isolated atom is equal to β_0 by construction. Similarly, the force on each atom j due to the SNAP potential can be expressed as a weighted sum over the derivatives w.r.t. $\mathbf{r}_j$ of the bispectrum components of each atom i.

$$\mathbf{F}^j_{SNAP} = -\nabla_j \sum_{i=1}^{N} E^i_{SNAP} = -\boldsymbol{\beta} \cdot \sum_{i=1}^{N} \frac{\partial \mathbf{B}^i}{\partial \mathbf{r}_j} \tag{8}$$

In this way, the total energy, forces, and also the stress tensor, can be written as linear functions of quantities related to the bispectrum components of the atoms. In addition to shifting the bispectrum components by B^i_{k0}, it also makes sense to set $\beta_0 = 0$, constraining the potential energy of an isolated atom to be zero. This ensures that SNAP correctly reproduces the cohesive energy of the reference solid structure, an important physical attribute of any general purpose interatomic potential. For multi-element systems, such as the tungsten-beryllium materials considered here, SNAP captures the effect of compositional differences in several ways. Firstly, the coefficients $\boldsymbol{\beta}$ are different for each element. Secondly, the contributions to the basis functions in Eq. 4 made by each atom depend on the element weight $w_{i'}$ and the effective atomic radius $R_{i'}$.

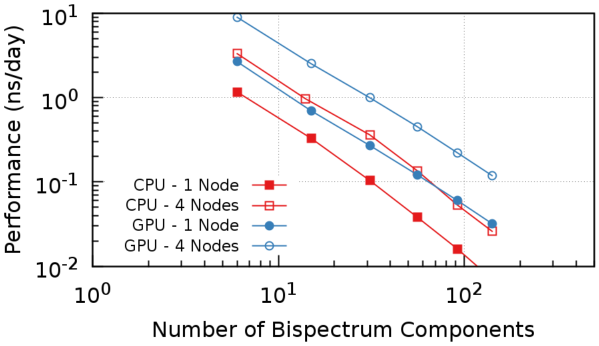


Figure 1. Simulation rate (ns/day) for an NVE MD simulation consisting of 31k atoms versus the number of descriptors used in the SNAP potential. The benchmark was run on one and four CPU or GPU nodes. Each CPU node consists of two Intel Broadwell E5-2695 v4 processors, with a total of 36 physical cores per node. The GPU node consisted of four NVIDIA P100 cards. Both the CPU and GPU systems use an Omni-path interconnect.

B. Computational Efficiency

Implementation of SNAP in LAMMPS[50,51] uses the KOKKOS library[52,53], allowing the code to run efficiently on diverse hardware architectures, including CPU, GPU, and many-core processors. The implementation also exploits LAMMPS highly-scalable MPI-based spatial decomposition scheme, allowing a single MD simulation to be distributed over a few nodes or an entire supercomputer[52]. The spatial resolution of the potential can be continuously improved by increasing the number of bispectrum components, systematically increasing the accuracy of the SNAP potential, at the price of greater computational cost. This is illustrated in Figure 1 where the number of bispectrum components is increased from 6 to 141, increasing the computational cost by over two orders of magnitude. Performance is reported as the amount of MD simulation time that can be calculated in a given amount of wall-clock time (ns/day). The data displayed here is for a benchmark problem consisting of 31,250 tungsten atoms, running molecular dynamics in the microcanonical ensemble with a timestep of 0.5 fs. Figure 1 compares a traditional CPU (Intel Broadwell) compute node to a modern multi-GPU (four NVIDIA P100's) compute node. SNAP scales comparably on either hardware, but there are significant performance gains when multiple GPU cards are assembled onto a single compute node.

III. TRAINING A MACHINE LEARNED MODEL

A. Constructing the Training Set

The present work is focused on generating a SNAP interatomic potential for tungsten-beryllium with an intended use in simulating plasma facing components in a fusion reactor. As such, a training set must be constructed that reflects the material properties relevant to this application space, but also is of general use to end users. Given the highly flexible nature of machine learned potentials, any reference model can be taken as a training set. However, we employ SNAP as a multiscale link between density functional theory (DFT) and MD, and as such will need a data base of expensive electronic structure calculations to properly train the model. Constructing a training set is a critical part of any machine learning endeavor because the constructed model will, by default, be best at interpolating between data it has already seen. Therefore, when it comes to an IAP, the more diverse the atomic configurations included in the training set the better suited for *general use* the resultant potential should be. Domain size limits within DFT imposes some restrictions of what types of training configurations can be included with an upper limit around a few hundred atoms. Atomic configurations within these size limitations need to be chosen such that they represent the material properties and application space of interest. There are no well-defined rules for how best to construct training set for ML-IAP generation. Physical insight and expert domain knowledge of the materials science application are needed to guide the selection of the DFT atomic configurations. Alternative methods such as learning *on-the-fly*[54–56] have been proposed as unsupervised approaches to training set construction, but this is still an area of active research.

Presently, we have chosen to curate the training set by hand. The constructed training set can be divided into three general categories: DFT calculations of pure tungsten, pure beryllium, and those containing both elements. Table I lists all of the training data used, as well as the number of energy (N_E) and force (N_F) points that each group contributes to the overall fit. Beginning with the pure tungsten training data, a number of configurations were taken from a data set previously used to fit a GAP potential for tungsten[57,58]. These are the Dislocations, isothermal *ab initio* MD, Elastic Deformations, Surfaces, Monovacancies, and two Γ-surface groups. Additional DFT calculations were carried out to add the Self-Interstitials, Liquids, Divacancy and Equation of State training groups to the set. Pure tungsten training calculations were performed with VASP[59–61] using a 600eV plane wave cutoff energy, approx. 0.015 Å^{-1} (depends on configuration) k-point spacing, a PBE-GGA exchange-correlation functional[62–64] and a pseudopotential that leaves the outermost s-,p- and d-orbitals to the be solved by the basis set. Additional details on how these training data were generated can be found in the supplemental material[65].

While there are any number of additional configurations that could be added, we believe the current training set for tungsten covers most of the bulk behavior (elastic deformations, equation of state, vacancies) as well as high energy configurations that would result from radiation damage(dislocations, interstitials, surfaces). In total, there are 9897 individual atomic configurations in the pure tungsten set, with over 10^6 force data points.

The beryllium training set has a very similar composition to that of tungsten, since the goal is to create a *general use* potential that is also tailored to simulate plasma facing materials. Equilibrium bulk properties of beryllium are captured through the Elastic Deformation, Equation of State and *ab initio* MD training groups. Together these groups contribute approximately 95% and 47% of the total energy and force training points, respectively. Conversely, the defect properties and lower symmetry environments of beryllium are collected in the Surfaces, Self-interstitials, Stacking Fault and Liquid groups. While fewer in number of configurations, these large atom count training structures contribute the majority of the force data points. All of the beryllium training data was also generated using VASP with the same simulation parameters as the tungsten data, with the chosen pseudopotential leaving just the outermost s-orbital electrons to the basis set.

Lastly, a set of training data was generated that focused on ordered inter-metallic phases of W-Be ranging in composition from equiatomic to WBe_{12}. In addition, multiple crystal structures of these proposed inter-metallic compounds were used in these calculations. For all training groups except Surface Adhesion, six different phases of W-Be were considered: B_2 (WBe), L1_2 (WBe_3), C_{14} (WBe_2), C_{15} (WBe_2), C_{36} (WBe_3), and D_2B (WBe_{12}). Surface Adhesion is a special training group that is strongly aligned with the target application of high energy Be implantation onto a W surface. This set of configurations included the binding of a single Be atom adsorbed onto (100) and (111) tungsten surfaces as well as multiple Be atoms adsorbed onto the same surface orientations.

The remaining columns in Table I, σ_E and σ_F, are the optimal training weights selected for the energies and forces in each training group. These group weights are scaled by the number of data points in the group, so they indicate the relative importance assigned to each group in the optimization process. The bolded values indicate the largest weight in each column, which can be interpreted as the *most important* type of training data for fitting the full W-Be SNAP potential. The details of this optimization process and how these optimal training weights were obtained will be discussed in the following section.

B. Optimization Methodology

Once a training set has been constructed, the goal of fitting a SNAP potential is to strike a balance between accurate reproduction of the training data (interpolated properties) and ability to describe structures that are too large to calculate using DFT (extrapolated properties). The simplest, and most common, interpolation error that can be optimized is the re-

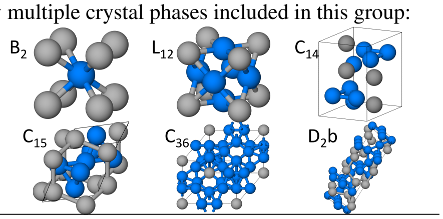
Description	N_E	N_F	σ_E	σ_F	Description	N_E	N_F	σ_E	σ_F	Description	N_E	N_F	σ_E	σ_F
W:					Be:					W-Be:				
Elastic Deform	2000	6000	$5 \cdot 10^0$	$6 \cdot 10^4$	Elastic Deform	4594	43260	$1 \cdot 10^5$	$\mathbf{1 \cdot 10^7}$	Elastic Deform†	3946	68040	$\mathbf{3 \cdot 10^5}$	$2 \cdot 10^3$
Equation of State	125	3468	$1 \cdot 10^{-1}$	$6 \cdot 10^4$	Equation of State	502	5418	$6 \cdot 10^4$	$3 \cdot 10^6$	Equation of State†	1113	39627	$2 \cdot 10^5$	$4 \cdot 10^4$
DFT-MD	60	23040	$3 \cdot 10^0$	$1 \cdot 10^4$	DFT-MD	909	130896	$\mathbf{2 \cdot 10^5}$	$2 \cdot 10^6$	DFT-MD†	3360	497124	$7 \cdot 10^4$	$6 \cdot 10^2$
Surfaces	180	334818	$\mathbf{1 \cdot 10^5}$	$3 \cdot 10^5$	Surfaces	90	17280	$1 \cdot 10^3$	$4 \cdot 10^5$	Surface Adhesion	381	112527	$2 \cdot 10^4$	$\mathbf{9 \cdot 10^4}$
Self-Interstitials	15	5805	$5 \cdot 10^{-2}$	$8 \cdot 10^2$	Self-Interstitials	179	137931	$3 \cdot 10^2$	$4 \cdot 10^5$	† multiple crystal phases included in this group:				
Liquids	27	3120	$4 \cdot 10^{-3}$	$3 \cdot 10^2$	Liquids	75	57600	$7 \cdot 10^1$	$7 \cdot 10^5$	 				
Dislocations	98	39690	$3 \cdot 10^0$	$9 \cdot 10^4$	Stacking Faults	6	864	$3 \cdot 10^0$	$2 \cdot 10^6$					
Monovacancy	420	183054	$2 \cdot 10^3$	$1 \cdot 10^5$										
Divacancy	39	6084	$1 \cdot 10^0$	$1 \cdot 10^3$										
Γ-Surface	6183	328338	$1 \cdot 10^0$	$1 \cdot 10^6$										
Γ-Surf.+Vacancy	750	105750	$4 \cdot 10^{-1}$	$\mathbf{3 \cdot 10^6}$										
Total	9897	1039167				6355	393249				8800	717318		

Table I. Training data used in the full W-Be SNAP fit, broken down by element type and group within each element. Each of the groups in the W-Be category contain configurations for multiple inter-metallic compounds, some of which are displayed in the inset. For each group the number of energy (N_E) and force (N_F) training points are given as well as the optimal training weight (σ_E, σ_F) selected for the full W-Be SNAP potential.

gression error. Equation 9 captures the general form of linear regression used here. $\hat{\boldsymbol{\beta}}$ minimizes the difference between the descriptor (D, bispectrum representation) prediction and reference (T, electronic structure) data. A regularization penalty of order n with weight γ_n can be applied to constrain the $\hat{\boldsymbol{\beta}}$ solution. Solutions with $n = 1$ enforce sparsity in the $\hat{\boldsymbol{\beta}}$ solution, while Tikhonov regularization[66] with $n = 2$ penalize against large values of $\hat{\boldsymbol{\beta}}$ which are hallmarks of an overfit solution. We have observed no improvement in overall accuracy when enforcing sparsity, and there is little risk of overfitting, because the number of bispectrum descriptors is far less than the number of training points($\mathcal{O}[10^6]$).

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(\|\boldsymbol{\epsilon} \circ (D\boldsymbol{\beta} - T)\|^2 - \gamma_n \|\boldsymbol{\beta}\|^n) \qquad (9)$$

Therefore, we solve Equation 9 with no regularization penalty, corresponding to the weighted linear least squares solution.

When fitting a SNAP potential, there are two different categories of fitting variables that are controlled by the optimizer. The first are called hyper-parameters and will directly modify the bispectrum components themselves. Examples of these are the radial cutoff ($R_{ii'}$), element densities ($w_{i'}$), and cutoff scale factor (α) of equations 2 and 3, respectively. A second set of fitting variables are the aforementioned group weights, $\boldsymbol{\epsilon}$, that scale each component of target space (T) in equation 9. There are far fewer hyper-parameters than group weights with the latter being as numerous as the user sees necessary to divide up the full training set into unique groups. In order to limit the number of free variables, we have chosen to optimize the hyperparameters and group weights for each element separately before tackling the mixed element training data.

DAKOTA[67] is used as the optimizer utilizing a single objective genetic algorithm (GA). Figure 2 visually displays the overall fitting procedure. Central to the overall fitting process is FitSNAP.py, which couples DAKOTA, LAMMPS, and the database of DFT training data. Following one pass through this optimization loop, a set of fitting parameters is provided from DAKOTA to FitSNAP, new bispectrum components are calculated by taking the coordinate information from the reference data and sending it to LAMMPS. Once all training configurations are converted into their respective bispectrum components, which forms D of Eq. 9, the energy and forces are parsed from the reference data to populate T. Solving for $\hat{\boldsymbol{\beta}}$, the linear regression is done using singular value decomposition and the energy and force errors (interpolation error) are reported back to DAKOTA as part of the objective function. At this point the candidate potential is used to run short MD simulations to evaluate material properties of interest. For example, while fitting the tungsten data the elastic constants and a few defect formation energies are calculated for each candidate and their percent error with respect to DFT is communicated back to DAKOTA. An equally weighted contribution from each of these material properties plus the regression errors is used to form the objective function for GA optimization.

As was mentioned previously, the optimization of the hyper-parameters and group weights were done separately for each element type, this is done to limit the number of free fitting variables. For a single element SNAP fit, the only hyper-parameter is the radial cutoff term since element density and ($w_{i'}$) and cutoff scale factor (α) are only needed to be modified to differentiate between element species. Optimization of $R_{ii'}$ was carried out by sampling values between $2.0\mathring{A}$ and $10.0\mathring{A}$ using a GA where regression errors and a subset of the full set of material properties (details on these in the next section) were used to determine the optimal radial cutoff. With this first step done, the conversion of the training data to bispectrum components (D in equation 9) can be precomputed for each candidate in the optimization loop(Figure 2) which significantly improves the throughput of the overall fitting process. Secondly, the training group weights for either pure element training are now optimized with this fixed $R_{ii'}$. Each generation of the GA fit consists of three-hundred candidate potentials and we observed minimal improvement

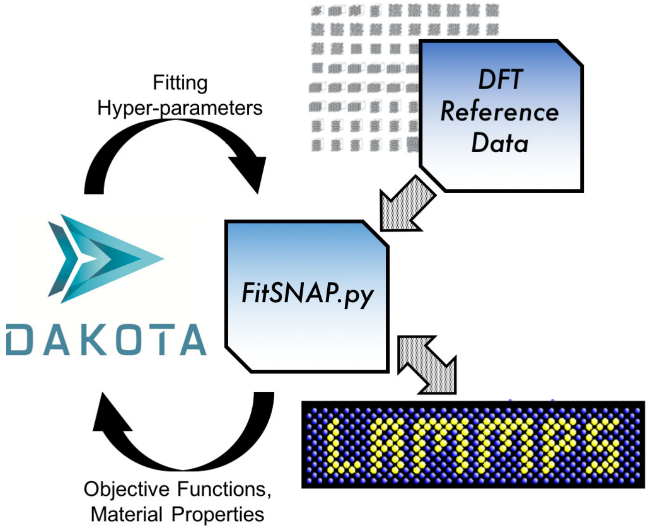


Figure 2. Workflow of fitting SNAP potentials. Each of the software tools, DAKOTA, LAMMPS and FitSNAP.py are developed at Sandia National Labs. The reference data can be generated with any code, we have used VASP in the current work.

of the overall fitness after seventy generations. One advantage of using a GA is that evaluations of candidate potentials (single pass through Figure 2) can be done simultaneously. As a result, the throughput of the overall fitting process can be distributed across a large super computer. Asynchronous evaluation tiling is used to distribute each evaluation to separate compute nodes. Typical jobs used fifty Intel Broadwell nodes simultaneously to generate and evaluate candidate potentials.

In order to parameterize the binary system, new hyper-parameters are introduced and thus need to be optimized before group weight optimization of the binary training data. Optimized groups weights from pure element fits are carried over, but a new $R_{ii'}$ (common to both elements), $w_{i'}$ and α (only Be terms modified) are needed. During this hyper-parameter optimization the regression errors(energy and forces), error in BCC-W, HCP-Be elastic constants as well as error in formation energies of ground state crystal structures for either pure phase were used to find optimal solutions. The final optimized hyper-parameters for the binary system were $R_{ii'} = 4.812$Å, $w_{Be} = 0.959$,$w_W = 1.0$, $\alpha_{Be} = 0.836$ and $\alpha_W = 1.0$.

C. Interpolated and Fitted Properties

The distribution of regression errors of the resultant best fit candidate are displayed in Figure 3. The three data series here represent each of the pure phase fits (W Fit and Be Fit, respectively) followed by the optimal fit to the entire W+Be training set. Due to the increased training set volume and the choice to use linear regression for SNAP potentials, the full W-Be fit has higher average errors, indicated by the dashed vertical lines, than either of the pure component fits. However, the fraction of the training data with error below the average is well above 50%, which indicates that the average errors are dominated by a relatively small number of outlier configurations that have exceptionally large energy or force errors. The average interpolation errors for the fit to the entire W-Be training set are 0.12 eV/atom and 0.31 eV/Å, respectively.

A detailed breakdown of the energy and force regression errors per group of training data is provided in Table II. For each of the energy and force columns, the highest errors are denoted as bold text while the lowest regression errors are underlined. It is interesting to note that within the pure W and pure Be training sets, the highest errors are reported in high energy, low coordination atomic environments which are the Liquids or Self-Interstitial training groups. Other than these best and worst fit training groups, the remainder of the training errors are relatively close to one another. Confirming what is shown in Figure 3, the average regression errors are lower in the pure W training data than Be or the combined W+Be set.

In addition to these interpolation errors, each candidate potential is used in a set of short MD simulations to determine its accuracy for material properties of interest. The reference values for these properties are taken from DFT and a percent error is reported back to DAKOTA as part of the optimization. The relative errors of these predictions are included in the objective function for hyperparameter and group weight optimization. For tungsten these properties are elastic constants, lattice parameter, cohesive energy, and the relaxed formation energies of six point defects in the BCC phase. Similarly for beryllium the HCP elastic constants, six point defect formation energies, and cohesive energies of five simple crystal structures are used as fitting objectives. These fitted material properties are displayed in Figure 4. The left and right panels show the percent errors with respect to the DFT predictions for the pure-W and pure-Be material properties, respectively. Intermediate, single element optimized potentials(denoted as W-SNAP and Be-SNAP) are shown in Figure 4 in addition to the final binary potential (WBe-SNAP). For the tungsten properties, the average percent error of C_{11},C_{12} and C_{44} is reported as "Elast. Const." along with the percent errors in the formation energy of four self-interstitial defects at the Octahedral site, Tetrahedral site, [110] and [111] oriented dumbbell defects. Lastly, the percent error in the formation energy of a single vacancy and divacancy(nearest neighbor positions) binding energy are also used as fitting objectives. The small divacancy binding energy (0.12eV) results in large percent errors for even small deviations and since all of these material properties are equally weighted during optimization we see this property as an indicator of overall fitness. This does have unwanted side effects though, as seen in the large percent error in the formation energy of a single vacancy.

Regarding the beryllium fitted properties, the average percent error in the bulk modulus, three shear moduli and a modulus corresponding to basal expansion under c-axis compression is reported as the "Avg. Moduli." Percent errors in the HCP, FCC and BCC cohesive energies are averaged and reported alongside formation energy error in five self-interstitial defects; Basal-octahedral, Octahedral, Basal-Split, Crowdion and Tetrahedral sites. Lastly, the formation energy error of a

Description	$\Delta E(eV/atom)$	$\Delta F(eV/\AA)$	Description	$\Delta E(eV/atom)$	$\Delta F(eV/\AA)$	Description	$\Delta E(eV/atom)$	$\Delta F(eV/\AA)$
W:			Be:			W-Be:		
Elastic Deform	$5.3 \cdot 10^{-2}$	$0.0 \cdot 10^{0}$	Elastic Deform	$7.6 \cdot 10^{-2}$	$2.3 \cdot 10^{-3}$	Elastic Deform†	$9.2 \cdot 10^{-2}$	$\underline{1.7 \cdot 10^{-1}}$
Equation of State	$1.4 \cdot 10^{-1}$	$\underline{4.0 \cdot 10^{-5}}$	Equation of State	$9.8 \cdot 10^{-2}$	$\underline{8.8 \cdot 10^{-4}}$	Equation of State†	$\mathbf{1.1 \cdot 10^{0}}$	$\mathbf{6.3 \cdot 10^{-1}}$
DFT-MD	$5.3 \cdot 10^{-2}$	$6.0 \cdot 10^{-2}$	DFT-MD	$3.6 \cdot 10^{-2}$	$8.2 \cdot 10^{-2}$	DFT-MD†	$\underline{7.9 \cdot 10^{-2}}$	$5.7 \cdot 10^{-1}$
Surfaces	$3.4 \cdot 10^{-2}$	$2.8 \cdot 10^{-1}$	Surfaces	$\underline{1.8 \cdot 10^{-2}}$	$5.0 \cdot 10^{-2}$	Surface Adhesion	$8.6 \cdot 10^{-2}$	$4.7 \cdot 10^{-1}$
Self-Interstitials	$4.6 \cdot 10^{-2}$	$9.5 \cdot 10^{-2}$	Self-Interstitials	$\mathbf{1.3 \cdot 10^{0}}$	$8.2 \cdot 10^{-2}$	† Multiple crystal phases included in this group		
Liquids	$\mathbf{2.9 \cdot 10^{-1}}$	$\mathbf{4.8 \cdot 10^{-1}}$	Liquids	$6.5 \cdot 10^{-2}$	$\mathbf{2.6 \cdot 10^{-1}}$			
Dislocations	$5.0 \cdot 10^{-2}$	$7.8 \cdot 10^{-2}$	Stacking Faults	$2.0 \cdot 10^{-2}$	$2.0 \cdot 10^{-3}$			
Monovacancy	$4.2 \cdot 10^{-2}$	$9.8 \cdot 10^{-2}$						
Divacancy	$\underline{2.9 \cdot 10^{-2}}$	$8.7 \cdot 10^{-2}$						
Γ-Surface	$4.6 \cdot 10^{-2}$	$2.5 \cdot 10^{-1}$						
Γ-Surf.+Vacancy	$4.3 \cdot 10^{-2}$	$1.7 \cdot 10^{-1}$						

Table II. Regression errors for training data used in the full W-Be SNAP fit, broken down by element type and group within each element. Within each elemental set of training data values are bolded and underlined for the highest and lowest error values, respectively. Force errors for the tungsten elastic deformations are zero due to the fact a single atom unit cell was used in these DFT calculations and all forces are identically zero.

single vacancy is included in the list of fitted properties. In both cases, results for a comparable empirical potential are shown (EAM[32,68] for W and BOP[33] for Be). A schematic of these interstitial defects an the numerical values for fitted properties are available in the supplemental material[65].

One of the primary flaws of the W-EAM potential[68] was the prediction of an attractive divacancy binding energy at the nearest neighbor position, whereas DFT predicts[69] a mild (0.12 eV) repulsive energy for this defect configuration. It is believed that vacancy clustering is a key step in surface morphology changes when tungsten is used as a plasma facing component[70]. Therefore, the sign of the divacancy binding energy plays a critical role in surface evolution. All of the targeted material properties are within 10% of the DFT predictions, with the exception of the vacancy formation energy which has an error of 23% or -0.74eV w.r.t. DFT.

With respect to the beryllium properties, the current SNAP potential is a significant improvement on the existing bond order potential[33]. All of the cohesive energies and point defect properties for the present SNAP potential are again within 10% of DFT. The exceptions to these positive SNAP results are the HCP elastic moduli. Our W-Be SNAP potential predicts C_{13} to be -22 GPa whereas DFT predicts[71] a value of 17 GPa. This subsequently makes for a poor description of the shear moduli which captures the basal expansion under compression along the c axis.

No additional fitting objectives were added for the binary system. Point defects of dissimilar element species were left as measures of extrapolation accuracy that will be discussed in the following section. The new SNAP W-Be potential has been added to the public distribution of the LAMMPS software package[51], and the training data is available upon request from the authors.

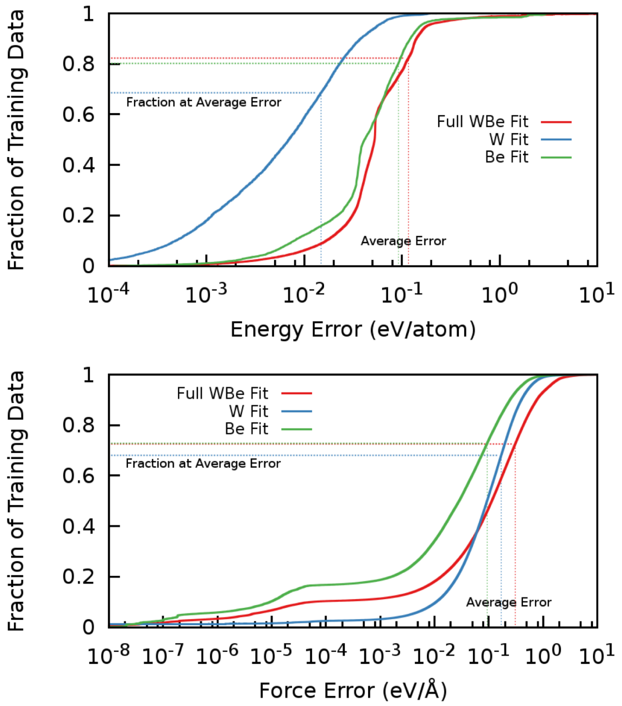


Figure 3. Distribution of the regression errors observed for the best fit candidates for each of the pure-W (blue), pure-Be (green) and full W-Be (red) training sets. **(Top)** Energy errors **(Bottom)** Force errors. In all cases, vertical dashed lines indicate the average regression error and horizontal dashed lines indicate the fraction of the training data with error lower than the average.

IV. BERYLLIUM IMPLANTATION RESULTS

To test the quality of the potential outside of the data included in the training set, molecular dynamics simulations of single beryllium implantations in tungsten were performed.

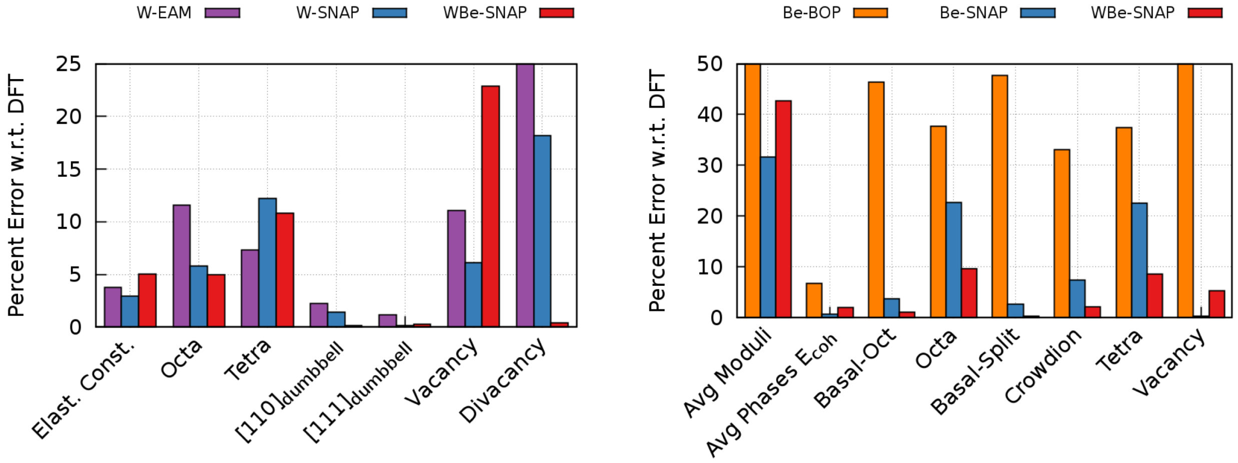


Figure 4. **(Left)** Tungsten material property predictions included in the SNAP optimization loop represented as percent error to the DFT prediction. Due to the small absolute value of the divacancy binding energy (0.12eV) the percent error for EAM is much larger than either SNAP potential displayed here. **(Right)** Beryllium material property predictions that are also included in the optimization procedure. Both SNAP potentials displayed here are significant improvements over the existing BOP predictions of self-interstitial formation energies.

Quantifying the implantation depth and lattice interaction of beryllium in tungsten will determine future diffusion and damage mechanisms that will affect overall tungsten diverter performance.

Simulations were performed using the LAMMPS[50] molecular dynamics package and the SNAP potential described in this work. The simulation cell consisted of a 3 nm x 3 nm x 9 nm tungsten slab with 3 nm of void space above the surface. Periodic boundary conditions were used in the x, [100] and y, [010] directions while a free surface boundary condition was used in the z, [001] direction. The tungsten was first equilibrated to a temperature of 1000 K by giving the atoms a velocity based on the Maxwell Boltzmann distribution. Dynamics were run with an NVE thermostat and a 1 fs timestep for 20 ps where velocity rescaling was performed for the first 5 ps and then turned off for the last 15 ps. After equilibration, a beryllium atom was placed 10 Å above the surface with random x and y coordinates. The beryllium atom was then given an energy of 75 eV in the z direction directly towards the surface and dynamics were performed with an NVE thermostat. During the implantation, a variable timestep was required to conserve energy due to the initially high beryllium velocity. The timestep was allowed to vary between 10^{-4} fs and 0.5 fs and was updated every 10 timesteps so that no atom moved more than 0.02 Å per time step. It was necessary to freeze the bottom two layers of atoms by setting their forces to zero to prevent the unwanted movement of the slab. The simulation was allowed to evolve for 3 ps and the beryllium location in the lattice was subsequently recorded unless it reflected from the surface. A total of 5,000 individual simulations were performed. A similar calculation for 75 eV beryllium implantation in tungsten was run in the Stopping and Range of Ions in Matter (SRIM) program[72] for $1 \cdot 10^{6}$ atoms for comparison. SRIM is a widely used binary collision approximation code that models the interaction of ions in matter and the outputs include the final distribution of ions in the target material as well as ion effects in the target such as sputtering, material damage, and ionization. Unlike MD, SRIM includes the electronic stopping of the ion. However, the target material is represented by a mean density and is assumed to be amorphous. Therefore, crystal dependent effects like channeling are neglected.

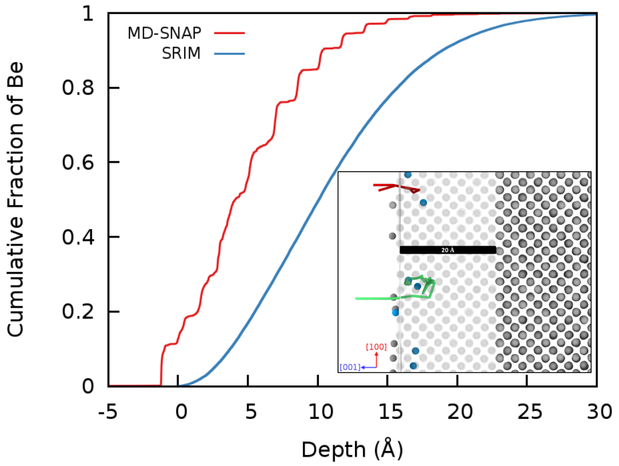


Figure 5. Plot of the cumulative depth distribution of 75 eV Be in tungsten at 1000 K using both MD (red) and SRIM (blue). Inset displays an atomistic snapshot of the Be (blue spheres) implanted onto the (001) surface of W (grey spheres). The red/green trajectory lines show the time history of a rejected/captured Be atom, respectively.

Of the total beryllium implantations performed with the newly generated SNAP potential, 35% implanted in the lattice while the other 65% reflected. A plot of the beryllium depth distributions for SNAP and SRIM[73] are shown in Figure 5 in red and blue respectively. The SNAP potential predicts the beryllium atoms to remain within 20 Å of the surface after implantation with about 12% of the beryllium atoms residing above the original surface, indicating a preference for the beryllium to be near the surface. While the SRIM profile is comparable to SNAP, SRIM predicts a slightly deeper depth profile and a lower reflection rate. Nevertheless, both distributions indicate that implanted beryllium remains near the surface after implantation. The distinct stepped profile produced by SNAP reflects the tendency of the beryllium atoms occupy particular interstitial sites within the tungsten matrix. SRIM does not capture this effect, since the material is modeled as a homogeneous isotropic material with no crystalline structure. While SRIM also includes electronic stopping and MD does not, the depth profiles is still more shallow for the SNAP potential. This indicates the importance of atomic collisions in the beryllium implantation process. Overall the two distributions are fairly consistent, both predict Be to implant within 3 nm of the surface , the slopes of the curves are comparable, and the curves are offset by about 5 angstroms at most. This is consistent with previous works investigating implantations in tungsten[74,75]. Many of the discrepancies observed are due to the different assumptions, namely the lack of crystal orientation in SRIM and surface effects, leading to the bumps and the initial jump in the cumulative fraction at the surface for the MD curve.

The inset in Figure 5 depicts the beryllium trajectory for a few different individual implantations. The red line traces the history of a beryllium atom that entered the lattice but subsequently escaped while the green line traces a beryllium atom that implanted. Captured beryllium diffuses rapidly during the brief thermalization process, as indicated by the jagged trajectory line, and eventually becomes trapped just under the surface. The impacting beryllium atoms interact with the tungsten lattice in a variety of ways including displacing a tungsten atom and subsequently occupying the vacant site, creating tungsten adatoms (see inset image of Figure 5), creating W-Be dumbbells, and sputtering tungsten atoms with a low sputtering yield of 0.006 W/Be. A breakdown of the relative contributions of SNAP versus ZBL interactions is given for a simple two-atom collision in the supplemental material[65].

Initial observations of the simulations indicated that implanted beryllium atoms typically resided in interstitial sites, substitutional sites, surface sites, or as $\langle 111\rangle$ or $\langle 110\rangle$ oriented W-Be dumbbells. For the case of $\langle 111\rangle$ W-Be dumbbell formation, the configuration is more like a series of oriented displacements in the $\langle 111\rangle$ direction with a beryllium at the center and typically two displaced tungsten atoms. Beryllium that substitutes a tungsten atom on the lattice results in the displaced tungsten atom typically residing on the surface as an adatom. All of the 12% of the beryllium atoms above the surface in the depth profile were identified to be at hollow sites. The number of implanted beryllium atoms at each site was quantified by extracting the lattice position and is listed in Table III. Overall the beryllium atoms preferred the $\langle 111\rangle$ dumbbell, followed by the substitutional site and the hollow site on the surface.

To determine how realistic the rate of occurrence of these beryllium interstitials are, a series of new DFT calculations has been performed to assess these defect formation energies.

	Implanted	Formation Energy (eV)		
Defect Type	Be Percent	DFT	SNAP	BOP[33]
[111] Dumbbell	41.2	4.30	3.66	0.67
Substitution	22.2	3.11	3.29	-2.00
Surf. Hollow Site	12.3	-1.05	-1.39	-3.52
Tetrahedral Inter.	10.4	4.13	4.20	-0.28
[110] Dumbbell	8.4	4.86	4.29	-0.03
Octahedral Inter.	5.3	3.00	5.11	0.34
Surf. Bridge Site	0.03	1.01	0.44	-1.30

Table III. Defect formation statistics for single, 75 eV, Be implantation onto a (001) surface of BCC tungsten with a comparison of formation energies for these Be interstitials in the W matrix. While SNAP was only trained for self-interstitial energies for either element type, its prediction of these multi-element defects are much closer to DFT than the empirical BOP potential.[33]

It is important to note that these formation energies were not included in the training data and are therefore a good test of how well the potential can truly predict properties relevant for this particular application. Values of the defect formation energies calculated using DFT and SNAP, as well as the existing BOP potential for comparison, are shown in Table III. The SNAP potential performs very well for most cases, with the exception of the octahedral formation energy and the $\langle 111\rangle$ dumbbell. Nevertheless, the new SNAP potential predicts formation energies much closer to DFT values than BOP. Furthermore, SNAP predicts the three lowest formation energies to be the surface hollow site, the substitutional site, and the $\langle 111\rangle$ dumbbell. These three defects are also the most frequently observed defects in the implantation simulations, indicating a general consistency between the MD results and what is energetically expected from the DFT defect formation energy calculations. While both SNAP and DFT predict the surface hollow site to have the lowest formation energy, the $\langle 111\rangle$ dumbbell as well as the substitutional defect are observed more frequently. This is a kinetic effect of the 75 eV implantation energy. Beryllium atoms that are not immediately reflected are more likely to be trapped in sub-surface defect sites than to bind at the surface hollow site.

V. CONNECTION TO MULTI-SCALE COMPUTATIONAL EFFORTS

These initial MD simulation results provide a first evaluation of the implanted beryllium profile, as well as identifying the initial fate of the beryllium once in the lattice. This advanced ML-IAP enables larger MD simulations that can be used to investigate longer time scale ($\mathcal{O}[10^{-1} - 10^{1}]\mu s$) evolution of the tungsten surface subjected to beryllium implantation. These simulations will reveal important physics related to the timescale associated with W-Be intermetallic phase formation, as well as local defect configurations that may serve as trapping sites for implanted hydrogen or helium atoms. Large-scale MD simulations can also provide important computational data for benchmarking longer time mesoscale or continuum simulation techniques.

The plasma-surface interactions (PSIs) occurring in the diverter and plasma facing components (PFCs) pose a critical scientific challenge that limits our ability to operate fusion machines by sustaining a steady-state burning plasma. The simulation paradigm of multiscale computational modeling relies on a parameter-passing framework in which the entire spatial and temporal domains are sub-partitioned into different regimes on the basis of the characteristic length and time scale of the physical phenomena involved. Such multiscale models attack the complex materials degradation issues from both a *bottom-up* atomistic-based approach simultaneously with a *top-down* continuum perspective, and focus on the hierarchical integration of kinetic processes of species reactions and diffusion to model microstructure evolution over experimental timescales. The simultaneous use of both an atomistic and continuum approach has furthered the development of scale-bridging or multi-scale integration, and has led to fundamental insight into helium-hydrogen synergies controlling PSI in tungsten, as well as the long-term microstructural evolution due to radiation damage in structural materials.[76]

First-principles, density functional theory (DFT) electronic structure methods as implemented in commercial and open-source codes[77–79] can be instrumental in providing interaction forces, basic thermodynamic and kinetic interactions and rates, which can be used in fitting interatomic potentials for molecular dynamics simulations, and are utilized where existing interatomic potentials are deemed inadequate. Unfortunately the limitation of such first principles methods relate to the lack of thermal fluctuations in DFT calculations of thermodynamics and migration barriers, as well as the very short timescales ($\mathcal{O}[10^2]$ ps) available for dynamic DFT-MD simulations. Moving past the size and time limitations of DFT, large-scale MD simulations can provide an extension to the *bottom up* multiscale modeling paradigm. MD simulations are only as accurate as the interatomic potentials, but can provide important physical insights on the dynamics of defect interactions, provided that such interaction dynamics occur on rapid, nanosecond timescales.

Furthermore, MD simulations can provide a computational database capable of benchmarking mesoscale or continuum scale models, as well as identifying key physical mechanisms that must be included in longer-time simulation techniques. The emerging multiscale modeling capabilities are very much in the early stages of development, and continued research activities are required to further develop this capability. In particular, the questions around mixed material formation including the timescale on which intermetallic phase separation occurs, how such phases and localized chemically complex defect arrangements influence hydrogen retention and permeation, require atomistic insight. These initial MD simulations, and the improvements in modeling chemically complex plasma exposed surfaces using the SNAP interatomic potentials, provide a key opportunity to investigate such complex and important PSI challenges.

VI. CONCLUSIONS

At the intersection of data-science and atomistic simulation of materials, the presented ML-IAP demonstrates the significant improvement over empirical IAP that can be provided by SNAP. This new SNAP W-Be potential improves upon the existing BOP for key material properties that are necessary for studying PFCs and ultimately this accuracy and scalability improvement will become a key component of multiscale simulation of PFCs. The results of the Be implantation simulations discussed here indicate a preference for surface adhesion and shallow depth profiles into tungsten. This SNAP W-Be potential will allow for further simulations targeting W-Be plasma material interactions, filling a critical need in the area of fusion energy research. The results presented here show consistency with DFT for important defect properties relevant to Be implantation in W. How these implantation defects affect helium and hydrogen trapping from the plasma, as well as long timescale dynamics of Be at W surfaces is the focus of future work. Lastly, the fitting methodology outlined here can be safely applied to any condensed phase system given suitable training data, though additional study is needed to validate the bispectrum as a physical descriptor of gaseous and molecular bonding environments.

ACKNOWLEDGMENTS

All authors gratefully acknowledge funding support is from the plasma surface interaction project of the Scientific Discovery through Advanced Computing (SciDAC) program, which is jointly sponsored by the Fusion Energy Sciences (FES) and the Advanced Scientific Computing Research (ASCR) programs within the U.S. Department of Energy Office of Science. Equal support of this work is from the Exascale Computing Project (No. 17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

[1] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, Jom **65**, 1501 (2013).

[2] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, Appl. Materials **1**, 011002 (2013).

[3] H. L. Zhuang and R. G. Hennig, The Journal of Physical Chemistry C **117**, 20440 (2013).

[4] P. Villars, H. Okamoto, and K. Cenzual, ASM International, Materials Park, OH, USA (2006).

[5] C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, The Journal of Physical Chemistry C **122**, 17575 (2018).

[6] M. S. Daw, S. M. Foiles, and M. I. Baskes, Materials Science Reports **9**, 251 (1993).

[7] M. Baskes, Physical review B **46**, 2727 (1992).

[8] S. B. Sinnott and D. W. Brenner, Mrs Bulletin **37**, 469 (2012).

[9] T. Liang, Y. K. Shin, Y.-T. Cheng, D. E. Yilmaz, K. G. Vishnu, O. Verners, C. Zou, S. R. Phillpot, S. B. Sinnott, and A. C. Van Duin, Annual Review of Materials Research **43**, 109 (2013).

[10] J. Tersoff, Physical Review B **38**, 9902 (1988).

[11] J. E. Lennard-Jones, Proceedings of the Physical Society **43**, 461 (1931).

[12] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, The Journal of Physical Chemistry C **121**, 511 (2016).

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

[14] M. Rupp, O. A. von Lilienfeld, and K. Burke, The Journal of Chemical Physics **148** (2018).

[15] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, The Journal of Chemical Physics **148**, 241722 (2018).

[16] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, Physical Review Materials **1**, 043603 (2017).

[17] A. P. Bartok, J. Kermode, N. Bernstein, and G. Csanyi, arXiv preprint arXiv:1805.01568 (2018).

[18] M. A. Wood and A. P. Thompson, The Journal of Chemical Physics **148**, 241721 (2018).

[19] S. J. Plimpton and A. P. Thompson, MRS bulletin **37**, 513 (2012).

[20] Z. Deng, C. Chen, X.-G. Li, and S. P. Ong, arXiv preprint arXiv:1901.08749 (2019).

[21] X.-G. Li, C. Hu, C. Chen, Z. Deng, J. Luo, and S. P. Ong, Physical Review B **98**, 094104 (2018).

[22] G. Federici, P. Andrew, P. Barabaschi, J. Brooks, R. Doerner, A. Geier, A. Herrmann, G. Janeschitz, K. Krieger, A. Kukushkin, *et al.*, Journal of Nuclear Materials **313**, 11 (2003).

[23] S. Brezinsek, J.-E. contributors, *et al.*, Journal of nuclear materials **463**, 11 (2015).

[24] G. Federici, C. H. Skinner, J. N. Brooks, J. P. Coad, C. Grisolia, A. A. Haasz, A. Hassanein, V. Philipps, C. S. Pitcher, J. Roth, *et al.*, Nuclear Fusion **41**, 1967 (2001).

[25] S. Brezinsek, T. Loarer, V. Philipps, H. Esser, S. Grünhagen, R. Smith, R. Felton, J. Banks, P. Belo, A. Boboc, *et al.*, Nuclear Fusion **53**, 083023 (2013).

[26] S. Brezinsek, S. Jachmich, M. Stamp, A. Meigs, J. Coenen, K. Krieger, C. Giroud, M. Groth, V. Philipps, S. Grünhagen, *et al.*, Journal of Nuclear Materials **438**, S303 (2013).

[27] H. Okamoto, L. Tanner, S. N. Naidu, and P. R. Rao, Indian Institute of Metals, Calcutta (1991).

[28] R. Doerner, M. Baldwin, and R. Causey, Journal of nuclear materials **342**, 63 (2005).

[29] R. Doerner, Journal of nuclear materials **363**, 32 (2007).

[30] M. Baldwin, R. Doerner, D. Nishijima, D. Buchenauer, W. Clift, R. Causey, and K. Schmid, Journal of nuclear materials **363**, 1179 (2007).

[31] C. Linsmeier, K. Ertl, J. Roth, A. Wiltner, K. Schmid, F. Kost, S. Bhattacharyya, M. Baldwin, and R. Doerner, Journal of nuclear materials **363**, 1129 (2007).

[32] M. A. Cusentino, K. D. Hammond, F. Sefta, N. Juslin, and B. D. Wirth, Journal of Nuclear Materials **463**, 347 (2015).

[33] C. Björkas, K. Henriksson, M. Probst, and K. Nordlund, Journal of Physics: Condensed Matter **22**, 352206 (2010).

[34] A. Lasa, K. Heinola, and K. Nordlund, Nuclear Fusion **54**, 083001 (2014).

[35] A. Lasa, K. Heinola, and K. Nordlund, Nuclear Fusion **54**, 123021 (2014).

[36] S. Plimpton, Computational Materials Science **4**, 361 (1995).

[37] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, npj Computational Materials **3**, 54 (2017).

[38] J. Behler and M. Parrinello, Physical review letters **98**, 146401 (2007).

[39] J. Behler, The Journal of chemical physics **145**, 170901 (2016).

[40] N. Lubbers, J. S. Smith, and K. Barros, The Journal of Chemical Physics **148**, 241715 (2018).

[41] H. Wang, L. Zhang, J. Han, and E. Weinan, Computer Physics Communications **228**, 178 (2018).

[42] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Physical review letters **104**, 136403 (2010).

[43] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, NPJ Computational Materials **3**, 37 (2017).

[44] V. Botu, J. Chapman, and R. Ramprasad, Computational Materials Science **129**, 332 (2017).

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

[46] R. Drautz, Physical Review B **99**, 014104 (2019).

[47] A. Seko, A. Togo, and I. Tanaka, arXiv preprint arXiv:1901.02118 (2019).

[48] J. Biersack and J. F. Ziegler, in *Ion implantation techniques* (Springer, Berlin, Heidelberg, 1982) pp. 122–156.

[49] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskii, *Quantum theory of angular momentum* (World Scientific, 1988).

[50] S. Plimpton, Journal of computational physics **117**, 1 (1995).

[51] LAMMPS, "LAMMPS molecular dynamics package," WWW site: `lammps.sandia.gov`.

[52] C. R. Trott, S. D. Hammond, and A. P. Thompson, in *International Supercomputing Conference* (Springer, 2014) pp. 19–34.

[53] T. I. Mattox, J. P. Larentzos, S. G. Moore, C. P. Stone, D. A. Ibanez, A. P. Thompson, M. Lísal, J. K. Brennan, and S. J. Plimpton, Molecular Physics **116**, 2061 (2018).

[54] G. Csányi, T. Albaret, M. Payne, and A. De Vita, Physical review letters **93**, 175503 (2004).

[55] Z. Li, J. R. Kermode, and A. De Vita, Physical review letters **114**, 096405 (2015).

[56] E. V. Podryabinkin and A. V. Shapeev, Computational Materials Science **140**, 171 (2017).

[57] W. J. Szlachta, A. P. Bartók, and G. Csányi, Physical Review B **90**, 104108 (2014).

[58] libAtoms.org, "libAtoms.org DFT data repository," WWW site: `http://www.libatoms.org/Home/TungstenTrainingConfigurations`.

[59] G. Kresse and J. Hafner, Physical Review B **47**, 558 (1993).

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

[61] G. Kresse and J. Furthmüller, Computational Materials Science **6**, 15 (1996).

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

[63] P. E. Blöchl, Physical review B **50**, 17953 (1994).

[64] G. Kresse and D. Joubert, Physical Review B **59**, 1758 (1999).

[65] See Supplemental Material at [URL will be inserted by publisher] for additional information regarding generation of the training set, MD calculations of defects, and other material properties predicted from the WBe-SNAP potential fit here. (2019).

[66] A. N. Tikhonov, A. Goncharsky, V. Stepanov, and A. G. Yagola, "Numerical methods for the solution of ill-posed problems," (2013).

[67] B. M. Adams, W. Bohnhoff, K. Dalbey, J. Eddy, M. Eldred, D. Gay, K. Haskell, P. D. Hough, and L. P. Swiler, Sandia Na-

tional Laboratories, Tech. Rep. SAND2010-2183 (2009).

[68] N. Juslin and B. Wirth, Journal of Nuclear Materials **432**, 61 (2013).

[69] P. M. Derlet, D. Nguyen-Manh, and S. Dudarev, Physical Review B **76**, 054107 (2007).

[70] P.-E. Lhuillier, T. Belhabib, P. Desgardin, B. Courtois, T. Sauvage, M.-F. Barthe, A.-L. Thomann, P. Brault, and Y. Tessier, Journal of Nuclear Materials **416**, 13 (2011).

[71] D. J. Silversmith and B. Averbach, Physical Review B **1**, 567 (1970).

[72] J. F. Ziegler, M. D. Ziegler, and J. P. Biersack, Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms **268**, 1818 (2010).

[73] J. F. Ziegler, M. D. Ziegler, and J. P. Biersack, *SRIM: the stopping and range of ions in matter* (Cadence Design Systems, 2008).

[74] K. D. Hammond and B. D. Wirth, Journal of applied physics **116**, 143301 (2014).

[75] V. Borovikov, A. F. Voter, and X.-Z. Tang, Journal of Nuclear Materials **447**, 254 (2014).

[76] J. Marian, C. S. Becquart, C. Domain, S. L. Dudarev, M. R. Gilbert, R. J. Kurtz, D. R. Mason, K. Nordlund, A. E. Sand, L. L. Snead, *et al.*, Nuclear Fusion **57**, 092008 (2017).

[77] G. Kresse and J. Hafner, Physical Review B **47**, 558 (1993).

[78] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, *et al.*, Journal of physics: Condensed matter **21**, 395502 (2009).

[79] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, Journal of Physics: Condensed Matter **14**, 2745 (2002).