Article

Explicit Multielement Extension of the Spectral Neighbor Analysis Potential for Chemically Complex Systems

Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

M. A. Cusentino, M. A. Wood, and A. P. Thompson*

Cite This: *J. Phys. Chem. A* 2020, 124, 5456–5464

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A natural extension of the descriptors used in the Spectral Neighbor Analysis Potential (SNAP) method is derived to treat atomic interactions in chemically complex systems. Atomic environment descriptors within SNAP are obtained from a basis function expansion of the weighted density of neighboring atoms. This new formulation instead partitions the neighbor density into partial densities for each chemical element, thus leading to explicit multielement descriptors. For N_{elem} chemical elements, the number of descriptors increases as $O(N_{elem}^3)$, while the computational cost of the force calculation as implemented in LAMMPS is limited to $O(N_{elem}^2)$ and the favorable linear scaling in the number of atoms is retained. We demonstrate these chemically aware descriptors by

producing an interatomic potential for indium phosphide capable of capturing high-energy defects that result from radiation damage cascades. This new explicit multielement SNAP method reproduces the relaxed defect formation energies with substantially greater accuracy than weighted-density SNAP, while retaining accurate representation of the bulk indium phosphide properties.
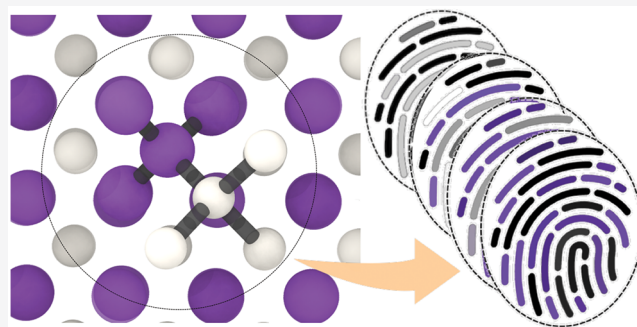
I. INTRODUCTION

Interatomic potentials (IAP) are an essential part of any classical molecular dynamics (MD) simulation and are also the leading approximation determining the physical accuracy of this method. The approximations that are inherent to all IAP (locality of forces, Born–Oppenheimer energy surface) are tolerated because this results in a computational cost that only scales linearly with the number of atoms, in contrast to cubic scaling of *ab initio* MD. Many IAP in use capture local interactions using functional forms that approximate known physical and chemical phenomena, such as covalent bonding,[1,2] electrostatic screening,[3] and electron density-mediated metallic bonding,[4,5] to name a few. In terms of developing new IAP, a decades long trend shows that much of the effort has focused on more accurate, but more computationally expensive potentials.[6]

A recent branch of this development incorporates advances in the field of data science wherein machine learning of IAP can provide an alternative to the aforementioned physics inspired potentials. Machine learned IAP (ML-IAP) take an alternate approach which is to forego a physically inspired model form in favor of a highly flexible functional based upon a generalized set of local atomic descriptors. These ML-IAP have an added requirement during parametrization relative to traditional IAP; that is, they need to be trained against a

database of energies and forces that usually come from a higher fidelity simulation, e.g., density functional theory (DFT). The contributions from each of the chosen local descriptors to the energy and forces are independently weighted in order to match a database of higher fidelity results. A variety of different descriptors to describe the local environment exist in the literature,[7] such as symmetry functions,[8,9] bispectrum components,[10] moment tensors,[11] and the Coulomb matrix.[12] At a minimum, a descriptor has to be invariant under translation, rotation, and permutation, of neighboring atoms. As is the case in many machine learning applications beyond IAP, the development and selection of the descriptor space used is of critical importance to the overall model performance. Significant effort in recent years has focused on developing atomic descriptors that can be used in, though not limited to, material property predictions,[13–18] alloy design,[19,20] and IAP used in MD.[7,8,10–12,21] Drautz has recently shown that many of these descriptors share a common mathematical foundation in
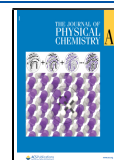
the atomic cluster expansion for the Born–Oppenheimer potential energy function.[22] A recent study comparing families of ML-IAP trained on shared training databases using different descriptors and training techniques showed broadly similar improvements in accuracy with increasing number of descriptors.[23] However, this comparative work by Zuo et al. was limited to parametrizations of single element systems. In this work we focus on the performance of the bispectrum components as atomic environment descriptors for multielement systems.

As is the case with all ML-IAP, SNAP reduction of regression errors (with respect to DFT training) can systematically be improved by including more descriptors, but Wood et al.[24] showed there are diminishing returns in accuracy even though greater computational cost is incurred in calculating these extra descriptors. These diminishing returns in accuracy hint that the underlying descriptor is incapable of capturing the full many-body nature of the energies and forces in the training set. In the original formulation of SNAP, the bispectrum descriptors only distinguish between neighbor atoms of different chemical elements based on their weighted contribution to the total atomic density.[24,25] This is similar in spirit to the construction of the density function within the embedded atom method for metal alloys.[26,27] For systems that show strong differences in bonding characteristics depending on the chemical identity of the atoms, this weighted-density (WD) approach is likely insufficient.

In this work we propose an explicit multielement (EME) SNAP descriptor formulation. In EME-SNAP, the descriptors are decomposed into separate contributions that depend on the partial densities of each chemical element in order to better express differences in the chemical makeup of the local atomic environment. In addition to the physical motivation to adapt these descriptors, the approach also adds more degrees of freedom to the model allowing for greater flexibility to fit the training data. A derivation of the EME descriptors and a discussion of their implementation into the LAMMPS MD software is provided in sections IIA and IIB. An example EME-SNAP potential and the associated training procedure is described in section IIC and a quantitative comparison of the accuracy of the WD-SNAP and EME-SNAP IAP is given in section III.

II. COMPUTATIONAL DETAILS

A. Explicit Multielement Bispectrum Descriptors. In the original weighted density (WD) SNAP formulation, the total density of neighbor atoms around a central atom i of element μ_i located at the origin is represented as a sum of δ-functions in a three-dimensional space:

$$\rho(\mathbf{r}) = w^{\text{self}}_{\mu_i}\delta(0) + \sum_{r_{ik} < R^{\mu_i\mu_k}_{\text{cut}}} f_{\text{c}}(r_{ik}; R^{\mu_i\mu_k}_{\text{cut}})w_{\mu_k}\delta(\mathbf{r}_{ik}) \tag{1}$$

where $\mathbf{r}_{ik}$ is the position of the neighbor atom k of element μ_k relative to the central atom i. The w_μ coefficients are dimensionless weights that are chosen to distinguish atoms of different chemical elements μ, while the central atom is assigned the weight $w^{\text{self}}_{\mu_i}$. The sum is over all atoms k within some cutoff distance $R^{\mu_i\mu_k}_{\text{cut}}$ that depends on the chemical identities of both the neighbor atom and the central atom. The switching function $f_{\text{c}}(r; R^{\mu_i\mu_k}_{cut})$ ensures that the contribution of each neighbor atom goes smoothly to zero at $R^{\mu_i\mu_k}_{\text{cut}}$. Following

Bartók et al.,[10] the radial distance r_{ik} is mapped to a third polar angle θ_0 defined by

$$\theta_0 = \theta_0^{\text{max}}\frac{r_{ik}}{R^{\mu_i\mu_k}_{\text{cut}}} \tag{2}$$

The additional angle θ_0 allows the set of points $\mathbf{r}_{ik}$ in the 3D ball of possible neighbor positions to be mapped on to the set of points (θ, ϕ, θ_0) on the unit 3-sphere. The neighbor density function can be expanded in the basis of 4D hyperspherical harmonic functions $\mathbf{U}^j$

$$\rho(\mathbf{r}) = \sum_{j=0,1/2,\dots}^{\infty} \mathbf{u}_j \cdot \mathbf{U}_j(\theta_0, \theta, \phi) \tag{3}$$

where $\mathbf{u}_j$ and $\mathbf{U}_j$ are rank $(2j + 1)$ complex square matrices. The · symbol indicates the scalar product of the two matrices. $\mathbf{u}_j$ are Fourier expansion coefficients given by the inner product of the neighbor density with the basis functions $\mathbf{U}_j$ of degree j. Because the neighbor density is a weighted sum of δ-functions, each expansion coefficient can be written as a sum over discrete values of the corresponding basis function,

$$\mathbf{u}_j = w^{\text{self}}_{\mu_i}\mathbf{U}_j(0) + \sum_{r_{ik} < R^{\mu_i\mu_k}_{\text{cut}}} f_{\text{c}}(r_{ik}; R^{\mu_i\mu_k}_{\text{cut}})w_{\mu_k}\mathbf{U}_j(\theta_0, \theta, \phi) \tag{4}$$

The expansion coefficients $\mathbf{u}_j$ are complex-valued and they are not directly useful as descriptors because they are not invariant under rotation of the polar coordinate frame. However, the following scalar triple products of expansion coefficients are real-valued and invariant under rotation:[10]

$$B_{j_1j_2j} = \frac{1}{2j + 1}\mathbf{u}_{j_1} \otimes_{j_1j_2j} \mathbf{u}_{j_2}\cdot(\mathbf{u}_j)^* \tag{5}$$

The symbol $\otimes_{j_1j_2j}$ indicates a Clebsch–Gordan product of matrices of degrees j_1 and j_2 that produces a matrix of degree j, as defined in our original formulation of SNAP.[21] The additional factor of $2j + 1$ renders $B_{j_1j_2j}$ invariant under permutation of the indices, which simplifies the calculation of gradients (see section IIB). These invariants are the components of the bispectrum. 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 bispectrum components defined here have been shown to be closely related to the four-body basis functions of the Atomic Cluster Expansion introduced by Drautz.[22]

In the WD-SNAP method, the potential energy of each atom is written as a linear or quadratic function of these geometric descriptors, as described below. This has proven to be an accurate and efficient method for constructing interatomic potentials for both single element and multielement systems. However, because neighbor atoms of different elements are distinguished solely by the magnitude of the factor $f_{\text{c}}(r_{ik}; R^{\mu_i\mu_k}_{\text{cut}})w_{\mu_k}$ that defines the effective weight of the contribution to $\mathbf{u}_j$, the WD-SNAP formulation does not strongly distinguish the chemical identities of neighbor atoms.

In order to achieve a more explicit representation of different chemical elements, a natural step is to partition the total neighbor density into partial densities for each element

$$\rho^\mu(\mathbf{r}) = w^{\text{self}}_{\mu_i\mu}\delta(0) + \sum_{r_{ik} < R^{\mu_i\mu_k}_{\text{cut}}} \delta_{\mu\mu_k}f_{\text{c}}(r_{ik}; R^{\mu_i\mu}_{\text{cut}})w_\mu\delta(\mathbf{r}_{ik}) \tag{6}$$

where $\delta_{\mu\mu_k}$ indicates that only neighbor atoms of element μ contribute to the partial density ρ^μ. The central atom of element μ_i contributes a partial self-weight $w^{\text{self}}_{\mu_i\mu}$ to ρ^μ. By requiring that the partial densities sum to the total density used in the original WD-SNAP formulation

$$\rho(\mathbf{r}) = \sum_{\mu=1}^{N_{\text{elem}}} \rho^\mu(\mathbf{r}) \tag{7}$$

it follows that $w^{\text{self}}_{\mu_i}$ in eq 1 is equal to the sum of the partial self-weights $w^{\text{self}}_{\mu_i\mu}$. In analogy with the WD-SNAP formulation, we set $w^{\text{self}}_{\mu_i\mu} = 1$ for all μ_i and μ.

The partial expansion coefficients for each element follow naturally from this definition

$$\mathbf{u}^\mu_j = w^{\text{self}}_{\mu_i\mu} \mathbf{U}_j(0) + \sum_{r_{ik} < R^{\mu_i\mu_k}_{\text{cut}}} \delta_{\mu\mu_k} f_c(r_{ik}; R^{\mu_i\mu_k}_{\text{cut}}) w_{\mu_k} \mathbf{U}_j(\theta_0, \theta, \phi) \tag{8}$$

The explicit multielement (EME) bispectrum can then be formed from products of the partial expansion coefficients

$$B^{\kappa\lambda\mu}_{j_1j_2j} = \frac{1}{2j+1} \mathbf{u}^\kappa_{j_1} \otimes_{j_1j_2j} \mathbf{u}^\lambda_{j_2} \cdot (\mathbf{u}^\mu_j)^* \tag{9}$$

The EME bispectrum components are indexed on ordered triplets of elements. Hence, for a two-element system, each total bispectrum component is partitioned into eight EME bispectrum components. In general, the bispectrum components are not invariant under permutation of the ordered triplet of elements. However, when two or all three of the bispectrum indices are equal, then certain EME bispectrum components will be equal to each other. These equivalences are expressed by the following identity

$$B^{\sigma(\kappa\lambda\mu)}_{j_1j_2j} = B^{\kappa\lambda\mu}_{j_1j_2j} \Leftrightarrow \sigma(j_1, j_2, j) = (j_1, j_2, j) \tag{10}$$

where σ is an element of the permutation group S_3. The sum of the EME bispectrum components over all ordered triplets of elements is exactly equal to the total bispectrum component defined in the WD-SNAP formulation

$$B_{j_1j_2j} = \sum_{\kappa,\lambda,\mu=1}^{N_{\text{elem}}} B^{\kappa\lambda\mu}_{j_1j_2j} \tag{11}$$

The EME bispectrum components defined in this way have a similar mathematical structure to the descriptors proposed by Drautz[22] in the multicomponent version of the Atomic Cluster Expansion (see Appendix A of ref 22).

In order to demonstrate the improved chemical sensitivity of the EME bispectrum components, we compare WD-SNAP and EME-SNAP descriptors for a phosphorus atom in two very different chemical environments. In the zinc blende ground state structure for bulk indium phosphide each phosphorus atom is covalently bonded to four indium neighbors. In contrast to this, replacing an indium atom with a phosphorus atom creates an antisite defect in which the phosphorus atom has four phosphorus neighbors. We calculated the EME bispectrum components for the eight ordered triplets that can be formed from In and P, and for all half-integer triples (j_1, j_2, j) in the range $0 \leq 2j_2 \leq 2j_1 \leq 2j \leq 2J_{\max} = 6$. For each ordered triplet there are 30 bispectrum components, of which seven belong to the power spectrum with indices $(j, 0, j)$, $0 \leq 2j \leq$

$2J_{\max} = 6$. The power spectrum descriptors are three-body functions, while the remaining bispectrum components are strictly four-body. To simplify the comparison, we omitted the switching function f_c and included only the four nearest neighbors of the phosphorus atoms by setting $R_{\text{cut}} = 4.1$ Å. In Figure 1 we show the difference between each of the
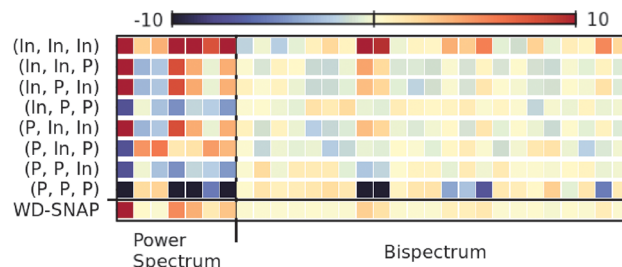

Figure 1. Difference in SNAP descriptors between two distinct chemical environments: the bulk phosphorus site and the phosphorus antisite defect. The WD-SNAP descriptor changes are shown in the last row, while the eight upper rows show the EME-SNAP descriptor changes. The distinction between the power spectrum (left panel) and bispectrum (right panel) is explained in the text.

descriptors in the two atomic environments. It can be seen that the EME bispectrum components retain a large amount of extra information that is lost when they are summed up to form the total bispectrum components of WD-SNAP. In particular, the (In, In, In) and (P, P, P) EME bispectrum components exhibit strong differences of opposite sign, which largely cancel out in the total bispectrum components. This behavior is most pronounced in the case of the power spectrum components (left panel).

B. EME-SNAP Potential Energy Function. Given the EME bispectrum components as descriptors of the neighborhood of each atom, it remains to express the potential energy of a configuration of N atoms in terms of these descriptors. As in our previous work, we decompose the energy of the system containing N atoms with positions $\mathbf{r}^N$ into a sum of local contributions that depend on the neighborhood of each atom and an additional reference energy E_{ref}

$$E(\mathbf{r}^N) = \sum_{i=1}^{N} E_i + E_{\text{ref}}(\mathbf{r}^N) \tag{12}$$

The reference energy is a convenient way to impose desired physical effects, such as long-range electrostatic interactions and strong short-range repulsion, for which well-established energy models exist. Including a reference potential is advantageous because it can correctly represent known limiting cases of atomic interactions, leaving the many-body effects to SNAP. The quality of the SNAP potential will somewhat depend on the choice of reference potential. For example, since the training set does not include highly compressed configurations, the reference potential will need to provide a good physical description of Pauli repulsion which dominates the interaction at close separation. In the case of the indium phosphide potential developed in this work, a previously described ZBL potential was used as the reference potential.[21,28] The local SNAP contribution E_i must capture all the additional effects that are not accounted for by the reference energy. We assume that the local energy can be expressed as a linear function of all the distinct bispectrum components up to some maximum order $J_{\max}$. For a particular

Table 1. Mean Absolute Errors for Energy and Forces for the Various Groups of DFT Training Data for the Branicio,[42] WD-SNAP, and EME-SNAP Potentials

category	N_{config}	N_{forces}	Branicio		WD-SNAP		EME-SNAP	
			E_{MAE} (eV/atom)	F_{MAE} (eV/Å)	E_{MAE} (eV/atom)	F_{MAE} (eV/Å)	E_{MAE} (eV/atom)	F_{MAE} (eV/Å)
bulk	1	2.4×10^1	4.8×10^{-4}		3.2×10^{-4}		5.8×10^{-4}	
defects	428	3.3×10^5	1.4×10^{-1}	3.0×10^{-1}	7.7×10^{-3}	1.5×10^{-1}	3.4×10^{-4}	8.5×10^{-2}
EOS	268	6.4×10^3	2.5×10^0	7.6×10^{-4}	1.1×10^{-2}		7.9×10^{-3}	8.2×10^{-4}
shear	346	8.3×10^3	9.6×10^{-1}	6.19×10^0	3.5×10^{-3}	1.5×10^{-1}	4.6×10^{-3}	1.2×10^0
strain	163	3.9×10^3	3.8×10^{-2}	3.6×10^{-5}	8.1×10^{-3}	2.4×10^{-5}	1.2×10^{-3}	2.4×10^{-5}

choice of J_{max}, we can list the N_B total bispectrum components in some arbitrary order as $B_1, ..., B_{N_B}$. We can further decompose each B_l into EME bispectrum components $B_l^{\kappa\lambda\mu}$ and express the energy as a linear function of these

$$E_i(\mathbf{B}_i) = \sum_{\kappa,\lambda,\mu} \sum_{l=1}^{N_B} \beta_{l,\mu_i}^{\kappa\lambda\mu}(B_{l,i}^{\kappa\lambda\mu} - B_{l0,\mu_i}^{\kappa\lambda\mu}) \tag{13}$$

where $\beta_{l,\mu_i}^{\kappa\lambda\mu}$ are the linear SNAP coefficients for atoms of element μ_i. As a computational convenience, each EME bispectrum component is shifted by the value for an isolated atom, $B_{l0,\mu_i}^{\kappa\lambda\mu}$, so that the local SNAP energy of an isolated atom is zero by construction. The force on each atom k of element μ_k is obtained by summing over all atoms of which it is a neighbor and all EME bispectrum components involving element μ_k

$$\mathbf{F}_k = -\sum_{i=1}^{N} \sum_{\kappa,\lambda}^{N_{elem}} \sum_{l=1}^{N_B} \beta_{l,\mu_i}^{\mu_k\kappa\lambda} \frac{\partial B_{l,i}^{\mu_k\kappa\lambda}}{\partial \mathbf{r}_k} + \beta_{l,\mu_i}^{\kappa\mu_k\lambda} \frac{\partial B_{l,i}^{\kappa\mu_k\lambda}}{\partial \mathbf{r}_k}$$
$$+ \beta_{l,\mu_i}^{\kappa\lambda\mu_k} \frac{\partial B_{l,i}^{\kappa\lambda\mu_k}}{\partial \mathbf{r}_k} \tag{14}$$

The double sum over elements κ and λ in this expression shows that the computational cost of evaluating forces increases as N_{elem}^2. The three terms correspond to the three different positions where element μ_k can appear in the chemical labelings of the EME bispectrum components.

By formulating the SNAP potential energy as a linear function of the EME bispectrum components, the problem of generating the interatomic potential has been reduced to that of choosing the best values for the linear SNAP coefficients. We can achieve this by writing the SNAP contributions to the total energy, the force on an atom, and the stress tensor as explicit functions of the unknown SNAP coefficients $\boldsymbol{\beta}_\mu$

$$E(\mathbf{r}^N) = \sum_{\mu=1}^{N_{elem}} \boldsymbol{\beta}_\mu \cdot \sum_{i \in \mu} \mathbf{B}_i \tag{15}$$

where $\boldsymbol{\beta}_\mu$ is the $(N_B \times N_{elem}^3)$-vector of SNAP coefficients for element μ and $\mathbf{B}_i$ is the $(N_B \times N_{elem}^3)$-vector of EME bispectrum components for atom i. The contribution of the SNAP energy to the force on atom k can be written in terms of the derivatives of the EME bispectrum components wrt $\mathbf{r}_k$, the position of atom k

$$\mathbf{F}_k = -\sum_{\mu=1}^{N_{elem}} \boldsymbol{\beta}_\mu \cdot \sum_{i \in \mu} \frac{\partial \mathbf{B}_i}{\partial \mathbf{r}_k} \tag{16}$$

where $\mathbf{F}_k$ is the force on atom k due to the SNAP energy. Finally, we can write the contribution of the SNAP energy to the stress tensor

$$\mathbf{W} = -\sum_{\mu=1}^{N_{elem}} \boldsymbol{\beta}_\mu \cdot \sum_{i \in \mu} \sum_{k=1}^{N} \mathbf{r}_k \otimes \frac{\partial \mathbf{B}_i}{\partial \mathbf{r}_k} \tag{17}$$

where $\mathbf{W}$ is the contribution of the SNAP energy to the stress tensor and $\otimes$ is the Cartesian outer product operator.

All three of these expressions consist of the vectors $\boldsymbol{\beta}_\mu$ of SNAP coefficients for each element multiplying a vector of quantities that are calculated from the EME bispectrum components of atoms in a configuration. This linear structure greatly simplifies the task of finding the best choice for $\boldsymbol{\beta}_\mu$. We can define a system of linear equations whose solution corresponds to an optimal choice for $\boldsymbol{\beta}_\mu$, in that it minimizes the sum of square differences between the above expressions and the corresponding quantum results defined for a large number of different atomic configurations. This is described in more detail in the following section.

C. Fitting Procedure. The process of fitting a ML-IAP has two key components, construction of a training set and optimization (i.e., learning) of free parameters in the model form. Both of these components will be detailed prior to the comparison of WD-SNAP and EME-SNAP descriptors. In order to demonstrate the improvement over the WD-SNAP form of the EME-SNAP descriptors, a training set has been constructed for InP that not only exercises chemically unique environment but also contains high-energy defects that are challenging for all IAP to capture correctly. Training for accurate defect properties is critical when the intended use of the IAP is to study radiation damage effects where collision cascades of sufficiently high energy leave behind high formation energy point defects.

Since the training of energies and forces for a ML-IAP is done against an electronic structure database, the training configurations are necessarily small $(N_{atoms} \lesssim 10^2)$ to make collecting a large number of them computationally tractable. InP training configurations were generated using the Vienna Ab Initio Simulation Package (VASP)[29−31] using a LDA exchange−correlation functional,[32] PAW pseudopotential[33,34] leaving out the outermost valence p- and d-orbitals of P and In, respectively. K-point grids were generated using the Monkhorst−Pack[35] scheme with a spacing between 0.17 and 0.72 Å^{-1} depending on system size, while a constant plane wave cutoff of 500 eV was used throughout.

In total, 1206 configurations were generated with atom counts per configuration ranging from 8 to 216. Curating training data for a SNAP potential involves grouping training data based on similarities between the configurations, which limits the number of free parameters in the weighted least-squares regression. Each defect type (interstitials, vacancies,

antisites and diantisites) was assigned a unique group such that these configurations could be weighted individually during optimization. In addition to defect configurations, we also included training groups to describe the properties of the bulk zinc blende structure. This includes configurations for uniform expansion and compression (equation of state), random cell shape modifications (shear group), and uniaxially strained (strain group) unit cells. Lastly, the ground state configuration for the bulk zinc blende structure is left as a separate group. The number of configurations (N_{config}) in each training data group is summarized in Table 1. Energies and forces are weighted separately, resulting in ten fitting degrees of freedom from the group weights. The weighting of different training groups is applied during the regression step within our fitting software FitSNAP[36] by modifying the diagonal matrix **w** in the system of equations:

$$\min(\|\mathbf{w}(A\mathbf{x}_\beta - T)\|^2 - \gamma_n\|\mathbf{x}_\beta\|) \tag{18}$$

where A is a matrix of bispectrum components computed from LAMMPS and T is a vector of energies and forces taken from VASP. A regularization penalty, γ_n, can be used to aid against overfitting the solution $\mathbf{x}_\beta$, but to date has not yielded better fits than use without.

In addition to weighting certain training data more/less heavily as part of the optimization, the radial cutoffs ($R_{cut}^{\mu_j\mu_k}$ of eq 6) and per-element density weights (w_β of eq 6) are included as free parameters but are considered hyperparameters since they directly affect the calculation of the descriptors. All of these parameters have been described in previous work,[21,24] except for the atomic energy difference, which warrants further explanation. The DFT energy is computed as a sum of contributions from different terms in the full electron−ion quantum system, most of which are nonzero for an isolated atom. The SNAP potential, like most classical potentials, is constructed so that each isolated In or P atom has zero energy. To reconcile these two conventions, it is necessary to shift the DFT energy data by an amount that depends only on the element type of each atom, which we call the atomic energy. We constrained the sum of the In and P atomic energies by fixing the energy of the ground state structure at −3.48 eV/atom, ensuring that the experimental cohesive energy[37] of InP is reproduced. The difference between the In and P atomic energies is an additional free parameter.

We automated the fitting process using the optimization software DAKOTA,[39] the overall workflow is schematically captured in Figure 2 and is the same when optimizing EME-SNAP or WD-SNAP ML-IAP. The hyperparameters and the group weights were optimized in two stages. First, the group weights were held fixed and the hyperparameters are optimized using a Single Objective Genetic Algorithm (SOGA). These hyperparameter search spaces were necessarily bounded, restricting the parameters to physical values, and to save computational time, these ranges and the optimal values for each hyperparameter are given in Supplemental Table 1. Once stable values for the hyperparameters were determined, a second SOGA optimization was performed on the training group weights which are allowed to vary over a large range (10^1−10^7).

For all DAKOTA driven optimizations, we have constructed a set of equally weighted objective functions that will be used to judge the quality of the fit. These objective functions are the
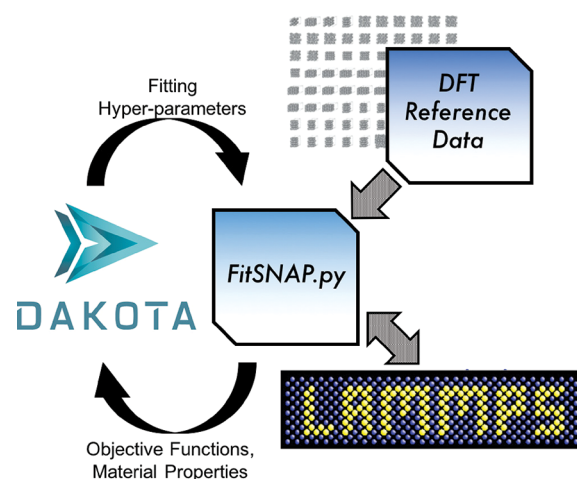

Figure 2. Schematic of the WD-SNAP and EME-SNAP fitting procedure. The FitSNAP[36] software package provides the overall workflow framework. LAMMPS[38] efficiently evaluates the WD-SNAP and EME-SNAP descriptors, while DAKOTA[39] performs the hyperparameter optimization. Reprinted figure with permission from ref 24. Copyright 2019 by the American Physical Society.

lattice parameter and cohesive energy of zinc blende InP as well as the difference between the DFT values and IAP predicted defect formation energies for the following stoichiometric pairs: In interstitial + In vacancy, P interstitial + P vacancy, In interstitial + P interstitial, In vacancy + P vacancy, In antisite + P antisite. Objective functions should be considered as *trained* properties rather then testing since their error minimization is included in the generation of optimal candidate potentials. Referring back to the workflow shown in Figure 2, one pass through the optimization loop proceeds as follows. For each set of candidate hyperparameters or group weights proposed by DAKOTA, linear regression was used to solve for the SNAP coefficients using FitSNAP.[40] With each new candidate potential, LAMMPS is used[38,41] to relax a full set of InP defect configurations as well as the bulk InP zinc blende structure. For each relaxation, the configuration was first annealed at 10 K for 0.1 ps before performing the minimization wherein the volume of the cell was also allowed to relax. This added annealing step combats a common pitfall of interatomic potential fitting in which training structures that are stable minima are represented by the potential as unstable or marginally stable saddle points. Relaxing the defect structure after adding small perturbations from equilibrium tests the local stability of defect configurations as predicted by the candidate potentials. The absolute error in these seven stoichiometric defect formation energies and the sum of the error in the cohesive energy and lattice parameter of zinc blende formed the eight distinct SOGA objective functions. A generation consisted of 200 candidates after which hybridization and mutation steps that adjust group weights are carried out, the global fit was considered converged after 7600 generations.

x‘

III. RESULTS AND DISCUSSION

To demonstrate the improvement provided by the EME-SNAP descriptors, optimized fits of either ML-IAP will be compared to available experimental data for InP properties near equilibrium as well as for high-energy point defects. For a

complete comparison, we will also include predictions from a prior empirical IAP from Branicio et al.[42] The Branicio potential for InP is a physics-motivated empirical model with electrostatic, van der Waals, and three-body interactions that take into account the covalent nature of the InP interaction.

To begin, Table 1 summarizes the errors for the reduced set of training groups outlined in the previous section. For each training group and IAP type the mean absolute error (MAE) is reported for both the energy (E_{MAE}) and forces (F_{MAE}). In all cases except for the training group corresponding to sheared geometries, the EME-SNAP potential reproduces the DFT values of energy and atomic force most accurately. The average energy and force errors across the entire training set for EME-SNAP are 3.3 meV/atom and 67 meV/Å, respectively, which are quite good for the training set size put forth here. In contrast, the average energy and force errors for the SNAP potential are 5.7 meV/atom and 75 meV/Å, respectively.

Our main focus was to test whether EME-SNAP could reproduce the defect formation energies more effectively than the WD-SNAP model. Figure 3 displays the defect formation
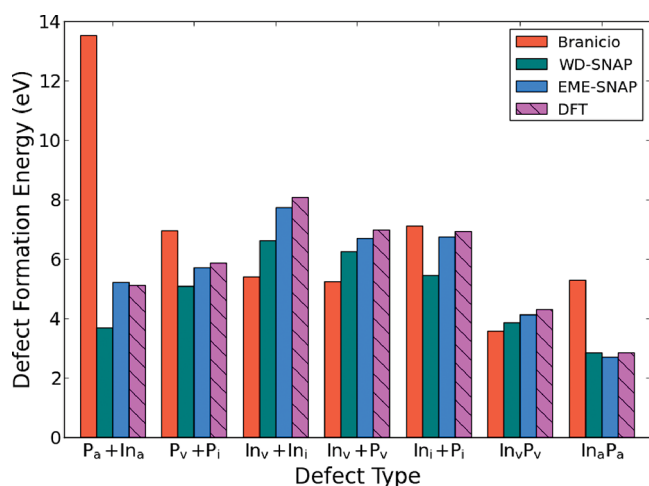

Figure 3. Relaxed defect formation energies for stoichiometric combinations of point defects in InP zinc blende structure for classical potentials compared to DFT. Orange, green, and blue represent the Branicio, SNAP, and EME-SNAP potentials, respectively. The pink hatched columns represent the relaxed DFT formation energies.

energies for EME-SNAP (blue) and DFT (pink crosshatched). Subscripts in this figure correspond to vacancy (v), interstital (i), and antisite (a) point defects. As described previously, the defect formation energies are combined in stoichiometric combinations. Overall, EME-SNAP well reproduces the formation energies with the largest deviation from DFT being about 0.35 eV. With EME-SNAP, there are now eight descriptors for each j, j_1, and j_2 triplet combination (see Figure 1) and these descriptors can clearly distinguish between local environments composed of varying compositions of multiple atomic species. This is especially useful for high-energy defects of lowered local symmetry compared to bulk InP. Importantly, the relative ordering of the seven formation energies is almost identical for DFT and EME-SNAP. The exception is that DFT predicts the combined $In_v + P_v$ formation energy to be about 0.06 eV higher than the combined $In_i + P_i$ defect, but EME-SNAP predicts the latter to be higher by 0.04 eV. This small difference will likely not drastically affect the defect population

distributions obtained in radiation damage molecular dynamics simulations. The largest discrepancies between DFT and EME-SNAP are for the combined $In_v + In_i$ and the combined $P_v + In_v$; both are underpredicted by about 0.35 and 0.30 eV, respectively. In general, the interstitial configurations tend to be most difficult to accurately reproduce. However, this level of disagreement is acceptable, since the relative ordering of defects is accurately preserved. This is the most important factor in predicting the correct defect populations in collision cascade simulations and further evolution of the cascade.

Figure 3 also displays the relaxed defect formation energies for an empirical potential by Branicio et al.[42] (red) and the WD-SNAP potential (blue). The Branicio IAP predicts defect formation energies poorly match DFT values with an average error of 2.46 eV, making this potential unsuitable for radiation damage studies. Regarding the ML-IAP fitted here, the WD-SNAP potential improves upon the Branicio potential but some of the defect formation energies still differ from DFT by as much as 1.4 eV. These discrepancies ultimately indicate the defects that would form during a collision cascade simulation using either the Branicio or WD-SNAP potential will most likely not be consistent with DFT. In contrast, the EME-SNAP defect formation energies are all within 0.35 eV of the DFT values and captures the overall trend of the DFT values.

While this maximal error value (0.35 eV) may appear rather larger, it occurs for the defect combination with the highest formation energy ($In_v + In_i$, 7.98 eV), and so corresponds to a relative error of less than 5%. Moreover, the EME-SNAP ordering of defect combinations from least to most stable is consistent with DFT predictions, in contrast to the other potentials. Many of the issues with the WD-SNAP potential stemmed from the initial configurations relaxing to distinctly different structures during annealing and minimization. With EME-SNAP, the decreases in energy during MD relaxation were in all cases small and the structure predicted by DFT was preserved.

In addition to defect formation energies, we also examined the IAP predictions for other bulk InP properties. Table 2 lists the zinc blende structure elastic constants for each of the potentials. Given the simplicity of these equilibrium properties, it is unsurprising that the elastic constants are reasonably consistent with both experimental and DFT values for all potentials. Still, there still are trade-offs in accuracy among

Table 2. Structural and Mechanical Properties of the InP Zincblende Structure, As Predicted by the Branicio,[42] WD-SNAP, and EME-SNAP Potentials, as Well as Experimental[43] and DFT Values[a]

	Branicio	WD-SNAP	EME-SNAP	expt	DFT
a_0 (Å)	5.83	5.83	5.83	5.83	5.84
C_{11} (GPa)	102.5	122.0	113.7	101.1	99.3
C_{12} (GPa)	57.3	90.6	70.9	56.1	55.4
C_{44} (GPa)	69.6	63.6	48.4	45.6	45.0
bulk mod (GPa)	72.3	101.1	85.1	71.1	70.1
shear mod (GPa)	22.6	15.7	21.4	22.5	21.9
Poisson ratio	0.36	0.43	0.38	0.36	0.36
CPU (ns/day)	94.3	2.5	1.1		
GPU (ns/day)	65.2	6.9	2.2		

[a]Additionally, computational speed is given in simulated time per day running a representative MD simulation on both CPU and GPU architectures.

each. The Poisson ratio and bulk and shear moduli of EME-SNAP differ from DFT predictions by 5.6%, 21.4%, and 2.3%, whereas WD-SNAP and the Branicio potential predict (19.4%, 44.2%, and 28.3%) and (0.0%, 3.1%, and 3.2%), respectively, in this triplet of important properties. Overall, EME-SNAP does reasonably well in reproducing the elastic constants predicted by both experiments and DFT, but not to the level that the Branicio potential could. However, these values could potentially be improved upon by including stress tensor training data and adding the DFT elastic constants as objective functions during the fitting process.

Another important property is the relative stability of different low-energy crystal structures. Cold-curve equations of state for the zinc blende (ZB) and rock salt (RS) InP structures are compared with DFT and plotted in Figure 4.
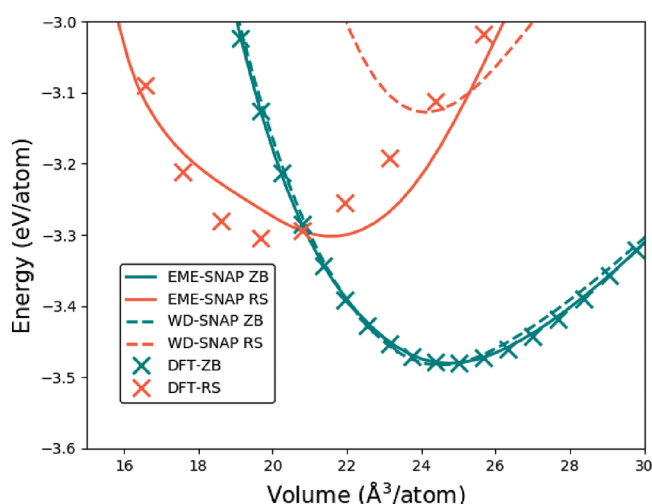
Figure 4. Energy versus volume for InP crystal structures calculated using the EME-SNAP potential (lines) and DFT (crosses). Green and orange represent zinc blende and rock salt structures, respectively.

These data were not included as training or objective functions and should be considered as test properties of the SNAP potentials. EME-SNAP correctly predicts ZB as the most stable structure and reproduces the experimental cohesive energy of −3.48 eV/atom at a volume of 24.4 Å^3/atom while the RS cohesive energy of −3.30 eV/atom matches exactly the DFT value. However, EME-SNAP predicts a slightly higher volume of 22.2 Å^3 compared to the DFT value of 19.69 Å^3. While not plotted in Figure 4, the wurtzite (WZ) ground state structure was also calculated. The DFT prediction of the WZ cohesive energy was found to be −3.47 eV/atom, slightly higher than the ZB structure. While the EME-SNAP potential does predict the WZ to be higher in energy than ZB, the predicted value is too high, being about −2.27 eV/atom. The EME-SNAP representation of the RS and WZ properties is sufficient, given that these phases were not represented in the training data. Again, the accuracy of these phases could be improved in future iterations of this potential by including them in the training data and by using the RS, ZB, and WZ cohesive energies as an objective functions during fitting.

Lastly, we have compared the performance between the three different InP potentials analyzed in this work. Timing data are given in Table 2 for a representative 1728 atom zinc blende structure running NVE dynamics. Both CPU (32 Intel Broadwell cores) and GPU (one NVIDIA P100) timings are presented, with interesting trends among the three potentials tested. Clearly the simpler Branicio potential runs faster than the SNAP potentials on both architectures, although it is less able to take advantage of greater processing power of the GPU. Both WD- and EME-SNAP see a more than 2-fold speedup going from 32 CPU cores to one GPU. On both architectures, the cost for the increased accuracy of the EME-SNAP potential is about a factor of 3 compared to WD-SNAP. More information about the details of the Kokkos-based CPU and GPU implementations of SNAP can be found on the LAMMPS Web site[38] and in a separate publication.[44]

While the EME-SNAP InP potential significantly improves the defect formation energies and reproduces some basic InP properties, additional training data will be needed to further improve the transferability of the potential for other target applications. For instance, adding liquid phase training data would be useful as a liquid will sample a wide variety of local chemical environments. The overall fitting process could also be improved by incorporating other properties as objective functions. In this study, the defect formation energies were the primary focus, since accurately reproducing these properties is essential for realistic simulations of radiation damage, the focus of future work.

IV. CONCLUSIONS

We have developed a natural extension to the SNAP interatomic potential form that improves the description of multicomponent systems. The new formulation, called EME-SNAP, uses chemically labeled descriptors that explicitly separate out contributions from different chemical elements in the atomic environment. This new method was applied to InP where previous potentials were unable to adequately represent defect formation energies that are essential for conducting molecular dynamics simulations of radiation damage effects. The new EME-SNAP potential reproduced relaxed defect formation energies to within 0.35 eV compared to DFT, whereas the original weighted density SNAP formulation exhibited discrepancies of more than 1 eV. This improvement indicates that EME-SNAP is better able to distinguish between different chemical environments. Other properties including volume dependence of energy in the zinc blende and rock salt crystal structures are also well modeled by EME-SNAP. The new EME-SNAP method shows promise in generating a potential suitable for collision cascade simulations. The potential presented in this work was developed strictly for modeling defect formation energies and configurations close to equilibrium. It has yet to be examined whether this new formulation will perform well for radiation damage simulations. The training set will need to be expanded to incorporate additional configurations to create a more general use potential capable of simulating the full cascade. Future work will focus on using the EME-SNAP form to develop a potential with a broader training set and deploying it in radiation damage simulations.

◼ ASSOCIATED CONTENT

ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.0c02450.

> Details on the search range and final values for the hyperparameter search for both WD-SNAP and EME-SNAP (PDF)

AUTHOR INFORMATION

Corresponding Author

    A. P. Thompson — *Center for Computing Research, Sandia National Laboratories, Albuquerque, New Mexico 87185, United States;* orcid.org/0000-0002-0324-9114; Email: athomps@sandia.gov

Authors

    M. A. Cusentino — *Center for Computing Research, Sandia National Laboratories, Albuquerque, New Mexico 87185, United States*

    M. A. Wood — *Center for Computing Research, Sandia National Laboratories, Albuquerque, New Mexico 87185, United States*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jpca.0c02450

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Stillinger, F.; Weber, T. Computer simulation of local order in condensed phases of silicon. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1985**, *31*, 5262.

(2) Tersoff, J. New empirical approach for the structure and energy of covalent systems. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 6991.

(3) Vashishta, P.; Kalia, R. K.; Rino, J. P.; Ebbsjö, I. Interaction potential for SiO_2: A molecular-dynamics study of structural correlations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1990**, *41*, 12197−12209.

(4) Daw, M. S.; Baskes, M. I. Semiempirical, quantum mechanical calculation of hydrogen embrittlement in metals. *Phys. Rev. Lett.* **1983**, *50*, 1285.

(5) Baskes, M. I. Application of the embedded-atom Method to covalent materials: A semiempirical potential for silicon. *Phys. Rev. Lett.* **1987**, *59*, 2666.

(6) Plimpton, S. J.; Thompson, A. P. Computational aspects of many-body potentials. *MRS Bull.* **2012**, *37*, 513.

(7) Bartók, A.; Kondor, R.; Csanyi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(8) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(9) Lindsey, R. K.; Fried, L. E.; Goldman, N. Chimes: A force matched potential with explicit three-body interactions for molten carbon. *J. Chem. Theory Comput.* **2017**, *13*, 6222−6229.

(10) Bartók, A.; Payne, M.; Kondor, R.; Csanyi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(11) Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153−1173.

(12) Rupp, M.; Tkatchenko, A.; Muller, K.; von Lilienfeld, A. O. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(13) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: Recent applications and prospects. *npj Comput. Mater.* **2017**, *3*, 54.

(14) Kim, C.; Chandrasekaran, A. H.; Tran, D.; Das, D.; Ramprasad, R. Polymer genome: A data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C* **2018**, *122*, 17575−17585.

(15) Nelson, L. J.; Hart, G. L. W.; Zhou, F.; Ozoliņš, V. Compressive sensing as a paradigm for building physics models. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 035125.

(16) Rajan, K. Materials informatics: The materials "gene" and big data. *Annu. Rev. Mater. Res.* **2015**, *45*, 153−169.

(17) Wang, Q.; Jain, A. A transferable machine-learning framework linking interstice distribution and plastic heterogeneity in metallic glasses. *Nat. Commun.* **2019**, *10*, 5537.

(18) Xie, S. R.; Stewart, G. R.; Hamlin, J. J.; Hirschfeld, P. J.; Hennig, R. G. Functional form of the superconducting critical temperature from machine learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *100*, 174513.

(19) Rosenbrock, C. W.; Gubaev, K.; Shapeev, A. V.; Pártay, L. B.; Bernstein, N.; Hart, G. L. W. Machine-learned interatomic potentials for alloys and alloy phase diagrams. *arXiv:1906.07816* **2019**.

(20) Nyshadham, C.; Rupp, M.; Bekker, B.; Shapeev, A. V.; Mueller, T.; Rosenbrock, C. W.; Csányi, G.; Wingate, D. W.; Hart, G. L. W. Machine-learned multi-system surrogate models for materials prediction. *npj Comput. Mater.* **2019**, *5*, 51.

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

(22) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *99*, 014104.

(23) Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csanyi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A.; et al. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **2020**, *124*, 731−745.

(24) Wood, M. A.; Cusentino, M. A.; Wirth, B. D.; Thompson, A. P. Data-driven material models for atomistic simulation. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *99*, 184305.

(25) Li, X.-G.; Hu, C.; Chen, C.; Deng, Z.; Luo, J.; Ong, S. P. Quantum-accurate spectral neighbor analysis potential models for Ni-Mo binary alloys and fcc metals. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2018**, *98*, 094104.

(26) Daw, M. S.; Foiles, S. M.; Baskes, M. I. The embedded-atom method: A review of theory and applications. *Mater. Sci. Rep.* **1993**, *9*, 251.

(27) Baskes, M. I. Modified embedded-atom potentials for cubic materials and impurities. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 2727.

(28) Ziegler, J.; Biersack, J. P.; Littmark, U. *The stopping and range of ions in matter*; Pergamon: New York, NY, 1985.

(29) Kresse, G.; Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *47*, 558.

(30) Kresse, G.; Furthmuller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 11169.

(31) Kresse, G.; Furthmuller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **1996**, *6*, 15.

(32) Perdew, J. P.; Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1981**, *23*, 5048.

(33) Blochl, P. E. Projector augmented-wave method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1994**, *50*, 17953.

(34) Kresse, G.; Joubert, J. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 1758.

(35) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976**, *13*, 5188.

(36) FitSNAP. http://github.com/FitSNAP/FitSNAP (accessed May 6, 2020).

(37) Harrison, W. A. *Electronic Structure and the properties of solids: The physics of the chemical bond*; Dover: New York, USA, 1989.

(38) LAMMPS Large-scale Atomic/Molecular Massively Parallel Simulator. http://lammps.sandia.gov (accessed May 6, 2020).

(39) Adams, B. M.; Bohnhoff, W.; Dalbey, K.; Eddy, J.; Eldred, M.; Gay, D.; Haskell, K.; Hough, P. D.; Swiler, L. P. *Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis*; Technical Report SAND 2010-2183; Sandia National Laboratories: Albuquerque, NM, 2009.

(40) Thompson, A. P.; Schultz, P. A.; Crozier, P. S.; Moore, S. G.; Swiler, L. P.; Stephens, J. A.; Foiles, S. M.; Tucker, G. J. *Automated algorithms for quantum-level accuracy in atomistic simulations: LDRD final report*; Technical Report SAND 2014-17862; Sandia National Laboratories: Albuquerque, NM, 2014.

(41) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1995**, *117*, 1−19.

(42) Branicio, P. S.; Rino, J. P.; Gan, C. K.; Tsuzuki, H. J. Interaction potential for indium phosphide: a molecular dynamics and first-principles study of the elastic constants, generalized stacking fault and surface energies. *J. Phys.: Condens. Matter* **2009**, *21*, 095002.

(43) Nichols, D. N.; Rimai, D. S.; Sladek, R. J. Elastic anharmonicity of InP: Its relationship to the high pressure transition. *Solid State Commun.* **1980**, *36*, 667.

(44) Gayatri, R.; Moore, S.; Weinberg, E.; Lubbers, N.; Anderson, S.; Deslippe, J.; Perez, P.; Thompson, A. P. Optimizing SNAP for future generation architectures using TestSNAP - A proxy app for LAMMPS/SNAP. *Supercomputing* **2020**, accepted.