Proper Orthogonal Descriptors for Efficient and Accurate Interatomic Potentials

Ngoc-Cuong Nguyen\textsuperscript{a}, Andrew Rohskopf\textsuperscript{b}

\textsuperscript{a}Center for Computational Engineering, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, 02139, MA, USA
\textsuperscript{b}Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, 02139, MA, USA

Abstract

We present the proper orthogonal descriptors for efficient and accuracy representation of the potential energy surface. The potential energy surface is represented as a many-body expansion of parametrized potentials in which the potentials are functions of atom positions and parameters. The Karhunen-Loève (KL) expansion is employed to decompose the parametrized potentials into a set of proper orthogonal descriptors (PODs). Because of the rapid convergence of the KL expansion, relevant snapshots can be sampled exhaustively to represent the atomic neighborhood environment accurately with a small number of descriptors. The proper orthogonal descriptors are used to develop interatomic potentials by using a linear expansion of the descriptors and determining the expansion coefficients from a weighted least-squares regression against a density functional theory (DFT) training set. We present a comprehensive evaluation of the POD potentials on previously published DFT data sets comprising Li, Mo, Cu, Ni, Si, Ge, and Ta elements. The data sets represent a diverse pool of metals, transition metals, and semiconductors. The accuracy of the POD potentials are comparable to that of state-of-the-art machine learning potentials such as the spectral neighbor analysis potential (SNAP) and the atomic cluster expansion (ACE).

Keywords: interatomic potentials, machine learning potentials, Karhunen-Loève expansion, proper orthogonal descriptors, atomic cluster expansion, spectral neighbor analysis potential
1. Introduction

Decomposing complex systems into a hierarchy of interacting sub-systems at different scales is a recurring theme in physics. The approach applied to atomic systems lends itself naturally to learning atomic interactions. For example, to calculate the total energy, one might just sum up independent contributions from each of its neighbors. Specifically, the potential energy \( E \) is constructed as a sum of local atomic energies \( E_i \) of all atoms \( i \) in the system

\[
E(\mathbf{r}_1, \ldots, \mathbf{r}_N) = \sum_{i=1}^{N} E_i(\mathbf{r}_1, \ldots, \mathbf{r}_N),
\]

where \( \mathbf{r}_i \) are Cartesian coordinates of atoms \( i \) in the system. Although the local energies \( E_i \) are generally functions in \( 3N \)-dimensional space, they are often expressed in much lower dimensional spaces by limiting the atomic interaction to local atomic environments around the atoms \( i \). Typically, the local energies take the form

\[
E_i = \sum_{m=1}^{M} \Psi_m(\mathbf{q}_i, \mathbf{c}_m), \quad i = 1, \ldots, N,
\]

where \( \mathbf{q}_i \) are the local coordinates attached to each atom \( i \), \( \{\mathbf{c}_m\}_{m=1}^{M} \) are sets of parameters for potential functions \( \{\Psi_m(\mathbf{q}, \mathbf{c})\}_{m=1}^{M} \), which together encode local atomic environments around atoms \( i \). The local coordinates \( \mathbf{q}_i \) include coordinates of the atom \( i \) and those of the neighbors within a cutoff distance from the atom \( i \). Since the number of neighbors around any atom \( i \) can often be orders of magnitude less than \( N \), the expansion (2) allows for substantial dimensionality reduction.

Over the years, a large number of empirical potentials have been developed to provide analytical forms of the potential functions \( \Psi_m \). Empirical potentials are usually derived from physical insights of electronic structure theories. Despite many years of efforts devoted to the development of empirical potentials, open problems still need to be addressed. Empirical potentials often suffer from not being able to reproduce properties of a system under different temperature, pressure and environmental conditions. More often than not, empirical potentials are recalibrated by reparameterizing the existing set of parameters or extending the energy formula by adding terms and then parameterizing the entire potential anew. This results in many different sets of parameters for widely-used empirical potentials such as EAM [1],
Stillinger-Weber [2], Tersoff [3], EDIP [4], REBO [5], ReaxFF [6]. Empirical potentials and their specific parametrization should be applied to problems for which they are known to produce consistently accurate results. Developing accurate interatomic potentials for a wide variety of systems is still an active and important area of research.

In recent years, a significant trend has emerged in the form of machine learning potentials, where the potential energy surface is described as a function of local environment descriptors that are invariant to translation, rotation, and permutation of atoms in the system. Examples of such potentials include the high-dimensional neural network potential (NNP) [7, 8, 9], the Gaussian approximation potential (GAP) [10, 11, 12], the spectral neighbor analysis potential (SNAP) [13, 14], moment tensor potentials (MTP) [15, 16], and the atomic cluster expansion [17, 18, 19]. Typical approach to training such potentials involves the generation of a sufficiently large and diverse data set of atomic configurations with corresponding energies, forces and stresses from DFT calculations, which are then used in the training of ML potentials based on one or several target metrics, such as minimizing the mean absolute or squared errors in predicted energies, forces, stresses, or derived properties (e.g., elastic constants). A recent work [20] presents a comprehensive evaluation of ML potentials based on four local environment descriptors - atom-centered symmetry functions (ACSF), smooth overlap of atomic positions (SOAP), the SNAP bispectrum components, and moment tensors. It is shown in [20] that ML potentials provide substantial improvement over EAM and Tersoff potentials in predicting energies and forces across diverse chemistries and configurations.

The descriptors play a central role in the accurate representation of the potential energy. They must meet several requirements. First and foremost, the descriptors must be invariant with respect to permutation, rotation, translation, and reflection of atoms in the system. Second, since the transformation from atom coordinates onto the descriptors has to be carried out for every atom, the evaluation should be fast, and the descriptors need to be differentiable with respect to the atomic positions to enable the calculation of analytic gradients for the forces. Third, in order to obtain accurate ML potentials, the descriptors must provide a detailed structural description of the local atomic environment. The challenge of finding suitable descriptors for ML potentials was recognized in earlier work [7]. Initially, the lack of suitable descriptors was the main obstacle to construct ML potentials for high-dimensional systems because multidimensional functions were required.
to fit the many-body interactions in electronic structure calculations with high precision. In recent years, significant progress has been made in developing a wide variety of descriptors for ML potentials. We refer to [21, 22, 19] for a comprehensive survey of methods to transform the coordinates of atoms onto a suitable set of descriptors.

In this paper, we present a new method to represent atomic neighborhood environments with the so-called proper orthogonal descriptors. The method is motivated by the reduced basis method [23, 24, 25, 26, 27] for parametrized partial differential equations. In spirit of the reduced basis method, we view the potential energy surface as a manifold of potential energies over parameters. Instead of calibrating or optimizing the parameters, we aim to construct an efficient and accurate approximation of the manifold of potential energies, thereby providing a rich and diverse representation of atomic interactions. To this end, we employ the Karhunen-Loève (KL) expansion [28, 29, 30] to generate orthogonal basis functions that optimally approximate the manifold of potential energies. These basis functions are then transformed onto a set of proper orthogonal descriptors (PODs) to represent atomic neighborhood environments. Owing to the exponential convergence of the KL expansion, accurate representation of local atomic environments can be achieved with a small number of descriptors. The proper orthogonal descriptors are used to develop ML potentials by using a linear expansion of the descriptors and determining the expansion coefficients from a weighted least-squares regression against DFT training set. We assess the POD potentials on diverse data sets of Li, Mo, Cu, Ni, Si, Ge, and Ta elements, and compare their performance to that of other linear models such as SNAP and ACE potentials.

The paper is organized as follows. In Section 2, we introduce proper orthogonal descriptors. In Section 3, we discuss the connection of our work with previous work. In Section 4, we describe the implementation of the proposed descriptors and the potential fitting. In Section 5, we present numerical results to assess the performance of the proposed method. Finally, in Section 6, we make a number of concluding remarks on the results as well as future work.

2. Proper Orthogonal Descriptors

2.1. Parametrized potential energy surface

Let \( \mathbf{r}_n \in \Omega \) be a position vector of an atom \( n \) in a physical domain \( \Omega \in \mathbb{R}^3 \). We consider a system of \( N \) atoms with \( N \) position vectors \( \mathbf{R} = \)
\((r_1, r_2, \ldots, r_N) \in \mathbb{R}^{3N}\). The potential energy surface (PES) of the system of \(N\) atoms can be expressed as a many-atom expansion of the form

\[
E(R, \eta, \mu) = V^{(0)} + \sum_i V^{(1)}(r_i; \mu^{(1)}) + \frac{1}{2} \sum_{i,j} V^{(2)}(r_i, r_j; \eta^{(2)}, \mu^{(2)}) + \frac{1}{6} \sum_{i,j,k} V^{(3)}(r_i, r_j, r_k; \eta^{(3)}, \mu^{(3)}) + \ldots
\]

where \(V^{(0)}\) is a constant energy offset, \(V^{(1)}\) is the one-body potential often used for representing external field or energy of isolated elements, and the higher-body potentials \(V^{(2)}, V^{(3)}, \ldots\) are symmetric, uniquely defined, and zero if two or more indices take identical values. The superscript on each potential denotes its body order. Note that each \(q\)-body potential \(V^{(q)}\) depends on \(\mu^{(q)}\) and \(\eta^{(q)}\), which are a set of parameters to fit the potential and a set of hyperparameters, respectively. Then \(\mu\) is a collection of all potential parameters \(\mu^{(1)}, \mu^{(2)}, \mu^{(3)}, \ldots\), and \(\eta\) is a collection of all hyperparameters \(\eta^{(2)}, \eta^{(3)}, \ldots\). The hyperparameters include inner and outer cut-off distances. Typically, lower-body potentials have longer cut-off distances than higher-body potentials. Note that the one-body potential \(V^{(1)}\) does not have hyperparameters since there is only one central atom \(i\) in the one-body interaction.

In the absence of external fields, the PES should not depend on the absolute position of atoms, but only on the relative positions. This means that the PES can be rewritten as a function of interatomic distances and angles between the bonds as

\[
E(R, \eta, \mu) = V^{(0)} + \sum_i V^{(1)}(\mu^{(1)}) + \frac{1}{2} \sum_{i,j} V^{(2)}(r_{ij}; \eta^{(2)}, \mu^{(2)}) + \frac{1}{6} \sum_{i,j,k} V^{(3)}(r_{ij}, r_{ik}, \theta_{ijk}; \eta^{(3)}, \mu^{(3)}) + \ldots
\]

where each potential is a function of interatomic distances \(r_{ij} = |r_i - r_j|\), bond angles \(\theta_{ijk}\), potential parameters, and the hyperparameters. A wide variety of interatomic potentials such as Lennard-Jones potential, Morse potential, Stillinger-Weber potential, angle potentials, and dihedral potentials have the form \(\text{(4)}\).

In other interatomic potentials such as EAM and Tersoff potentials, the many-body interactions are embedded into the terms of a pair potential in
which the nature of the interaction is modified by the local environment of
the atom via the bond order parameter, coordination number, or electron
density. For the EAM potential, the energy is a nonlinear function of a sum
of functions of the interatomic distance. The Tersoff potential includes an
explicit dependence on bond angles to describe directional bond formation,
so that the energy is written as a nonlinear function of a sum of functions of
the interatomic distance and bond angle.

Interatomic potentials rely on parameters to learn relationship between
atomic environments and interactions. Since interatomic potentials are ap-
proximations by nature, their parameters need to be set to some reference
values or fitted against experimental and/or DFT data sets by necessity. In
simple potentials such as the Lennard-Jones and Morse potential, the pa-
rameters can be set to match the equilibrium bond length and bond strength
of a dimer molecule or the surface energy of a solid. Many-body potentials
often contain many parameters with limited interpretability and need to be
optimized by fitting their parameters against a larger set of data. Typically,
potential fitting involves solving a regression problem to find optimal param-
eters, $\mu^*$, which minimize a certain loss function of the predicted quantities
and data. The accuracy of a fitted potential at different conditions other than
those used in the fitting procedure is often measured in terms of transferabil-
ity of the potential. It is desirable to make the potential transferable, so that
it can describe properties that are different from those it was fitted to. Since
the fitted potential depends on the data set used to fit it, different data sets
will yield different optimal parameters and thus different fitted potentials.
When fitting the same functional form on $Q$ different data sets, we would
obtain $Q$ different optimized potentials, $E(R, \eta, \mu^*_q)$, $1 \leq q \leq Q$. These opti-
mized potentials are typically intended to predict properties that are similar
to those they are fitted to. Inaccurate predictions can occur when they are
used to predict out-of-fitting properties. Moreover, the hyperparameters are
left to be defined by users of the potentials.

Instead of finding optimal parameters for the PES, we draw inspira-
tion from the reduced basis method for parametrized partial differential equa-
tions and propose to view the parametrized PES as a parametric manifold of potential energies

$$\mathcal{M} = \{E(R, \eta, \mu) \mid \mu \in \Omega^\mu \}$$

where $\Omega^\mu$ is a parameter domain in which $\mu$ resides. The parametric manifold
$\mathcal{M}$ contains potential energy surfaces for all values of $\mu \in \Omega^\mu$. Therefore,
the parametric manifold yields a much richer and more transferable atomic representation than any particular individual PES $E(R, \eta, \mu)$. Our goal in this section is to construct an efficient and accurate approximation to the parametric manifold $\mathcal{M}$. To this end, we apply the KL expansion to snapshots (different values of parameters $\mu$) of the parametric manifold to obtain a set of proper orthogonal descriptors (PODs). The descriptors can then be employed in a fitting scheme, such as linear regression, to develop accurate and transferable interatomic potentials. The first step to crafting the PODs is to define a parametric manifold and its domain of parameters.

### 2.2. Two-body proper orthogonal descriptors

To define a parametric manifold for two-body interactions, we adopt the usual assumption that the direct interaction between two atoms vanishes smoothly when their distance is greater than the outer cutoff distance $r_{\text{max}}$. Furthermore, we assume that two atoms can not get closer than the inner cut-off distance $r_{\text{min}}$ due to the Pauli repulsion principle. Letting $r \in (r_{\text{min}}, r_{\text{max}})$, we introduce the following parametrized radial functions

$$\phi(r, r_{\text{min}}, r_{\text{max}}, \alpha, \beta) = \sin(\alpha \pi x) / (r - r_{\text{min}}), \quad \varphi(r, \gamma) = 1 / r^\gamma,$$

where the scaled distance function $x$ is defined below to enrich the two-body manifold

$$x(r, r_{\text{min}}, r_{\text{max}}, \beta) = e^{-\beta(r-r_{\text{min}})/(r_{\text{max}}-r_{\text{min}})} - 1 / e^{-\beta} - 1. \quad (7)$$

We introduce the following function as a convex combination of the two functions in (6)

$$\psi(r, r_{\text{min}}, r_{\text{max}}, \alpha, \beta, \gamma, \kappa) = \kappa \phi(r, r_{\text{min}}, r_{\text{max}}, \alpha, \beta) + (1 - \kappa) \varphi(r, \gamma). \quad (8)$$

We see that $\psi$ is a function of distance $r$, cut-off distances $r_{\text{min}}$ and $r_{\text{max}}$, and parameters $\alpha, \beta, \gamma, \kappa$. Together these parameters allow the function $\psi$ to characterize a diverse spectrum of two-body interactions within the cut-off interval $(r_{\text{min}}, r_{\text{max}})$.

Next, the two-body potential is defined as follows

$$V^{(2)}(r_{ij}, \eta^{(2)}, \mu^{(2)}) = f_c(r_{ij}, \eta^{(2)}) \psi(r_{ij}, \eta^{(2)}, \mu^{(2)}) \quad (9)$$

where $\eta_1^{(2)} = r_{\text{min}}, \eta_2^{(2)} = r_{\text{max}}, \mu_1^{(2)} = \alpha, \mu_2^{(2)} = \beta, \mu_3^{(2)} = \gamma,$ and $\mu_4^{(2)} = \kappa$. Here the cut-off function $f_c(r_{ij}, \eta^{(2)})$ is used to ensure the smooth vanishing of the
potential and its derivative for $r_{ij} \geq r_{\text{max}}$. The following cut-off function has been widely used in many interatomic potentials:

$$f_c(r_{ij}, r_{\text{min}}, r_{\text{max}}) = 0.5 \left(1 + \cos \frac{\pi(r_{ij} - r_{\text{min}})}{r_{\text{max}} - r_{\text{min}}} \right).$$  \hspace{1cm} (10)

Here we propose a new cut-off function

$$f_c(r_{ij}, r_{\text{min}}, r_{\text{max}}) = \exp \left(1 - \frac{1}{\sqrt{\left(1 - \frac{(r-r_{\text{min}})^3}{(r_{\text{max}}-r_{\text{min}})^3}\right)^2 + \epsilon}}\right)$$ \hspace{1cm} (11)

where $\epsilon = 10^{-6}$. As shown in Figure 1, the new cut-off function (11) has considerably higher values than (10) on almost the entire interaction distance. As a result, (11) can enable the potentials to capture more atomic interactions than (10). We empirically observe that the cut-off function (11) often yields more accurate results in terms of both energy and force predictions than (10).

Based on the two-body parametrized potential (9), we construct the proper orthogonal descriptors as follows. We assume that we are given $L$ parameter tuples $\mu^{(2)}_\ell, 1 \leq \ell \leq L$. The selection of these parameters is explained in Section 4. We introduce the set of snapshots on $(r_{\text{min}}, r_{\text{max}})$ as

$$\xi_\ell(r_{ij}, \eta^{(2)}_\ell) = V^{(2)}(r_{ij}, \eta^{(2)}_\ell, \mu^{(2)}_\ell), \quad \ell = 1, \ldots, L.$$ \hspace{1cm} (12)

Figure 1: Cut-off functions [10] and [11] as a function of $r_{ij}$ for $r_{\text{min}} = 1\text{Å}$ and $r_{\text{max}} = 5\text{Å}$. 
We note here that these snapshots are not evaluated on any particular set of atomic configurations in a training set; the snapshots are calculated on the interval $r_{\text{min}} \leq r_{ij} \leq r_{\text{max}}$ to ensure adequate sampling of the PES for different parameters and hyperparameters. We assume that the number of snapshots $L$ is sufficiently large to approximate the parametric manifold of potential energies well. Typically, we choose $L$ between 10 and 30. We employ the Karhunen-Loève (KL) expansion [28] to generate an orthogonal basis set which is known to be optimal for representation of the snapshot family $\{\xi_\ell\}_{\ell=1}^L$. The details of the KL expansion are given in Appendix A for reference. In particular, the two-body orthogonal basis functions are computed as follows

$$U^{(2)}_k(r_{ij}, \eta^{(2)}) = \sum_{\ell=1}^L a_{\ell k} \xi_\ell(r_{ij}, \eta^{(2)}), \quad k = 1, \ldots, K, \quad (13)$$

where the matrix components $a_{\ell k}$ consist of eigenvector components of the eigenvalue problem (A.7), which finds the components in the space of snapshots with the largest spatial variability between snapshots. The number of basis functions $K$ is less than $L$, and consist of the largest eigenvalues. Figure 2 shows the orthogonal basis functions as well as the eigenvalues for $r_{\text{min}} = 1\,\text{Å}$ and $r_{\text{max}} = 5\,\text{Å}$, for a family of 40 snapshots. We observe the exponential decay of the eigenvalues, which is why we can take $K < L$. For example, 10 basis functions ($K = 10$) capture more than 99.99% the energy content of the family of 40 snapshots ($L = 40$).

Finally, the two-body proper orthogonal descriptors at each atom $i$ are computed by summing the orthogonal basis functions over the neighbors of atom $i$ as

$$D^{(2)}_{ik}(\eta^{(2)}) = \sum_j U^{(2)}_k(r_{ij}, \eta^{(2)}), \quad 1 \leq i \leq N, 1 \leq k \leq K. \quad (14)$$

Owing to the rapid convergence of the KL expansion, only a small number of descriptors is needed to obtain accurate approximation. We therefore recommend to use no more than 10 two-body proper orthogonal descriptors. This procedure for creating two-body PODs is summarized in Figure 3 with more details on the implementation in Section 4.1. The logic behind the procedure in Figure 3 also applies to three-body PODs, which we describe next.
2.3. Three-body proper orthogonal descriptors

In order to provide proper orthogonal descriptors for three-body interactions, we need to introduce a three-body parametrized potential. In particular, the three-body potential is defined as a product of radial and angular functions as follows

\[ V^{(3)}(r_{ij}, r_{ik}, \theta_{ijk}, \eta^{(3)}, \mu^{(3)}) = \psi(r_{ij}, r_{\text{min}}^{(3)}, r_{\text{max}}^{(3)}, \alpha, \beta, \gamma, \kappa) f_c(r_{ij}, r_{\text{min}}^{(3)}, r_{\text{max}}^{(3)}) \]

\[ \times \psi(r_{ik}, r_{\text{min}}^{(3)}, r_{\text{max}}^{(3)}, \alpha, \beta, \gamma, \kappa) f_c(r_{ik}, r_{\text{min}}^{(3)}, r_{\text{max}}^{(3)}) \]

\[ \cos(\sigma \theta_{ijk} + \zeta) \]  

(15)

where \( \sigma \) is the periodic multiplicity, \( \zeta \) is the equilibrium angle, \( \eta^{(3)} = (r_{\text{min}}^{(3)}, r_{\text{max}}^{(3)}) \), and \( \mu^{(3)} = (\alpha, \beta, \gamma, \kappa, \sigma, \zeta) \). The three-body potential [15] provides an angular fingerprint of the atomic environment through the bond angles \( \theta_{ijk} \) formed with each pair of neighbors \( j \) and \( k \). Compared to the two-body potential [9], the three-body potential [15] has two extra parameters (\( \sigma, \zeta \)) associated with the angular component. Note that \( r_{\text{min}}^{(3)} \) and \( r_{\text{max}}^{(3)} \) are the inner and outer cutoff distances for the three-body potential, respectively.

Let \( \varrho = (\alpha, \beta, \gamma, \kappa) \). We assume that we are given \( L^r \) parameter tuples \( \varrho_\ell, 1 \leq \ell \leq L^r \). We introduce the following set of snapshots on \( (r_{\text{min}}^{(3)}, r_{\text{max}}^{(3)}) \):

\[ \zeta_\ell(r_{ij}, r_{\text{min}}^{(3)}, r_{\text{max}}^{(3)}) = \psi(r_{ij}, r_{\text{min}}^{(3)}, r_{\text{max}}^{(3)}, \varrho_\ell) f_c(r_{ij}, r_{\text{min}}^{(3)}, r_{\text{max}}^{(3)}), \quad 1 \leq \ell \leq L^r. \]  

(16)
Figure 3: Procedure for making two-body proper orthogonal descriptors. First we define the parametric manifold and its parameter domain by specifying the functional forms and their sets of parameters. These sets are created by specifying the bounds $P_\alpha$ and $P_\gamma$ for which parameters $\alpha \in [1, P_\alpha]$ and $\gamma \in [1, P_\gamma]$ take integer values. The parameter $\beta$ can take $P_\beta$ values in the range $[\beta_{\min}, \beta_{\max}]$. We let $\kappa$ be 0 or 1, thus resulting in $L = P_\alpha P_\beta + P_\gamma$ possible 2-body convex functions $\psi(r, \eta, \mu_l) = \kappa \phi(r, \eta, \mu_l) + (1 - \kappa) \varphi(r, \gamma_l)$. 

1. Define the parametric manifold domain, which determines the number of snapshots $L$, e.g.,

$$P_\alpha = 5 \to \phi(r, \alpha) \in \left\{ \sin(x), \sin(2x), \sin(3x), \sin(4x), \sin(5x) \right\}$$

$$P_\beta = 2, \beta_{\min} = 0.01, \beta_{\max} = 1.0 \to x(r, \beta) = \left\{ \frac{\mathrm{exp}[\alpha])(r_{\min})/(r_{\max} - r_{\min})\right\}^{1}, \frac{\mathrm{exp}[\alpha])(r_{\min})/(r_{\max} - r_{\min})\right\}^{1}$$

$$P_\gamma = 12 \to \phi(r, \gamma) \in \left\{ \frac{1}{\gamma_{\max}}, \frac{1}{\gamma_{\max}}, ..., \frac{1}{\gamma_{\max}} \right\}$$

This creates $L = P_\alpha P_\beta + P_\gamma$ possible 2-body convex functions $\psi(r, \eta, \mu_l) = \kappa \phi(r, \eta, \mu_l) + (1 - \kappa) \varphi(r, \gamma_l)$.

2. Calculate the two-body potential as a function of $r$ for each snapshot.

$$\xi_l(r, \eta) = f_l(r, \eta) \psi(r, \eta, \mu_l)$$

3. Calculate the inner product of all pairs of snapshots to form a matrix $C$ with elements:

$$C_{ij} = \frac{1}{L} \left( \xi_i[r, \eta], \xi_j[r, \eta] \right), \ 1 \leq i, j \leq L$$

4. Solve the eigenvalue problem $Ca = \lambda a$ where $a$ is a matrix whose columns are eigenvectors of $C$.

5. Calculate $K < L$ orthogonal basis functions where indices $k \in [1, ..., K]$ include the largest eigenvalues of $C$.

$$U_k(r, \eta) = \sum_i a_{ik} \xi_i(r, \eta)$$

6. Calculate the proper orthogonal descriptors for each atom $l$ by summing over neighbors $j$:

$$D_{lk}(\eta) = \sum_j U_k(r_{lj}, \eta)$$
We apply the Karhunen-Loève (KL) expansion to this set of snapshots to obtain orthogonal basis functions as follows

\[ U^r_k(r_{ij}, r_{\min}^{(3)}, r_{\max}^{(3)}) = \sum_{\ell=1}^{L^r} a^r_{\ell k} \zeta_\ell(r_{ij}, r_{\min}^{(3)}, r_{\max}^{(3)}), \quad k = 1, \ldots, K^r, \]  

(17)

where the matrix components \( a^r_{\ell k} \) consists of eigenvectors of the eigenvalue problem \((A, \zeta)\), and the number of basis functions \( K^r \) consists of the top eigenvalues like with the two-body basis functions. Similarly, we use the KL expansion to construct a set of orthogonal basis functions for the parametrized angular function \( \cos(\sigma \theta_{ijk} + \zeta) \) defined on \((0, \pi)\). We denote the resulting angular basis functions as \( U^a_n(\theta_{ijk}), 1 \leq n \leq K^a \), where \( K^a \) is the number of angular basis functions.

The orthogonal basis functions for the parametrized potential \((15)\) are computed as follows

\[ U^{(3)}_{mn}(r_{ij}, r_{ik}, \theta_{ijk}, \eta^{(3)}) = U^r_m(r_{ij}, \eta^{(3)})U^r_m(r_{ik}, \eta^{(3)})U^a_n(\theta_{ijk}), \]  

(18)

for \( 1 \leq m \leq K^r, 1 \leq n \leq K^a \). Finally, the three-body proper orthogonal descriptors at each atom \( i \) are obtained by summing \((18)\) over the neighbors \( j \) and \( k \) of atom \( i \) as

\[ D^{(3)}_{imn}(\eta^{(3)}) = \sum_{j,k} U^{(3)}_{mn}(r_{ij}, r_{ik}, \theta_{ijk}, \eta^{(3)}), \quad 1 \leq m \leq K^r, 1 \leq n \leq K^a. \]  

(19)

The number of three-body descriptors is \( K^r K^a \). Using 50–100 three-body descriptors is sufficient in most cases. We therefore recommend to use no more than 100 three-body descriptors, because of the rapid convergence of the KL expansion like we saw with the two-body descriptors.

2.4. Linear expansion of the descriptors

All proper orthogonal descriptors are invariant with respect to translation and rotation of the atomic environment because they depend on the atomic distances and angles. Furthermore, because of the sum over all neighbors, they are invariant with respect to any permutation of chemically equivalent atoms in the environment. As a result, they can be used to define an expansion for the atomic energies that are invariant with respect to translation,
rotation, and permutation. The most simple expansion is a linear combination of the descriptors leading to the following expression of the atomic energies

\[ E_i(\eta) = c^{(1)} D^{(1)}_i + \sum_m c^{(2)}_m D^{(2)}_{im}(\eta^{(2)}) + \sum_{mn} c^{(3)}_{mn} D^{(3)}_{imn}(\eta^{(3)}), \quad 1 \leq i \leq N, \quad (20) \]

where \( D^{(1)}_i, D^{(2)}_{im}, D^{(3)}_{imn} \) are the descriptors associated with one-body, two-body three-body parametrized potentials, respectively, and \( c^{(1)}, c^{(2)}_m, c^{(3)}_{mn} \) are their respective expansion coefficients. In a more compact notation that implies summation over descriptor indices the atomic energies \( E_i \) in \((20)\) can be written as

\[ E_i(\eta) = \sum_{m=1}^{M} c_m D_{im}(\eta), \quad 1 \leq i \leq N. \quad (21) \]

Here \( M \) is the total number of descriptors including one-body, two-body, and three-body descriptors. The potential energy is then obtained by summing local atomic energies \( E_i \) for all atoms \( i \) in the system

\[ E(\eta) = \sum_{i=1}^{N} E_i(\eta). \quad (22) \]

Because the descriptors are one-body, two-body, and three-body terms, the resulting potential energy \((22)\) is a three-body PES.

In Section 4, we will describe how to implement an efficient calculation of energy and forces. We will also describe an fitting procedure to determine the expansion coefficients, as well as an optimization formulation to optimize the hyperparameters.

3. Relation to Other Descriptors

In the section we will discuss some of the popular descriptors and the relation among them, as well as their connection to our method described in the previous section.

3.1. Atom-centered symmetry functions and angular Fourier series

The atom-centered symmetry function (ACSF) descriptors of Behler and Parrinello \( [7, 8, 9] \) are derived from the radial function

\[ G^{(2)}_i = \sum_j e^{-\eta(r_{ij}-r_s)^2} f_c(r_{ij}) \quad (23) \]
and the angular function

\[ G_i^{(3)} = 2^{1-\zeta} \sum_{j,k} (1 + \lambda \cos \theta_{ijk})^\zeta \cdot e^{-\eta(r_{ij}^2 + r_{ik}^2 + r_{jk}^2)} f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}) \]  

(24)

where the cut-off function is given by (10). Different values of the parameters \( \eta, r_s, \zeta, \lambda \) can be used to generate the ACSF descriptors. Similar to the ACSF descriptors (24), the angular Fourier series (AFS) descriptors are formed by using the Chebyshev polynomials of the angular component as \( \text{AFS}_{i}^{(3)l} = g_n(r_{ij})g_n(r_{ik}) \cos(l\theta_{ijk}) \), where \( g_n(r) \) are radial basis functions introduced in [21].

Both ACSF and AFS are related to our method. However, because ACSF basis functions are generated directly from snapshots of the symmetry functions (23) and (24), they are not orthogonal. Moreover, both ACSF and AFS do not make use of the KL expansion to reduce the number of basis functions.

3.2. Power spectrum and bispectrum descriptors

The starting point for constructing the power spectrum and bispectrum descriptors is to consider the neighborhood density at location \( r \) around a central atom \( i \):

\[ \rho_i(r) = \sum_j \delta(r - r_{ij}), \]  

(25)

where \( \delta \) is the Dirac-delta function. Expanding the neighbor density in terms of the spherical harmonics \( Y_{lm}(\hat{r}) \) and radial basis functions \( g_k(r) \) yields

\[ \rho_i(r) = \sum_{k=0}^{l} \sum_{l=0}^{l} \sum_{m=-l}^{l} c_{iklm} g_k(r) Y_{lm}(\hat{r}), \]  

(26)

where the expansion coefficients are given by the inner products of the neighborhood density with \( g_k(r) Y_{lm}(\hat{r}) \) and expressed as

\[ c_{iklm} = \sum_j g_k(r_{ij}) Y_{lm}(\hat{r}_{ij}). \]  

(27)

The power spectrum and bispectrum descriptors are obtained as follows

\[ p_{ikl_{1}l_{2}l_{3}} = \sum_{m=-l}^{l} (c_{ik1m})^* c_{ik2lm}, \]  

\[ b_{ikl_{1}l_{2}l_{3}l_{4}} = \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} \sum_{m_3=-l_3}^{l_3} (c_{ik1m_1})^* C_{m_1m_2m_3}^{l_1l_2l_3} c_{ik2m_2} c_{ik3m_3}, \]  

(28)
where $C_{m_1m_2m_3}^{l_1l_2l_3}$ are the Clebsch-Gordan coefficients and the asterisk denotes the complex conjugation [21].

The power spectrum descriptors are three-body, while the bispectrum descriptors are four-body. An advantage of the power spectrum and bispectrum descriptors is that their evaluation scales linearly with the number of neighbors. However, they require computing a number of inner products and triple products. As the power spectrum descriptors can be expressed as products of polynomials of $\cos(\theta_{ijk})$ and radial basis functions, they are related to the ACFS, AFS, and POD descriptors.

3.3. Smooth overlap of atomic positions

Instead of the Dirac-delta functions, the SOAP method [21] constructs the neighborhood density using Gaussians expanded in terms of spherical harmonic functions as follows

$$
\rho_i(r) = \sum_j \exp(-\alpha |r - r_{ij}|^2) = \sum_{klm} c_{iklm} g_k(r) Y_{lm}(\hat{r}).
$$

(29)

The expansion coefficients are inner products of the neighborhood density with $g_k(r) Y_{lm}(\hat{r})$ and can be computed by noting that

$$
\exp(-\alpha |r - r_{ij}|^2) = 4\pi \exp(-\alpha (r^2 + r_{ij}^2)) \sum_{lm} I_l(2\alpha rr_{ij}) Y_{lm}(\hat{r}) Y_{lm}(\hat{r}_{ij})
$$

(30)

where $I_l$ is a modified spherical Bessel function of the first kind.

The SOAP descriptors are equivalent to using the power or bispectrum descriptors together with Gaussian atomic neighbor density contributions and a dot product covariance kernel. The Gaussian width $\alpha$ controls the smoothness of the similarity measure between atomic neighbor environments. The ability to controlling the smoothness of the similarity measure is an important advantage of SOAP over the power or bispectrum descriptors. SOAP has been used for the construction of a variety of accurate and transferable interatomic potentials [31, 32, 33, 34].

3.4. Spectral neighbor analysis potential

The SNAP descriptors [13] are constructed from hyperspherical harmonics on the unit 3-sphere. A 3-dimensional vector $r$ is mapped onto the unit 3-sphere $S^3$ defined by three angles $(\phi, \theta, \omega)$ as follows

$$
\phi = \arctan(y/x), \quad \theta = \arccos(z/|r|), \quad \omega = \pi |r|/r_0.
$$

(31)
In [21] it is recommended to use \( r_0 = \frac{4}{3} r_{\text{cut}} \) to cover most of the sphere surface, while still excluding the region near the south pole. The Laplace–Beltrami operator on \( S^3 \) is defined as

\[
\Delta_{S^3} = \frac{1}{\sin^2 \omega} \frac{\partial}{\partial \omega} \left( \sin^2 \omega \frac{\partial}{\partial \omega} \right) + \frac{1}{\sin^2 \omega} \Delta_{S^2},
\]

(32)

where \( \Delta_{S^2} \) is the Laplace–Beltrami operator on the unit 2-sphere \( S^2 \). The eigenfunctions of (32) are the hyperspherical harmonics

\[
U_{jmm'}(\omega, \phi, \theta) = (\sin \omega)^m G_{j-m}^{m+1}(\cos \omega) Y_{mm'}(\theta, \phi)
\]

(33)

where \( G_{j-m}^{m+1} \) are the Gegenbauer polynomials, and \( Y_{mm'} \) are the usual spherical harmonics. These hyperspherical harmonics are also known as the Wigner matrices [13].

To make the contribution from atoms at \( r = r_{\text{cut}} \) vanish smoothly to zero, it is necessary to augment the atomic neighbor density function (25) with a cutoff function

\[
\rho_i(r) = \delta(r) + \sum_j f_c(r_i) \delta(r - r_{ij}).
\]

(34)

This atomic neighborhood density is expanded in terms of hyperspherical harmonics on the unit 3-sphere as

\[
\rho_i(r) = \sum_{j=0, 1/2, \ldots}^\infty \sum_{m=-j}^j \sum_{m'=-j}^j c_{ijmm'} U_{jmm'}(\omega, \phi, \theta)
\]

(35)

where

\[
c_{ijmm'} = U_{jmm'}(0, 0, 0) + \sum_j f_c(r_{ij}) U_{jmm'}(\omega_{ij}, \phi_{ij}, \theta_{ij}).
\]

(36)

The coefficients \( c_{ijmm'} \) are not directly useful as descriptors, because they are not invariant under rotation of the reference frame. However, the following bispectrum components are invariant with respect to permutation, translation and rotation

\[
b_{ij_1j_2j_3} = \sum_{m_1, m_1' = -j_1}^{j_1} \sum_{m_2, m_2' = -j_2}^{j_2} \sum_{m_3, m_3' = -j}^{j} (c_{ijmm'})^* H_{j_{ij_1}m_1m_1'}^{jj'}_{j_{ij_2}m_2m_2'} c_{ij_3m_3m_3'} c_{ij_2m_2m_2'},
\]

(37)

where the coupling constants are related to the Clebsch–Gordan coefficients as follows

\[
H_{j_{ij_1}m_1m_1'}^{jj'}_{j_{ij_2}m_2m_2'} = C_{jj_1j_2}^{mm_1m_2} C_{jjj_2}^{m'm_3m_3'}.
\]

(38)
We see from (33) that the hyperspherical harmonics can be expressed as a radial function time a spherical harmonic $U_{jm\ell}(\omega, \phi, \theta) = g_k(r)Y_{m\ell}(\theta, \phi)$ with $g_k(r) = (\sin \omega)^m G_{j-m}^{m+1}(\cos \omega)$, where the index $k$ has a one-to-one correspondence to $(j, m)$. As a result, SNAP can be viewed as a variant of the bispectrum and SOAP descriptors. Recently, SNAP has been extended to quadratic SNAP [14] and multi-element explicit SNAP [35].

3.5. Atomic cluster expansion

The atomic cluster expansion (ACE) is recently developed in [17, 18] as a hierarchical and complete descriptor of the local atomic environment due to its arbitrary number of body orders. This method extends the power and bispectrum construction to obtain a complete set of invariant descriptors. Similar to (27), the expansion coefficients in the ACE method are given by

$$A_{in\ell m} = \sum_j R_{nl}(r_{ij})Y_{lm}(\hat{r}_{ij})$$  \hspace{1cm} (39)

where the radial basis functions $R_{nl}(r)$ are constructed from the Chebyshev polynomials of the first kind and the cosine cut-off function. The ACE descriptors are then computed as

$$B_{in}^{(1)} = A_{in00},$$  \hspace{1cm} (40)$$

$$B_{in1n2l}^{(2)} = \sum_{m=-l}^{l} (-1)^m A_{in1lm} A_{in2l-m},$$  \hspace{1cm} (41)$$

$$B_{in1n2n3}^{(3)} = \sum_{l_1}^{l_1} \sum_{l_2}^{l_2} \sum_{l_3}^{l_3} W_{m1m2m3}^{l_1l_2l_3} A_{in1l_1m_1} A_{in2l_2m_2} A_{in3l_3m_3},$$  \hspace{1cm} (42)$$

$$B_{in1n2n3n4}^{(4)} = \sum_{m1m2m3m4} W_{m1m2m3m4}^{l_1l_2l_3l_4} A_{in1l_1m_1} A_{in2l_2m_2} A_{in3l_3m_3} A_{in4l_4m_4},$$  \hspace{1cm} (43)$$

which are in that order the two-body, three-body, four-body, and five-body descriptors, respectively. The coupling constants $W_{m1m2m3}^{l_1l_2l_3}$ and $W_{m1m2m3m4}^{l_1l_2l_3l_4}$ are calculated from the Clebsch–Gordan coefficients. The paper [17] describes a systematic approach to defining many-body ACE descriptors.

The ACE method has a number of attractive features. First, the method encompasses several well-known descriptors since ACSF, SOAP, power spectrum, bispectrum, and SNAP can be expanded using the ACE descriptors.
Second, the method is complete in the sense that it is possible to converge cluster functional to arbitrary accuracy. And third, the ACE descriptors scale linearly with the number of neighbors irrespective of the order of the expansion. The ACE method has been analyzed in [19] and applied to many different problems [36, 37, 38].

4. Implementation

In this section, we describe the implementation of the proper orthogonal descriptors and the potential fitting procedure.

4.1. Precomputation of the eigenvectors and eigenvalues

To compute the eigenvectors in (13), we need to define the parameter space and choose $L$ parameter points $\mu_\ell^{(2)} = (\alpha_\ell, \beta_\ell, \gamma_\ell, \kappa_\ell), 1 \leq \ell \leq L$. The parameters $\alpha \in [1, P_\alpha]$ and $\gamma \in [1, P_\gamma]$ are integers, where $P_\alpha$ and $P_\gamma$ are the highest degrees for $\alpha$ and $\gamma$, respectively. We recommend to uniformly choose $P_\beta$ different values of $\beta$ in the interval $[\beta_{\min}, \beta_{\max}]$, where $\beta_{\min} = 0$ and $\beta_{\max} = 5$. The parameter $\kappa$ can be set either 0 or 1. Hence, the total number of parameter points is $L = P_\alpha P_\beta + P_\gamma$. Although $P_\alpha$, $P_\beta$, $P_\gamma$ can be chosen conservatively large, we recommend to choose $P_\alpha \leq 8$, $P_\beta \leq 4$, $P_\gamma \leq 12$. In many cases, $P_\alpha = 6$, $P_\beta = 3$, $P_\gamma = 8$ are adequate. Once the parameter points are chosen, we compute the inner products of pairs of snapshots, $\{\xi_\ell(r, r_{\min}, r_{\max})\}_{\ell=1}^L$ to form the eigenvalue problem [A.7]. At this point, $r_{\min}$ and $r_{\max}$ are assumed to be fixed. As discussed later in section 4.4, both $r_{\min}$ and $r_{\max}$ are found by solving an outer optimization problem.

The eigenvectors in (17) can be computed using the same procedure as described above. Finally, we compute the eigenvectors for the angular component $\cos(\sigma \theta + \zeta)$ on the domain $\theta \in (0, \pi)$ in a similar manner. Typically, we choose $\sigma$ to be integers in $[1, P_\sigma]$ while fixing $\zeta = 0$, where $P_\sigma$ is the highest degree for $\sigma$. We recommend to choose $P_\sigma \leq 10$ in most cases.

4.2. Calculation of energy and forces

The radial basis functions in (13) are first calculated. Owing to the exponential convergence of the KL expansion, the number of basis functions $K$ is considerably smaller than the number of snapshots $L$. Therefore, the two-body descriptors (14) can be computed by summing the radial basis functions over neighbors of every atom $i$. Next, we compute the radial and angular basis functions, which are then used to compute the three-body
basis functions (18). Finally, the three-body descriptors (19) are obtained by summing the three-body basis functions over neighbors $j$ and $k$ of every atom $i$.

We can now calculate the local atomic energies for the linear expansion of the descriptors (21). The potential energy is then obtained by summing the atomic energies as

$$ E(\eta) = \sum_{m=1}^{M} c_m \sum_{i=1}^{N} D_{im}(\eta) = \sum_{m=1}^{M} c_m d_m(\eta), \quad (44) $$

where $d_m(\eta) = \sum_{i=1}^{N} D_{im}(\eta)$ are the global descriptors. To calculate forces, we note that

$$ F(\eta) = -\nabla E(\eta) = -\sum_{m=1}^{M} c_m \sum_{i=1}^{N} \nabla D_{im}(\eta) = -\sum_{m=1}^{M} c_m \nabla d_m(\eta), \quad (45) $$

where $\nabla d_m(\eta)$ are derivatives of the global descriptors with respect to atom positions. For the two-body global descriptors, their derivatives are computed from the derivatives of the radial basis functions (13) with respect to $r_{ij}$ for all neighbor pairs $i,j$. For the three-body global descriptors, their derivatives are computed from the derivatives of the radial basis functions with respect to $r_{ij}$ and $r_{ik}$, and the derivatives of the angular basis functions with respect to $\theta_{ijk}$.

### 4.3. Weighted least-squares regression

With the availability of high-performance computers and highly optimized DFT codes, it is not difficult to generate DFT data for hundreds or thousands of configurations of atoms. Let $J_{\text{train}}$ be the total number of training configurations. Let $\{E^\text{DFT}_j\}_{j=1}^{J_{\text{train}}}$ and $\{F^\text{DFT}_j\}_{j=1}^{J_{\text{train}}}$ be the DFT energies and forces for $J_{\text{train}}$ configurations. Next, we calculate the global descriptors and their derivatives for all training configurations. We then form the matrix $A \in \mathbb{R}^{J_{\text{train}} \times M}$, where the $j$th row of $A$ consists of the global descriptors for the $j$th configuration. Similarly, we form the matrix $B \in \mathbb{R}^{N \times M\times J_{\text{train}}}$ by stacking the derivatives of the global descriptors for all training configurations from top to bottom, where $N = 3\sum_{j=1}^{J_{\text{train}}} N_j$ and $N_j$ is the number of atoms for the $j$th configuration. Note that both $A$ and $B$ depends on $\eta$.

When the hyperparameters $\eta$ are known, the coefficient vector $c$ of the linear expansion (21) is found by solving the following least-squares problem

$$ \min_{c \in \mathbb{R}^M} \beta E \| A(\eta)c - E^\text{DFT} \|^2 + \beta F \| B(\eta)c - F^\text{DFT} \|^2, \quad (46) $$

19
where $\beta^E$ is the energy weight and $\beta^F$ is the force weight. Here $E^{\text{DFT}}$ is a vector of $J_{\text{train}}$ DFT energies and $F^{\text{DFT}}$ is a vector of $N_J$ DFT forces. In our implementation, we fix $\beta^F = 1$ and let $\beta^E$ be the only scalar weight to be dependent on specific problems.

4.4. Optimization of hyperparameters

The hyperparameters contain inner and outer cut-off distances for two-body and three-body interactions. They can have a critical effect on the prediction performance of the resulting potential. We observe through our numerical experiences that longer cut-off distances do not necessarily translate to better performance. Furthermore, setting them to some heuristic values may lead to an interatomic potential that can be considerably less accurate than the potential obtained by optimizing the cut-off distances. Therefore, it is crucial to optimize the hyperparameters.

Let $\Omega^\eta$ be a domain in which the hyperparameters reside. We consider solving the following optimization problem

$$
\min_{\eta \in \Omega^\eta} w \| A(\eta)c(\eta) - E^{\text{DFT}} \|_{\text{MAE}} + (1 - w) \| B(\eta)c(\eta) - F^{\text{DFT}} \|_{\text{MAE}}
$$

(47)

where $c(\eta)$ is the solution of the least-squares problem (46), $w \in [0, 1]$ is a given scalar weight, and $\| \cdot \|_{\text{MAE}}$ is defined as

$$
\| e \|_{\text{MAE}} = \frac{1}{n} \sum_i |e_i|.
$$

(48)

The problem (47) is an outer optimization problem, while (46) is an inner optimization problem.

Instead of directly solving the problem (47), we solve a surrogate problem. To this end, we approximate the loss function in (47) using polynomial tensor-product interpolation. In particular, we divide the domain $\Omega^\eta$ into $Q$ grid points $\{ \eta_q \}_{q=1}^Q$ and compute the loss function at those grid points. This requires us to solve the least-squares problem (47) $Q$ times for $\eta = \eta_q, q = 1, \ldots, Q$. We then fit the loss values at the grid points to a tensor-product multivariate polynomial of certain degree to construct a surrogate model of the loss function. Finally, the gradient descent algorithm is used to minimize the surrogate model to obtain the hyperparameters. In order to be able to reach the global minimum, we start our gradient descent algorithm with a large number of initial guesses randomly sampled in $\Omega^\eta$. This is necessary because the loss function (47) may have multiple local minima.
5. Numerical Experiments

In this section, we assess the performance of the ACE, SNAP, as well as POD descriptors for fitting interatomic potentials on previously published data sets for many different kinds of elements. In all cases, the weighted least-square regression \( \beta_E = 100 \) and \( \beta_F = 1 \) for the various descriptors considered here. For the ACE and POD methods, both the inner and outer cut-off distances are optimized by solving a polynomial interpolation surrogate model of the problem \( w = 0.8 \) and 100 randomly sampled initial guesses. For the SNAP method, however, only the outer cut-off radius is optimized. We shall evaluate these methods on the basis of mean absolute errors of the predicted energy and forces versus the number of descriptors. To facilitate the reproduction of our work, the code and the results presented in this section are made available as an open-source software on Github (https://github.com/cesmix-mit/mlp).

5.1. Tantalum element

The Ta data set contains a wide range of configurations to adequately sample the important regions of the potential energy surface [13]. The data set includes 12 different groups such as surfaces, bulk structures, defects, elastics for BCC, FCC, and A15 crystal structures, and high temperature liquid. Displaced configurations were constructed by randomly displacing atoms from their equilibrium lattice sites in supercells of the A15, BCC, and FCC crystal structures. Elastic configurations were constructed by applying random strains to primitive cells of the BCC and FCC crystals. The configurations of type “GSF” consist of both relaxed and unrelaxed generalized stacking faults along the [110] and [112] crystallographic directions. Liquid configurations were taken from a high-temperature quantum molecular dynamics simulations of molten tantalum. Surface configurations consisted of relaxed and unrelaxed [100], [110], [111], and [112] BCC surfaces. Volume configurations consisted of primitive cells of A15, BCC, and FCC crystal structures, compressed or expanded isotropically over a wide range of densities. The database was used to create a SNAP potential [13] which successfully describes a wide range of properties such as energetic properties of solid tantalum phases, the size and shape of the Peierls barrier for screw dislocation motion in BCC tantalum, as well as both the structure of molten tantalum and its melting point.
Herein the same database is used to fit ACE, POD, and SNAP potentials. We associate a weight of 100.0 with energy, 1.0 with force, and 0.0 with stress for all configurations in the database. This is different from the SNAP potential presented in [13], that specified different weights for each type of configuration. Furthermore, while the Ziegler–Biersack–Littmark (ZBL) empirical potential was used as the reference potential for the SNAP potential in [13], no reference potential is used for ACE, POD, and SNAP potentials in this paper. The SNAP potential constructed in this paper has considerably smaller energy errors than the one in [13], while having similar force errors. In particular, the MAE energy error of our SNAP potential can be close to 1 meV/Atom, while that of the SNAP potential in [13] is about 100 meV/atom.

| M  | r_{min} | r_{max} | M  | r_{min} | r_{max} | M  | r_{max} |
|----|---------|---------|----|---------|---------|----|---------|
| 6  | 1.29    | 4.46    | 6  | 1.04    | 4.65    | 6  | 5.00    |
| 16 | 0.92    | 4.42    | 16 | 1.42    | 4.49    | 15 | 4.38    |
| 32 | 1.19    | 4.53    | 32 | 1.14    | 5.00    | 31 | 4.33    |
| 57 | 1.48    | 4.50    | 57 | 0.90    | 4.68    | 56 | 4.61    |
| 93 | 0.99    | 4.43    | 92 | 0.90    | 4.75    | 92 | 4.64    |
| 142| 0.98    | 5.00    | 143| 0.94    | 4.84    | 141| 4.60    |

Table 1: Optimized inner and outer cut-off distances of the ACE, POD, and SNAP potentials as a function of the number of descriptors for the Ta element.

Table 1 provides the optimized cutoff distances of the ACE, POD, and SNAP potentials for several different values of the number of descriptors. The optimized cutoff distances for different potentials appear quite similar. In all cases, the optimized cutoff distances fall between the second nearest neighbor (2NN) and 3NN distances of Ta element. These observations are consistent with the previous SNAP potential [13] for Ta element, which has the cutoff radius of 4.67 Å. While the optimized cutoff distances vary with the number of descriptors, the variation appears not monotonic. Figure 4 shows the polynomial interpolation surrogates of the loss function for ACE and POD potentials. We see that in both cases the surrogate loss functions have multiple local minima. Nonetheless, the surrogate loss functions are smooth. Using 100 randomly sampled initial guesses in the gradient descent algorithm, we can find the global minimum in both cases, which are indicated by the red circles in Figure 4. The results also show significant influence of
the cutoff distances on the value of the loss function and thus on the accuracy of the resulting potential. Therefore, it is crucial to rigorously optimize these hyperparameters since heuristic means of choosing the hyperparameters may lead to suboptimal performance.

Figure 4: Surrogate loss function of ACE and POD potentials as a function of inner and outer cutoff distances for the Ta element. The red circles indicate the optimal value of the cutoff distances for which the surrogate loss function is minimum.

(a) ACE potential with $M = 57$ descriptors. (b) POD potential with $M = 143$ descriptors.

Figure 5 depicts training errors in energies and forces predicted using ACE, SNAP, POD potentials for different values of the number of the descriptors. The same results are also tabulated in Table 2 for quantitative comparison of the potentials. We observe fast convergence of the energy and force errors as the number of descriptors increases from 16 to 60. However, the convergence becomes slower as the number of descriptors increases above 60. Both POD and SNAP have very similar convergence of the energy error and yield faster convergence of the energy error than ACE. The convergence of the force error is faster for POD than ACE and SNAP, as POD has smaller force errors than ACE and SNAP in most cases. The MAE of the energy reaches about 1.07 meV/atom for POD with $M = 143$, 1.27 meV/atom for SNAP with $M = 141$, and 3.21 meV/atom for ACE with $M = 142$.

Table 3 provides the training errors in energy and forces for each of the 12 groups. The Surface group has the the highest mean absolute error in energy, while the Liquid group has the highest mean absolute error in forces. The liquid structures depend strongly on the repulsive interactions that occur when two atoms approach each other. Consequently, it is more difficult for potential models to predict forces of the liquid structures since the liquid
Figure 5: Energy errors in log scale and force errors in linear scale for ACE, SNAP, and POD potentials as a function of the number of descriptors for the Ta element.

| Configuration | ACE MAE(e) | ACE MAE(f) | SNAP MAE(e) | SNAP MAE(f) | POD MAE(e) | POD MAE(f) |
|---------------|------------|------------|-------------|-------------|------------|------------|
| 6             | 106.52     | 195.51     | 225.13      | 415.27      | 101.14     | 228.50     |
| 16            | 29.91      | 231.13     | 27.91       | 206.63      | 27.26      | 137.34     |
| 32            | 16.83      | 153.45     | 9.16        | 159.89      | 9.16       | 121.93     |
| 57            | 5.91       | 100.75     | 3.94        | 114.01      | 3.38       | 92.83      |
| 93            | 8.29       | 100.39     | 2.27        | 85.06       | 1.70       | 69.59      |
| 142           | 3.21       | 87.16      | 1.27        | 67.21       | 1.07       | 58.91      |

Table 2: Training errors in energies and forces for ACE, SNAP, and POD potentials as a function of the number of descriptors for the Ta element. The units of the mean absolute errors of the energy and force are meV/atom and meV/Å, respectively.

configurations are very different from those of the equilibrium solid crystals. It is also more difficult to predict energies of surface configurations because the surfaces of BBC crystals tend to be rather open with surface atoms exhibiting rather low coordination numbers. The MAE force errors for Volume A15, Volume BCC, and Volume FCC are zero, because all of the structures in Volume A15, Volume BCC, and Volume FCC are at equilibrium states and thus have zero atomic forces. Figure 5 plots the energy per atom computed with DFT, ACE, SNAP, and POD as a function of volume per atom for the BCC and FCC crystals. We see that the energies predicted using ACE, SNAP, POD potentials are extremely accurate.
Table 3: Energy and force errors obtained using ACE, SNAP, POD potentials with the largest number of descriptors for different training groups for the Ta element. The units of the mean absolute errors of the energy and force are meV/atom and meV/Å, respectively.

| Training group     | ACE MAE(e) | ACE MAE(f) | SNAP MAE(e) | SNAP MAE(f) | POD MAE(e) | POD MAE(f) |
|--------------------|------------|------------|-------------|-------------|------------|------------|
| Displaced A15      | 1.542      | 99.06      | 0.826       | 76.16       | 0.432      | 57.63      |
| Displaced BCC      | 9.211      | 148.4      | 2.365       | 104.9       | 0.879      | 100.4      |
| Displaced FCC      | 4.509      | 98.47      | 0.261       | 49.66       | 0.708      | 40.05      |
| Elastic BCC        | 0.600      | 47.19      | 0.243       | 32.80       | 0.219      | 42.76      |
| Elastic FCC        | 0.325      | 39.78      | 0.280       | 36.84       | 0.350      | 41.14      |
| GSF 110            | 4.988      | 32.33      | 2.252       | 23.33       | 1.870      | 23.60      |
| GSF 112            | 8.051      | 74.08      | 2.915       | 45.56       | 2.882      | 45.57      |
| Liquid             | 8.940      | 312.2      | 3.558       | 304.3       | 1.113      | 239.6      |
| Surface            | 12.39      | 48.30      | 9.713       | 53.35       | 10.68      | 42.66      |
| Volume A15         | 4.718      | 0.000      | 2.689       | 0.000       | 2.113      | 0.000      |
| Volume BCC         | 7.676      | 0.000      | 2.336       | 0.000       | 1.087      | 0.000      |
| Volume FCC         | 7.539      | 0.000      | 1.775       | 0.000       | 1.444      | 0.000      |

Figure 6: Energy per atom versus volume per atom for BCC and FCC crystal structures of the Ta element computed using DFT and ACE, SNAP, POD potentials with the largest number of descriptors.
5.2. Li, Mo, Ni, Cu, Si and Ge elements

In the recent paper [20], a diverse DFT data set was generated for Li, Mo, Ni, Cu, Si and Ge elements to evaluate and compare the performance of machine learning potentials based on four different descriptors. These elements were chosen to represent a variety of chemistries (main group metal, transition metal and semiconductor), crystal structures (bcc, fcc, and diamond) and bonding types (metallic and covalent). For each element, a set of configurations with diverse coverage of atomic local environment space were generated. For each element, the full data set were divided into a training set and a test set in order to assess the ML potentials on both training and test configurations. We refer to [20] for additional details about the DFT data set, as well as insightful results on the performance and cost assessment of the various ML potentials studied therein.

We use ACE, SNAP, and POD descriptors to fit interatomic potentials for Li, Mo, Ni, Cu, Si and Ge elements. The energy errors on the test sets as a function of the number of descriptors are shown in Figure 7. We observe that the energy errors drop quickly when the number of descriptors increases up to 60. When the number of descriptors increases above 60, the convergence becomes slower. These observations are similar to those of the Ta element presented earlier. To quantitatively compare the descriptors, we present in Table 4 the energy errors for three different values of the number of descriptors for both the training and test sets. The training errors are comparable to the testing errors across different descriptors, indicating no over-fitting for the potentials. The POD potentials generally have lower MAEs in energies than ACE and SNAP potentials for Ni, Cu, Li, Mo elements. The ACE potentials generally have lower energy errors than the SNAP and POD potentials for Ge element. All the potentials show extremely good performance for Ni, Cu, and Li elements, reaching about 1 meV/atom or less for the MAE of the energy. They have about 5 meV/atom in the MAE of the energy for Mo, Si, Ge elements, which are generally below the limits of DFT error.

Figure 8 depicts the force errors as a function of the number of the descriptors for the test sets, while Table 5 tabulates the force errors for three different values of the number of descriptors for both the training and test sets. We note that the POD potentials have smaller force errors than the ACE and SNAP potentials across all elements for $M \leq 60$. In general, POD outperforms both ACE and SNAP when the number of descriptors is smaller than 60. However, ACE and SNAP achieve similar errors to POD when the number of descriptors is large enough. Again, all the potentials exhibit low
Figure 7: Test errors in energies predicted using ACE, SNAP, and POD potentials as a function of the number of descriptors for Ni, Cu, Li, Mo, Si, Ge elements.

| Descriptors | Data set | Ni     | Cu     | Li     | Mo     | Si     | Ge     |
|-------------|----------|--------|--------|--------|--------|--------|--------|
| ACE         | Training | 2.374  | 1.718  | 1.198  | 13.52  | 18.28  | 11.81  |
| M = 16      | Test     | 2.453  | 1.754  | 1.199  | 12.19  | 21.02  | 12.71  |
| SNAP        | Training | 3.903  | 3.633  | 2.107  | 10.37  | 10.92  | 10.23  |
| M = 15      | Test     | 4.704  | 5.234  | 2.783  | 11.57  | 9.428  | 10.03  |
| POD         | Training | 1.482  | 1.143  | 0.754  | 11.16  | 12.27  | 8.745  |
| M = 16      | Test     | 1.784  | 1.338  | 0.931  | 11.87  | 11.77  | 8.972  |

| Descriptors | Data set | Ni     | Cu     | Li     | Mo     | Si     | Ge     |
|-------------|----------|--------|--------|--------|--------|--------|--------|
| ACE         | Training | 0.970  | 0.779  | 0.721  | 8.626  | 8.953  | 6.819  |
| M = 57      | Test     | 0.970  | 0.654  | 0.894  | 8.483  | 6.761  | 5.693  |
| SNAP        | Training | 1.128  | 1.001  | 0.818  | 5.510  | 4.657  | 7.916  |
| M = 56      | Test     | 1.339  | 0.824  | 0.909  | 6.314  | 6.170  | 8.835  |
| POD         | Training | 0.699  | 0.427  | 0.601  | 4.110  | 5.948  | 6.759  |
| M = 57      | Test     | 0.746  | 0.359  | 0.607  | 4.699  | 5.425  | 6.526  |

Table 4: Mean absolute errors in energies (meV/atom) predicted using the ACE, SNAP, POD methods for the training and test sets of Ni, Cu, Li, Mo, Si, Ge elements.
Figure 8: Test errors in forces predicted using the ACE, SNAP, and POD methods as a function of the number of descriptors for Ni, Cu, Li, Mo, Si, Ge elements.

test errors on forces (on the order of 10 meV/Å) for the Ni, Cu, and Li datasets. Significantly higher force errors (on the order of 100 meV/Å) are observed for the Mo, Si, and Ge datasets.

In terms of chemistries, we find that the lowest MAEs in energies and forces are observed for fcc crystal structures (Ni and Cu elements), followed by bcc structures (Li and Mo elements), and that the highest MAEs are observed for diamond structures (Si and Ge elements). These observations are consistent with the results reported in [20].

6. Conclusion

We have introduced a new method called proper orthogonal descriptors (POD) to generate invariant descriptors that are appropriate for fitting potential energy surfaces. The method is based on the KL expansion of the radial and angular components of the parametrized potentials. In our work, the parametrized potentials are designed to provide a rich and diverse representation of two-body and three-body interactions. Owing to the exponential convergence of the KL expansion, the number of descriptors needed to reach convergence in energy and force errors is often small.

We have evaluated the performance of POD in light of recently developed
Table 5: Mean absolute errors in forces (meV/Å) predicted using the ACE, SNAP, POD methods for the training and test sets of Ni, Cu, Li, Mo, Si, Ge elements.

| Descriptors | Data set | Ni    | Cu    | Li    | Mo    | Si    | Ge    |
|-------------|----------|-------|-------|-------|-------|-------|-------|
| ACE         | Training | 37.90 | 21.34 | 23.39 | 234.9 | 205.7 | 162.9 |
| M = 16      | Test     | 41.63 | 24.00 | 24.57 | 229.5 | 230.5 | 166.5 |
| SNAP        | Training | 71.98 | 71.09 | 38.96 | 306.6 | 146.2 | 102.1 |
| M = 15      | Test     | 89.48 | 86.80 | 41.55 | 301.2 | 162.9 | 104.2 |
| POD         | Training | 26.91 | 16.46 | 19.12 | 211.1 | 127.8 | 99.90 |
| M = 16      | Test     | 29.84 | 19.88 | 20.63 | 209.4 | 138.2 | 101.2 |
| ACE         | Training | 21.93 | 10.39 | 16.94 | 185.8 | 115.2 | 97.38 |
| M = 57      | Test     | 25.10 | 11.75 | 17.87 | 188.6 | 126.1 | 100.8 |
| SNAP        | Training | 27.23 | 22.82 | 20.65 | 143.9 | 98.77 | 85.17 |
| M = 56      | Test     | 32.14 | 26.02 | 22.30 | 148.3 | 109.0 | 86.48 |
| POD         | Training | 16.73 | 8.355 | 11.80 | 120.3 | 91.67 | 80.06 |
| M = 57      | Test     | 19.18 | 9.671 | 12.46 | 120.6 | 102.2 | 82.71 |
| ACE         | Training | 15.85 | 7.577 | 12.99 | 149.4 | 81.72 | 69.32 |
| M = 142     | Test     | 18.01 | 8.488 | 14.04 | 155.9 | 92.94 | 72.97 |
| SNAP        | Training | 21.44 | 14.69 | 16.11 | 118.8 | 85.78 | 74.69 |
| M = 141     | Test     | 25.66 | 17.73 | 17.60 | 125.3 | 97.11 | 74.63 |
| POD         | Training | 15.01 | 7.676 | 11.61 | 107.1 | 82.94 | 72.71 |
| M = 143     | Test     | 17.67 | 8.875 | 12.54 | 109.2 | 94.54 | 76.25 |

state-of-the-art methods such as SNAP and ACE descriptors for many different elements spanning different crystal structures (fcc, bcc, and diamond), chemistries (main group metals, transition metals and semiconductors) and bonding (metallic and covalent). This evaluation is carried out on the basis of mean absolute errors of the predicted energies and forces versus the number of descriptors.

For all the three methods, increasing the number of descriptors leads to higher accuracy in predicting energies and forces. The convergence in energy and force prediction is fast as the number of descriptors increases from 10 to 60. When the number of descriptors is above 60, the convergence appears somewhat slower. We observe that the errors for Ni, Cu, and Li elements are considerably smaller than those for Mo, Si, and Ge elements. For Ta element, the MAE in energies reaches 3.21, 1.27, and 1.07 meV/atom for ACE, SNAP, and POD, respectively. POD exhibits smaller MAEs in energies and forces than ACE and SNAP in a majority of cases. Nonetheless, we find that all
methods are able to achieve near-DFT accuracy in predicting energies and forces for all elements studied.

We find that the performance of a fitted potential can be improved by optimizing the hyperparameters. To this end, we have formulated a bilevel minimization problem in which the actual loss function is replaced with a polynomial interpolation surrogate. The gradient descent algorithm is used with randomly sampled initial guesses to minimize the surrogate loss function to obtain the hyperparameters and the expansion coefficients.

The ideas presented in this paper can be extended to both empirical and machine learning potentials. The two-body orthogonal basis functions developed herein can be used as radial basis functions in power spectrum, bispectrum, SOAP and ACE descriptors. Moreover, instead of fitting the parameters of empirical potentials, one can use the proposed method to construct an orthogonal basis for the parametric manifold of empirical potentials. We believe that this approach can improve not only the accuracy but also the transferability of many existing empirical potentials.

Acknowledgements

We would like to thank all members of the CESMIX center at MIT for fruitful and invaluable discussions leading to the ideas presented in this work. We want to thank Jaime Peraire, Youssef Marzouk, Nicolas Hadjiconstantinou, Dionysios Sema, William Moses, Jayanth Mohan, Dallas Foster, Valentin Churavy, Mathew Swisher for the many discussions we have on a wide ranging of different topics related to this work. We would also like to thank Christoph Ortner for making the ACE package available and Aidan Thompson for discussions about the SNAP potential. We gratefully acknowledge the United States Department of Energy under contract DE-NA0003965 for supporting this work.

Appendix A. Karhunen-Loève Expansion

We describe the KL expansion to generate an orthogonal basis set \( \{ \phi_n(x) \}_{n=1}^{N} \) from a given family of snapshots \( \{ \zeta_k(x) \}_{k=1}^{K} \). First, a two-point spatial correlation function is defined as

\[
\mathcal{K}(x, x') = \frac{1}{K} \sum_{k=1}^{K} \zeta_k(x)\zeta_k(x'),
\]

(A.1)
which accepts the following spectral decomposition

\[ K(x, x') = \sum_{k=1}^{K} \lambda_k \phi_k(x) \phi_k(x') . \]  

(A.2)

Here the set of basis functions \( \phi_k(x) \), \( 1 \leq k \leq K \), are ordered such that the associated eigenvalues

\[ \lambda_k = \frac{1}{K} \sum_{l=1}^{K} (\phi_k(x), \zeta_l(x))^2 \]  

(A.3)

satisfy \( \lambda_k \geq \lambda_{k+1} \), where \((\cdot, \cdot)\) denotes an appropriate inner product of two functions.

Next, for a given \( N \leq K \), the KL procedure consists in finding \( \phi_n, 1 \leq n \leq N \), so as to maximize the captured energy

\[ \max e_N = \sum_{n=1}^{N} \left( \frac{1}{K} \sum_{k=1}^{K} (\phi_n(x), \zeta_k(x))^2 \right) = \sum_{n=1}^{N} \lambda_n , \]  

(A.4)

subject to the constraints \((\phi_n(x), \phi_n'(x)) = \delta_{nn'}, 1 \leq n, n' \leq N \). The first few basis functions thus represent the main energy-containing structures in the snapshots, with their relative importance quantified by \( \lambda_n \). It can be shown that the problem (A.4) amounts to solve the eigenfunction equation

\[ (K(x, x'), \phi(x')) = \lambda \phi(x) \]  

(A.5)

for the first \( N \) eigenfunctions.

The method of snapshots \cite{28} expresses a typical empirical eigenfunction \( \phi(x) \) as a linear combination of the \( \zeta_k \)

\[ \phi(x) = \sum_{k=1}^{K} a_k \zeta_k(x) . \]  

(A.6)

Inserting this representation and (A.1) into (A.5), we immediately obtain

\[ Ca = \lambda a , \]  

(A.7)

where \( C \in \mathbb{R}^{K \times K} \) is given by \( C_{ij} = \frac{1}{K} (\zeta_i(x), \zeta_j(x)) \), \( 1 \leq i, j \leq K \). The eigenproblem (A.7) can then be solved for the first \( N \) eigenvectors from which the KL basis functions \( \phi_n, 1 \leq n \leq N \), are constructed by (A.6).
The optimality of the KL basis can be shown by considering an approximate representation \( \hat{\zeta}_k(x) = \sum_{n=1}^{N} \gamma_{Nn}^k \phi_n(x) \) of \( \zeta_k(x) \) for an arbitrary set of orthonormal basis functions, \( \{\phi_n\}_{n=1}^{N} \), and demonstrating that the KL basis is a minimizer of the error minimization problem

\[
\min \sum_{k=1}^{K} \left\| \zeta_k(x) - \sum_{n=1}^{N} \gamma_{Nn}^k \phi_n(x) \right\|^2. \tag{A.8}
\]

Indeed, this minimization problem is equivalent to the maximization problem \([A.4]\), which in turn asserts the optimality of \( \{\phi_n(x)\}_{n=1}^{N} \). Furthermore, the average least-squares error can be calculated as

\[
\frac{1}{K} \sum_{k=1}^{K} \left\| \zeta_k(x) - \sum_{n=1}^{N} (\phi_n(x), \zeta_k(x)) \phi_n \right\|^2 = \sum_{j=N+1}^{K} \lambda_j. \tag{A.9}
\]

Expression \([A.9]\) gives us an idea for choosing \( N \) as the smallest integer such that \( \sum_{n=1}^{N} \lambda_n / \sum_{k=1}^{K} \lambda_k \geq 0.9999 \).

The KL expansion is closely related to the proper orthogonal decomposition (POD) \([30]\) and indeed equivalent to the POD when the snapshots \( \xi_k \) are finite-size vectors instead of functions. In this connection, the POD can be viewed as a discrete version of the KL expansion.

References

[1] M. S. Daw, M. I. Baskes, Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals, Physical Review B 29 (12) (1984) 6443–6453. doi:10.1103/PhysRevB.29.6443.

[2] F. H. Stillinger, T. A. Weber, Computer simulation of local order in condensed phases of silicon, Physical Review B 31 (8) (1985) 5262–5271. doi:10.1103/PhysRevB.31.5262.

[3] J. Tersoff, New empirical approach for the structure and energy of covalent systems, Physical Review B 37 (12) (1988) 6991–7000. doi:10.1103/PhysRevB.37.6991.

[4] M. Z. Bazant, E. Kaxiras, J. Justo, Environment-dependent interatomic potential for bulk silicon, Physical Review B - Condensed Matter and Materials Physics 56 (14) (1997) 8542–8552. arXiv:9704137, doi:10.1103/PhysRevB.56.8542.
[5] D. W. Brenner, O. A. Shenderova, J. A. Harrison, S. J. Stuart, B. Ni, S. B. Sinnott, A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons, Journal of Physics Condensed Matter 14 (4) (2002) 783–802. doi:10.1088/0953-8984/14/4/312.

[6] A. C. Van Duin, S. Dasgupta, F. Lorant, W. A. Goddard, ReaxFF: A reactive force field for hydrocarbons, Journal of Physical Chemistry A 105 (41) (2001) 9396–9409. doi:10.1021/jp004368u.

[7] J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Physical Review Letters 98 (14) (2007). doi:10.1103/PhysRevLett.98.146401.

[8] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, Journal of Chemical Physics 134 (7) (2011). doi:10.1063/1.3553717.

[9] J. Behler, Representing potential energy surfaces by high-dimensional neural network potentials (2014). doi:10.1088/0953-8984/26/18/183001.

[10] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, Physical Review Letters 104 (13) (2010). arXiv:0910.1019, doi:10.1103/PhysRevLett.104.136403.

[11] S. Fujikake, V. L. Deringer, T. H. Lee, M. Krynski, S. R. Elliott, G. Csányi, Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures, Journal of Chemical Physics 148 (24) (2018). arXiv:1712.04472, doi:10.1063/1.5016317.

[12] W. J. Szlachta, A. P. Bartók, G. Csányi, Accuracy and transferability of Gaussian approximation potential models for tungsten, Physical Review B - Condensed Matter and Materials Physics 90 (10) (2014). doi:10.1103/PhysRevB.90.104108.

[13] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, G. J. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, Journal of Computational Physics 285 (2015) 316–330. doi:10.1016/j.jcp.2014.12.018.

33
[14] M. A. Wood, A. P. Thompson, Extending the accuracy of the SNAP interatomic potential form, Journal of Chemical Physics 148 (24) (2018). arXiv:1711.11131, doi:10.1063/1.5017641

[15] I. I. Novoselov, A. V. Yanilkin, A. V. Shapeev, E. V. Podryabinkin, Moment tensor potentials as a promising tool to study diffusion processes, Computational Materials Science 164 (2019) 46–56. arXiv:1812.02946, doi:10.1016/j.commatsci.2019.03.049

[16] A. V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials, Multiscale Modeling and Simulation 14 (3) (2016) 1153–1173. arXiv:1512.06054, doi:10.1137/15M1054183

[17] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, Physical Review B 99 (1) (2019). doi:10.1103/PhysRevB.99.014104

[18] R. Drautz, Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer, Physical Review B 102 (2) (2020). arXiv:2003.00221, doi:10.1103/PhysRevB.102.024104

[19] G. Dusson, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord, C. Ortner, Atomic cluster expansion: Completeness, efficiency and stability, Journal of Computational Physics 454 (2022) 110946. arXiv:1911.03550, doi:10.1016/j.jcp.2022.110946

URL https://www.sciencedirect.com/science/article/pii/S0021999122000080

[20] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, S. P. Ong, Performance and Cost Assessment of Machine Learning Interatomic Potentials, Journal of Physical Chemistry A 124 (4) (2020) 731–745. arXiv:1906.08888, doi:10.1021/acs.jpca.9b08723

[21] A. P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Physical Review B - Condensed Matter and Materials Physics 87 (18) (2013). arXiv:1209.3140, doi:10.1103/PhysRevB.87.184115.
[22] J. Behler, Perspective: Machine learning potentials for atomistic simula-
tions, Journal of Chemical Physics 145 (17) (2016). doi:10.1063/1.
4966192.

[23] M. Barrault, N. C. Nguyen, Y. Maday, A. T. Patera, An “Empir-
ical Interpolation” Method: Application to Efficient Reduced-Basis
Discretization of Partial Differential Equations, Comptes Rendus Math-
ematique 339 (9) (2004) 667–672. doi:10.1016/j.crma.2004.08.006.
URL http://linkinghub.elsevier.com/retrieve/pii/
S1631073X04004248http://www.sciencedirect.com/science/
article/pii/S1631073X04004248

[24] S. Boyaval, C. L. Bris, Y. Maday, N. C. Nguyen, A. T. Patera, A
reduced basis approach for variational problems with stochastic param-
eters: Application to heat conduction with variable Robin coefficient,
Computer Methods in Applied Mechanics and Engineering 198 (41-44)
(2009) 3187–3206. doi:10.1016/j.cma.2009.05.019.
URL http://linkinghub.elsevier.com/retrieve/pii/
S0045782509002114

[25] M. a. Grepl, Y. Maday, N. C. Nguyen, A. T. Patera, Efficient reduced-
basis treatment of nonaffine and nonlinear partial differential equations,
ESAIM: Mathematical Modelling and Numerical Analysis 41 (3) (2007)
575–605. doi:10.1051/m2an:2007031.
URL http://www.esaim-m2an.org/10.1051/m2an:2007031

[26] G. Rozza, D. B. P. Huynh, A. T. Patera, Reduced basis approximation
and a posteriori error estimation for affinely parametrized elliptic coer-
cive partial differential equations: Application to transport and contin-
umum mechanics, Archives Computational Methods in Engineering 15 (4)
(2008) 229–275.

[27] N. C. Nguyen, A multiscale reduced-basis method for parametrized
elliptic partial differential equations with multiple scales, Journal of
Computational Physics 227 (23) (2008) 9807–9822.
URL http://www.scopus.com/inward/record.url?
ed=2-s2.0-53349103005{&}partnerID=40{&}md5=
64852a2fac4b34c956379f980e2cf264

35
[28] L. Sirovich, Turbulence and the Dynamics of Coherent Structures, Part 1: Coherent Structures, Quarterly of Applied Mathematics 45 (3) (1987) 561–571.

[29] R. Everson, L. Sirovich, Karhunen-Loeve procedure for gappy data, Opt. Soc. Am. A 12 (8) (1995) 1657–1664.

[30] K. Willcox, J. Peraire, Balanced Model Reduction via the Proper Orthogonal Decomposition, AIAA 40 (11) (2002) 2323.

[31] A. P. Bart´ok, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, M. Ceriotti, Machine learning unifies the modeling of materials and molecules, Science Advances 3 (12) (2017). arXiv:1706.00179, doi:10.1126/sciadv.1701816.

[32] A. P. Bartók, J. Kermode, N. Bernstein, G. Csányi, Machine Learning a General-Purpose Interatomic Potential for Silicon, Physical Review X 8 (4) (2018). arXiv:1805.01568 doi:10.1103/PhysRevX.8.041048.

[33] V. L. Deringer, G. Csányi, Machine learning based interatomic potential for amorphous carbon, Physical Review B 95 (9) (2017). arXiv:1611.03277 doi:10.1103/PhysRevB.95.094203.

[34] D. Dragoni, T. D. Daff, G. Csányi, N. Marzari, Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron, Physical Review Materials 2 (1) (2018). arXiv:1706.10229 doi:10.1103/PhysRevMaterials.2.013808.

[35] M. A. Cusentino, M. A. Wood, A. P. Thompson, Explicit Multielement Extension of the Spectral Neighbor Analysis Potential for Chemically Complex Systems, The Journal of Physical Chemistry A 124 (26) (2020) 5456–5464. doi:10.1021/acs.jpca.0c02450. URL https://doi.org/10.1021/acs.jpca.0c02450.

[36] D. P. Kovács, C. V. D. Oord, J. Kucera, A. E. Allen, D. J. Cole, C. Ortner, G. Csányi, Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE, Journal of Chemical Theory and Computation 17 (12) (2021) 7696–7711. doi:10.1021/acs.jctc.1c00647.
[37] Y. Lysogorskiy, C. van der Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, R. Drautz, Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon, npj Computational Materials 7 (1) (2021). arXiv:2103.00814 doi:10.1038/s41524-021-00559-9.

[38] A. Bochkarev, Y. Lysogorskiy, S. Menon, M. Qamar, M. Mrovec, R. Drautz, Efficient parametrization of the atomic cluster expansion Physical Review Materials 6 (1) (2022) 13804. doi:10.1103/PhysRevMaterials.6.013804. URL https://link.aps.org/doi/10.1103/PhysRevMaterials.6.013804

[39] N. C. Nguyen, A. T. Patera, J. Peraire, A ‘best points’ interpolation method for efficient approximation of parametrized functions. International Journal for Numerical Methods in Engineering 73 (4) (2008) 521–543. doi:10.1002/nme.2086. URL http://doi.wiley.com/10.1002/nme.2086