Towards better efficiency of interatomic linear Machine Learning potentials

Alexandra Goryaeva, Jean-Bernanrd Maillet, Mihai Cosmin Marinica

Version of Record: https://www.sciencedirect.com/science/article/pii/S0927025619302605
Manuscript_dd3013f15ce94eb532b7db8e303f0e68

Towards better efficiency of interatomic linear Machine Learning potentials

Alexandra M. Goryaeva[a], Jean-Bernard Maillet[b], Mihai-Cosmin Marinica[a]

[a] *DEN - Service de Recherches de Métallurgie Physique, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France*
[b] *CEA - DAM, DIF, F-91297 Arpajon Cedex, France*

Abstract

Interatomic machine learning potentials have achieved maturity and became worthwhile alternative to conventional interatomic potentials. In this work we profile some characteristics of linear machine learning methods. Being numerically fast and easy to implement, these methods offer many advantages and appear to be very attractive for large length and time scale calculations. However, we emphasize that in order to be accurate on some target properties these methods eventually yield overfitting. This feature is rather independent of training database and descriptor accuracy. At the same time, the major weakness of these potentials, i.e., lower accuracy with respect to the kernel potentials, proves to be their strength: within the confidence limits of the potential fitting, one can rely on less accurate but faster descriptors in order to boost the numerical efficiency. Here, we propose a hybrid type of atomic descriptor that combines the original forms of radial and spectral descriptors. Flexibility in choice of mixing proportions between the two descriptors ensures a user defined control over accuracy / numerical efficiency of the resulting hybrid descriptor form. The performance and features of the above linear machine learning potentials are investigated for the interatomic interactions in metals of primary importance for fusion and fission applications, Fe and W. The suggested hybrid approach opens many avenues in the field of linear machine learning potentials that up to now are preferentially coupled with more robust and computationally expensive spectral descriptors.

Keywords: Interatomic potentials, Machine learning, Descriptors, SNAP, Molecular dynamics

1. Introduction

An accurate description of condensed matter properties requires a precise knowledge of the material at the atomic scale. Generally, the properties of materials are well described by *ab initio* methods, however, their numerical cost increases drastically with the system size, length

Email addresses: alexandra.goryaeva@cea.fr (Alexandra M. Goryaeva), mihai-cosmin.marinica@cea.fr (Mihai-Cosmin Marinica)

and time scale. Thus, the study of complex atomic configurations and their physical properties far from the equilibrium state (e.g., of irradiated systems, which typically contain in the order of 10^5-10^6 atoms) is often pushed beyond the scope of the first principle methods.

For metallic materials, the search for alternative theoretical methods has led to the development of various many-body interatomic models that span from second moment of tight binding approximation (1; 2; 3) to embedded-atom model (EAM) (4; 5), modified embedded-atom model (MEAM) (6), and higher order tight binding approximations (7; 8; 9; 10; 11; 12). Based on well-established physical models, these potentials use rigid and non-linear functional forms with few adjustable parameters. The optimization of the function parameters is a keen process, which is typically done using relatively small databases that contain few tens of *ab initio* or experimental observables. This approach ensures high computational efficiency, however, it implies a lack of transferability.

In the last decade, a prompt progress in machine learning (ML) methods has fueled a development of "flexible" interatomic potentials that can learn on the fly to predict atomic energies and/or forces. To a certain degree, both approaches, conventional and machine learning, are similar: an output (observable) is extrapolated from a given input (database). Compared to the traditional approach, the process of ML potential design differs by few points. Firstly, for ML potentials, the database is on an equal footing with the fitting formalism. As such, the definition of the potential is framed by these two characteristics. Consequently, ML potentials require using an extensive training database because its content has a strong impact on accuracy and transferability of the potential. Secondly, before performing the fit, all the database configurations, with different number of atoms, are mapped into a unique space called descriptor space. This projection is ensured by a descriptor or feature function, which also influence the potential accuracy and, therefore, represents an important intrinsic characteristic of the potential. And lastly, the fitting algorithm is performed in the descriptor space and the (statistical) ML procedure of the fit defines the performance and limitations of the potential. Below we briefly describe each of the points that define the performance of ML potentials, namely the database, the type and dimensionality of descriptors and ML procedures.

As for any interatomic potential, the target observables to extrapolate using a ML potential are mainly defined by the interests of the materials science community. The objective characteristics commonly include formation and migration energies of defects, phonon frequencies, elastic constants, etc. Very often the observables can be described by smooth continuous functions, e.g., by the total energy $E(\mathbf{r})$ of the system, which is a function of $3N$ atomic coordinates ($\mathbf{r} \in \mathbb{R}^{3N}$), with N being the number of atoms. However, fitting the observables of the system

directly as a function of atomic coordinates $\mathbf{r}$ using statistical ML methods is impractical. Here, ML approaches propose a notion of atomic descriptors, also known as fingerprints. The descriptors enable a specific representation of the crystal structure and its invariant description with respect to physical principles and symmetry including permutation of atoms, transnational and rotational invariance. These functions allow to capture the essential features of the individual atomic environments in a space of lower dimensionality, referred hereafter as the descriptor space. Thus, instead of using $\mathbb{R}^{3N}$ dimensional description of the local atomic environments, one employs a lower space $\mathbb{R}^{K}$ with $K < N$.

The first attempt to couple artificial intelligence and atomic-scale materials science was proposed by Behler and Parrinello in 2007 (13). This seminal paper has introduced the notion of descriptors adapted to the complexity of atomistic materials science (e.g., the radial $\mathbf{G}_2$ and angular $\mathbf{G}_3$ functions). Afterwards, Bartók et al. (14; 15) developed atomic descriptors based on spectral analysis of local atomic environments. More recently, Shapeev et al. (16; 17) introduced a new class of descriptors that are well suited for the moment tensor potentials (MTP). These potentials are based on moment tensors similar to the inertia tensors of local atomic environments. Besides the functions listed above, there are also the so-called similarity distance descriptors, which describe the distances between the atomic environments: Smooth Overlap of Atomic Positions (SOAP) (15), Atomic Configuration Distance (ACD) (18), and Graph Approximated Energy (GRAPE) (19).

The vast majority of existing ML models for atomistic calculations are based on neural networks (NN) (20; 21) or kernel methods (22; 23; 24; 25; 26; 27) (with the Gaussian function being the most common kernel). Both approaches provide a good numerical accuracy, however, application of NN potentials for molecular dynamics (MD) simulations is often challenging because of numerical instabilities that occur during simultaneous fit of the function and its gradient (i.e., of energies and forces). In contrast to NN, kernel methods are widely used for MD simulations. Some of the kernel models are formalized in the ever-growing field of the statistical on-the-fly learning methods (25; 27), while the others are built in the form of potentials such as the Adaptative Generalizable Neighborhood Informed (AGNI) (24) and Gaussian Approximation Potentials (GAP) (14; 28). The latter is the most robust and commonly used version.

Besides the NN and kernel methods, there is a so-called spectral neighbour analysis potential (SNAP) (29), which assumes the linear relationship between atomic energies and components of the spectral descriptor, namely of the bispectrum coefficients described in refs. (14; 15). The coefficients of such a potential are determined using weighted linear regression against the training database in the descriptor space. This method is referred hereafter as linear machine

learning (LML).

Computational cost and accuracy of ML potentials depend on several factors that are contingent on the underlying ML model and type of atomic descriptor. Similarly to any conventional potential, the numerical efficiency of ML potentials scales with the number of atoms N. The numerical cost of kernel potentials additionally scales with the number of local atomic environments in the training database. Moreover, the accuracy of the prediction strongly depends on the size and content of the database. In contrast to the kernel methods, the size of the database does not impact the numerical cost of the LML and NN potentials. However, for NN methods, the CPU cost also depends on the architecture of the network (e.g., on the number of hidden layers and number of neurons). Beyond the factors mentioned above, the numerical efficiency of all ML potentials is impacted by the time needed to compute a descriptor. The majority of existing kernel and LML potentials for metals (e.g., refs. (14; 22; 28; 29; 30; 31; 32)) are designed based on the robust and time consuming descriptors, like bispectrum (14; 15) or SOAP (15). Being combined with these descriptors, kernel (GAP) and LML (SNAP) potentials are, respectively, four and two orders of magnitude slower than EAM potentials (33). With such a numerical cost of kernel ML potentials, it is computationally expensive to have an access to free energy methods where the number of force evaluations exceeds 10^5 per atom. Thus, taking into account high computational cost of the kernel methods and limited application of NN for MD simulations, the LML potentials with a good tradeoff between accuracy and computational efficiency are the the good candidates for the free energy calculations. The main advantages of this method (compared, for instance, to kernel methods) is that the size of the training database does not affect the numerical cost of the potential. The LML potential is only defined by the set of $K + 1$ parameters, with K being the dimension of the descriptor function. Moreover, the LML method is relatively simple to implement and it allows for simultaneous fit of energies, forces and stress tensor, crucial for modeling large defects and irradiated systems.

In this work we aim to investigate the performance and limits of the LML formalism combined with different descriptors. The utility of accurate and slow descriptors, such as bispectrum or SOAP, are justified for the kernel based ML algorithms and NN. However, for the relatively simple LML algorithm, the inherent training / test error is inevitably higher than that in more complex and computationally expensive methods. Here, we examine the efficiency of different atomic descriptors coupled with LML regression in order to define the confidence region and limits of the potential (while targeting to reproduce lattice parameters and elastic constants of bcc metals). We suggest using less accurate descriptors that have the advantage to boost the numerical evaluation while preserving the training / test error at the level of complex descriptors

coupled with LML. We propose a method to build the so-called hybrid descriptors by mixing precise and computationally expensive descriptors, like bispectrum SO(4), of low dimensionality with numerically fast and less accurate ones, like Behler-Parinello radial function $\mathbf{G}_2$. Further, we demonstrate how the balance between the numerical cost and accuracy of such descriptors can be controlled while varying the proportions between heavy and fast descriptors.

2. Methods

2.1. Databases

In this work we fit the LML potentials using DFT databases for Fe and W (Table 1). The datasets include atomic structures of bcc bulk from molecular statics and molecular dynamics calculations, liquid configurations and the structures of point, planar and linear defects. For one configuration (simulation cell) with N atoms, the maximum number of the data points to fit is $(n_E + n_F + n_S) = 3N + 7$, which results from the simultaneous fit of energy ($n_E = 1$), forces ($n_F = 3N$) and six independent components of virial stress ($n_S = 6$). The W database is significantly bigger than that of Fe. Overall, the W database contains 262203 $(n_E + n_F + n_S)$ datapoints, whilst the Fe dataset contains 20577 points available for the test and train. The detailed content of the databases and the number of data points and configurations used for the train and test are reported in Table 1.

Table 1: Composition of Fe and W databases employed for fitting LML potentials. In the column with properties to fit, E corresponds to energies, F stands for forces and S for stress. The number of data points to fit is defined by $(n_E + n_F + n_S)$ with n_E, n_F and n_S being the number of points with energies, forces and stress, respectively. The train / test describes the number of training and testing configurations, respectively.

DB class	Fe				W (30)			
	Atoms per cell	Properties to fit	$n_E + n_F + n_S$ train / test	Configurations train / test	Atoms per cell	Properties to fit	$n_E + n_F + n_S$ train / test	Configurations train / test
Bcc unit cell	2	ES	840/483	120/69	1	ES	7336/6664	1048/952
Elasticity	2	ES	231/462	33/66	–	–	–	–
Bcc - MD	128	F	11520/3840	30/10	128	EF	11550/11550	30/30
Liquid	100	F	2100/900	7/3	–	–	–	–
Free surfaces	14-48	E	4/4	4/4	12	E	90/90	90/90
γ-surfaces	12	E	120/49	120/49	12	E	3000/3183	3000/3183
Vacancies	126-127	E	3/3	3/3	47-127	EF	89070/89070	585/585
Self-interstitials	129-132	E	9/9	9/9	–	–	–	–
Dislocations	–	–	–	–	135	EF	4060/36540	10/90
Total			14827/5750	326/213			115106/147097	4763/4930

The Fe database is calculated with VASP (34) using PAW pseudopotential with GGA-PBE functional that accounts for 14 valence electrons $3p^6 3d^7 4s^1$. The plane-wave cutoff energy is set to 400 eV. The Brillouin zone is sampled with the Monkhorst-Pack scheme such as the smallets spacing between the k-points is 0.0175 Å^{-1}.

For W, we employ the open access database from libAtoms.org (Cambridge University), which was previously used for fitting a GAP potential (30). This database was calculated with CASTEP (35) using ultrasoft pseudopotential and GGA-PBE functional with $5s^2 5p^6 5d^4 6s^2$ electrons. More details about the W database calculations can be found in (30).

2.2. Descriptor functions

2.2.1. Behler-Parinello descriptor functions

The Behler-Parinello (BP) symmetry functions were originally introduced in 2007 (13) and since then they are widely used as descriptors coupled with NN potentials (36) for various materials, such as aluminum (24; 37) and its alloys (38), carbon allotropes (graphite-diamond) (39), sodium (40; 41), germanium telluride (42), water (43; 44), etc.

The most simple and fast BP descriptor is the two-body symmetry function $\mathbf{G}_2^a$ that takes into account only radial atomic environment. Its form is, in fact, somewhat similar to the radial distribution function. Using the $\mathbf{G}_2^a$ function, the environment of the a^{th} atom can be written as:

$$G_2^a(\eta, R_s) = \sum_{b \in v(a)} e^{-\eta(r_{ab} - R_s)^2} w_b f_c\left(r_{ab}\right) , \tag{1}$$

where $v(a)$ denotes the set of all the neighbors of the a^{th} atom within a cutoff function $f_c\left(r_{ab}\right)$ (detailed below); parameters R_s and η control the position and width of the Gaussian with respect to the central atom; w_b is a weight factor, commonly taken as atomic mass, which allows to discriminate between various types of atoms. Using the grids of $\eta = \eta_1, \ldots, \eta_{g_1}$ and $R_s = R_{s,1}, \ldots, R_{s,g_2}$, the local environment of the a^{th} atom is projected into descriptor function space of dimension $j = dim(\mathbf{G}_2^a) = g_1 \times g_2$. Most commonly, these values are taken as $g_2 = 1$ and $R_s = 0$ whilst the grid η has $5 - 200$ values between 10^{-2} and 1.0.

In order to go beyond the radial function $\mathbf{G}_2^a$ (Eq. 1), it is necessary to consider angular many-body interactions. It implies considering triplets of atoms, which allows to take into account the angular correlations between the atoms. Using the angle $\theta_{abc} = \mathbf{r}_{ab}\mathbf{r}_{ac}/r_{ab}r_{ac}$ between the three atoms a, b and c (centered on the atom a), one can define the $\mathbf{G}_3^a$ descriptor as follows (13):

$$G_3^a(\eta, \lambda, \zeta) = 2^{1-\zeta} \sum_{b,c \neq a} (1 + \lambda \cos\theta_{abc})^{\zeta} \cdot e^{-\eta\left(r_{ab}^2 + r_{ac}^2 + r_{bc}^2\right)} \cdot f_c'\left(r_{ab}, r_{ac}, r_{bc}\right) , \tag{2}$$

where the parameter ζ controls the angular components of atomic environments; $\lambda = \pm 1$ defines the extreme positions of cosine function; and the function $f_c'(r_{ab}, r_{ac}, r_{bc}) = f_c(r_{ab}) \cdot f_c(r_{ac}) \cdot f_c(r_{bc})$. The η and ζ grids together with λ give the dimension of the descriptor function space $j = dim(\mathbf{G}_3^a)$.

The cutoff function on interatomic distances $f_c(r_{ab})$ describes the relevant local atomic environment within the distance $r \leq R_{cut}$ according to the following expression:

$$f_c(r) = \frac{1}{2}\left[\cos\left(\pi \frac{r}{R_{cut}}\right) + 1\right]. \tag{3}$$

The cutoff distance R_{cut} should be sufficiently large to include several nearest neighbors. In the literature, the R_{cut} values are commonly taken between 4 and 12 Å, dependently on investigated material and desirable convergence criterion (e.g., refs. (13; 39; 42; 45)). In this work, we employ the threshold values of 5 and 8 Å both for Fe and W, which is large enough for these materials.

2.2.2. Spectral descriptor functions

An alternative approach that accounts both for the radial and angular information information was proposed by Bartók et al. (14; 15). The spectral approach implies a quantitative evaluation of the density of neighbors centered on one atom. The neighbor density $\rho_a(\mathbf{r})$ for each atom a can be considered as a sum of δ functions at all positions of neighboring atoms up to a cutoff distance, weighted depending on the species involved:

$$\rho_a(\mathbf{r}) = w_a \delta(\mathbf{r}) + \sum_{b \in v(a)} f_c(r_{ab}) w_b \delta(\mathbf{r} - \mathbf{r}_{ab}). \tag{4}$$

The power spectrum of the local atomic density is computed by projection on some basis functions, such as, angular spherical or hyperspherical functions. The concept of power spectrum can be extended to the bispectrum (15; 46; 47). The latter is the direct product of the power spectrum and can be computed by coupling two different angular channels. The descriptors that are based exclusively on angular spherical functions Y_{lm} contain only radial information and, consequently, they are incomplete. However, the descriptor based on hyspherical functions, like bispectrum SO(4) (14; 15), was shown to be complete (14; 15; 46; 47), i.e., able to determine the local atomic environments uniquely, including the local symmetry operations like translation, rotation, inversion and permutation of particles. Instead of Cartesian coordinates $\mathbf{r}$, this approach relies on the bispectrum components of the 4D hyperspherical harmonics projected onto the R^4-sphere (θ, ϕ, θ_0). The relation between polar and Cartesian coordinates is bijective (48).

It is interesting to note that these functions are the transformations matrices for standard spherical harmonics under rotations by angle $\varphi = 2\theta_0$ around the axis defined by angles θ and

ϕ. Consequently, the functions can be easily related to the Wigner D-matrix, which gives the irreducible representations of the SO(3) group (48). The neighbor density function $\rho_a(\mathbf{r})$ of the a^{th} atom can be described in 4D hyperspherical harmonics as follows:

$$\rho_a(\mathbf{r}) = \sum_{j=0,1/2,\ldots}^{\infty} \sum_{m=-j}^{j} \sum_{m'=-j}^{j} c^a_{j;mm'} U_{j;mm'}(\theta_0, \theta, \phi \equiv \mathbf{r}) \,. \tag{5}$$

The expansion coefficients $c^a_{j;mm'}$ (also called power spectrum coefficients) for the a^{th} atom are given by the inner product between the density and hyperspherical functions:

$$c^a_{j,mm'} = w_a U_{j,mm'}(0,0,0) + \sum_{b \in v(a)} f_c(r_{ab}) w_b U_{j,mm'}(\theta^{ab}_0, \theta^{ab}, \phi^{ab} \equiv \mathbf{r}_{ab}) \,, \tag{6}$$

where, the j values can only be positive integer or half-integer, i.e., $j = 0, \frac{1}{2}, 1, \frac{3}{2}$, etc. For the maximal value of the angular moment j_{max}, the total number of components $U_{j;mm'}$ is $\sum_{j=0}^{j_{max}} (2j+1)^2$. The polar angles $\theta^{ab}_0, \theta^{ab}, \phi^{ab}$ are the representation on the $\mathbf{r}_{ab}$ vector on the $\mathbb{R}^4$-sphere. The direct evaluation of these functions is commonly performed using recurrence relations (see Appendix A for details).

The bispectrum SO(4) coefficients $B^a_{ll_1l_2}$ (15) of each atom a can be then written as:

$$B^a_{ll_1l_2} = \sum_{m',m=-l}^{l} \sum_{m'_1,m_1=-l_1}^{l_1} \sum_{m'_2,m_2=-l_2}^{l_2} c^{a*}_{j;m'm} C^{ll_1l_2}_{m'm'_1m'_2} C^{ll_1l_2}_{mm_1m_2} c^a_{l_1;m'_1m_1} c^a_{l_2;m'_2m_2} \,, \tag{7}$$

where $C^{ll_1l_2}_{mm_1m_2}$ are the Clebsch - Gordan coefficients. The number of these coefficients can be reduced while applying the selection rules (14; 29; 48). In this work we rely on the selection rules suggested in (48), as implemented in the MiLaDy package (49).

Brute application of the Clebsch-Gordan selection rules yields the bispectrum components with $l_1 \neq l_2$. In the frame of low-dimensional bispectrum analysis of signal problems, this choice was shown to be overcomplete (46). Therefore, some research papers dealing with the bispectrum descriptor for atomic systems eventually rely on the diagonal components only, for which $l_1 = l_2$ (15; 28). Here, we use the same selection approach that accounts for diagonal components and provides, for instance, 26 independent components for $j_{max} = 7/2$ (vs 40 components without diagonal constraint, as was used in refs. (29; 32; 50)), instead of the full set of $(2\times7/2+1)^3 = 512$ bispectrum coefficients.

2.2.3. Hybrid descriptor functions

The choice of atomic descriptor functions has an impact on the accuracy of the potential as well as on its numerical cost. Here, we propose a hybrid form of atomic descriptors that combines slow and numerically accurate descriptors with fast and less accurate ones. In this

case, the choice of mixing proportions between the the two descriptors is defined by user and should be driven by the the desirable precision of the fit as well as by the target physical problem to address.

In many total energy models, including few angular components is sufficient to ensure the accuracy of the underlying atomic interactions. As such, low values of the angular monent j_{max} of bispectrum b-SO(4) are sufficient for the angular description. However, this low values of j_{max} will degrade the radial description provided by the spectral descriptor. In order to ensure a robust radial description of the structure, one can additionally introduce a radial descriptor (e.g., $\mathbf{G}_2$) contributing to the hybrid function.

In this work we design a hybrid descriptor $\mathbf{D}$, called hereafter G2B4, which is built on the radial descriptor $\mathbf{G}_2$ (Eq. 1) and complemented with the bispectrum SO(4) coefficients $\mathbf{B}$ (Eq. 7) of relatively low dimensionality ($j_{max} = 1.5-3.5$). Adding low dimensionality bispectrum allows enriching the radial descriptor with some angular information, and, consequently, increasing the descriptor accuracy, without drastically rising the associated computational cost (for more details, see the section 3.3). The new descriptor G2B4 is built as a direct sum $\mathbf{D} = \mathbf{G}_2 \oplus \mathbf{B}$ and its dimension $K = dim(\mathbf{D})$ is equal to the sum $K = dim(\mathbf{G}_2) + dim(\mathbf{B})$. This mixing option is not unique and other possibilities can be explored. Our choice is mainly driven by the aim to examine the performance of hybrid descriptors that are built by the functions with the extreme limits of low / high accuracy and low / high computational cost.

2.3. Linear regression in the descriptors space

Designing a ML potential via linear fit in the descriptor space, i.e., LML, was originally introduced by Thompson and co-workers (29) in the framework of SNAP method, which links the bispectrum SO(4) coefficients $B^a_{ll_1l_2}$ (Eq. 7) to the total energy of the system. A similar regression algorithm was also employed in refs. (16; 17) where MTP potential uses invariant polynomials as descriptors. In this study we test how the linear formalism combines with other descriptors, e.g., with the BP functions (Eqs.1, 2) and the hybrid descriptor G2B4 (section 2.2.3).

Assuming a linear relation between the atomic energy and descriptor coefficients (29), the LML potential energy E^a_{LML} of each atom a can be computed as a sum over the K descriptor components $\mathbf{D}^a = (D^a_1, \ldots, D^a_K)$ multiplied by the corresponding parameters $(\beta_1, \ldots, \beta_K)^T$ along

each direction:

$$E^a_{LML}(\mathbf{D}^a) = \beta_0 + \sum_{k=1}^{K} \beta_k D^a_k = (1, \mathbf{D}^a) \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}, \tag{8}$$

where β_0 is the constant energy contribution. The energy descriptor of the local environment of the atom a is the $\mathbb{R}^{1\times(K+1)}$ vector $(1, \mathbf{D}^a)$. Then, assuming that the energy of the atomic system with N atoms can be represented as a sum of the local contributions from each atom (29), the total energy E_{LML} can be written as:

$$E_{LML} = N\beta_0 + \sum_{k=1}^{K} \beta_k \sum_{a=1}^{N} D^a_k = \left(N, \sum_{a=1}^{N} \mathbf{D}^a \right) \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}. \tag{9}$$

The total energy descriptor becomes $\mathbf{D}_E = (N, \sum_a \mathbf{D}^a)$ vector in the $\mathbb{R}^{1\times(K+1)}$ space. The energy of the corresponding system is obtained by a linear regression using the $K+1$ parameters mentioned above. The forces acting on each atom and the stress tensor of the system can be obtained from the derivatives of the energy (Eq. 9) with respect to the atomic positions $\mathbf{r}$. Thus, the force $\mathbf{F}^a_{LML}$ acting on the atom b in the direction α is:

$$F^{b\alpha}_{LML} = -\nabla_{b\alpha}\mathbf{E}_{LML} = \sum_{k=1}^{K} \beta_k \left(-\sum_{a=1}^{N} \frac{\partial D^a_k}{\partial x_{b\alpha}} \right) = \left(0, -\sum_{a=1}^{N} \frac{\partial \mathbf{D}^a}{\partial x_{b\alpha}} \right) \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}. \tag{10}$$

The descriptor for the force on the b^{th} atom in direction α becomes $\mathbf{D}^{b\alpha}_F = (0, -\sum_{a=1}^{N} \partial \mathbf{D}^a / \partial x_{b\alpha}) \in \mathbb{R}^{1\times(K+1)}$. As in the case of energy (Eq. 9), the atomic forces are in linear relation with the force descriptors and they are characterized by the same $K+1$ regression parameters. The force descriptors account for the neighbours of the b^{th} atom. In practical numerical implementation, the descriptor functions have a finite cutoff radius and, due to the periodic boundary conditions, several replicas of the system can be included into the cutoff region. Thus, the sum over N atoms should be replaced by the sum of the neighbours of the b^{th} atom:

$$\sum_{a=1}^{N} \rightarrow \sum_{a\in v(b)}. \tag{11}$$

Based on the similar protocols for the definition of virial stress, the stress descriptors $\mathbf{D}_S$ can be further deduced.

Thus, the evaluations of the descriptors and their derivatives fully determine the energy, force and stress descriptors. The derivatives of the BP descriptors $\mathbf{G}_2$ and $\mathbf{G}_3$ (Eqs. 1, 2) are extensively discussed in the literature (20; 36; 39; 51). The derivatives of the bispectrum coefficients $B^a_{ll_1l_2}$ (Eq. 7) are provided in the Appendix A.

2.4. LML potential fitting procedure

In practice, fitting a LML potential for energy, forces and stress, implies finding an optimal set of $K+1$ linear coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)^T$, where $K = dim(\mathbf{D})$ is the descriptor dimension. While dealing with a linear formalism, one can formulate this task as solving a system of linear equations $\mathbf{A}\cdot\boldsymbol{\beta} = \mathbf{y}$ for the coefficients $\boldsymbol{\beta}$ (29). In this system, the vector $\mathbf{y}$ contains the information on the target energy, forces and stress tensor from the DFT database (Table 1). Its dimensions are $(n_E + n_F + n_S) \times 1$, where n_E, n_F and n_S are the number of energies, forces and stress components of all the atomic configurations in the training dataset, respectively. The matrix $\mathbf{A}$ holds the corresponding descriptors $\mathbf{D}_{E,F,S}$ for energy, force and stress and adopts the dimensions $((n_E + n_F + n_S) \times (K+1))$. The explicit structure of this set of linear equations can be found elsewhere (29; 32).

Such a system of equations is overdetermined and, therefore, has no unique solution. The optimal solution $\boldsymbol{\beta}$ is imposed by minimizing an objective function $\mathbb{L}^2$, e.g., $J(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|^2$. The objective function is differentiable and the solution satisfies the equation $\nabla\ J(\boldsymbol{\beta}) = 0$, giving (52):

$$\boldsymbol{\beta} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}\,. \tag{12}$$

In order to fit some observables with higher accuracy compared to the other properties, a weights diagonal matrix $\mathbf{W}$ with dimensions $(n_E+n_F+n_S)\times(n_E+n_F+n_S)$ should be introduced. Each element $W_{i,i}$ in this matrix contains the weight that corresponds each individual observable to fit. Its value controls the accuracy of the fit for that particular property. Accounting for the weights, the objective function can be written as $J(\boldsymbol{\beta}) = \left\|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{A}\boldsymbol{\beta})\right\|^2$ and the solution for the potential parameters becomes:

$$\boldsymbol{\beta} = \left(\mathbf{A}^T\mathbf{W}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{W}\mathbf{y}\,. \tag{13}$$

Overall, previous LML studies (29; 32; 50; 53) favor the weighted solutions (Eq. 13) to the non-weigthed fit (Eq. 12) . In the section 3.3 we will discuss the effect of weighted fit on accuracy of target observables and on the quality of the fit for other observables.

3. Results and discussion

3.1. Computational cost of the LML method combined with different descriptors

In this section we estimate the CPU cost of the LML method coupled with different atomic descriptors. Once a descriptor and its derivatives are computed, numerical evaluation of observables, such as energies (Eqs. 8-9), forces (Eq. 10) and stress is fast (comparable with that of a standard empirical potential). Being independent of the database size, numerical cost of the LML potentials is controlled by the expense of the descriptor functions, which is strongly dependent on the descriptor type and its dimensionality. Here, we compare the cost of some widely used atomic descriptors with that of our hybrid descriptor G2B4: $\mathbf{D} = \mathbf{G}_2 \oplus \mathbf{B}$ with different proportions between the radial and bispectrum components. The calculations are performed using a single Intel Broadwell core and the recorded computational time is normalized by that of bispectrum b-SO(4) with j_{max}=3.5 and R_{cut}=5 Å, commonly used for the SNAP potentials of bcc metals (29; 32; 50; 53). From Figure 1 one can see that the hybrid descriptor can be 8-10 times faster than the reference bispectrum b-SO(4) (see, e.g., the difference between G2B4 curves with $j_{max} = 1.5$ and b-SO(4) curves with $j_{max} = 1.5$; or G2B4 curves with $j_{max} = 2.5$ and b-SO(4) curves with $j_{max} = 4.0$). The computational cost of hybrid descriptors is rather controlled by the expense of bispectrum. This feature can be seen from the two gray G2B4 curves with the bispectrum angular moment j_{max}=1.5 (Fig. 1). Although, the dimensionality of the radial functions that contribute to the G2B4 descriptors differs almost by a factor of three, the numerical cost the two hybrid descriptors is very similar (10-15 % difference).

The ability these descriptors in combination with the LML method to compute the ground state properties (e.g., unit cell parameters and elastic properties) of Fe and W are further examined in the section 3.3.

3.2. Simple LML regression without weights

In the literature one consistently finds the weighted solutions for the linear SNAP potentials (29; 32; 50; 53), while the efficiency and limitations of simple LML fit are never discussed. Here we test the performance of the SNAP potential (LML + bispectrum b-SO(4) with $j_{max} = 3.5$ and R_{cut}=5 Å) without applying any regression weights. The exact ratios between the train/test configurations for different database classes (e.g., primitive cell, free surfaces, point defects, etc.) are provided in Table 1. Atomic configurations from all classes are present in the train / test data sets. The results of the fit are reported in Figure 2. For the performed fit, we consistently find the test errors similar to those from training (Fig. 2). Regardless of the database size (small Fe or big W), the mean square errors (MSE) are always within 20 meV/atom, 100 meV/Å and

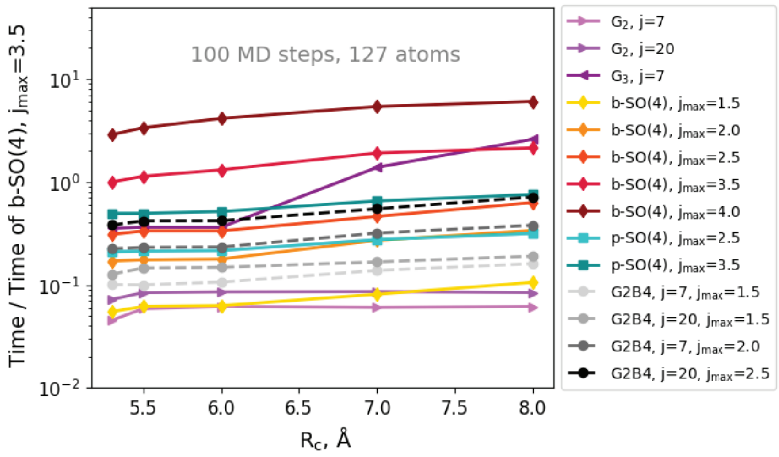


Figure 1: Computational cost of the atomic descriptors (on a single Intel Broadwell core) normalized by that of bispectrum b-SO(4) with j_{max}=3.5 and R_{cut}=5. The timings do not comprise evaluation of the nearest neighbor environment, i.e., they are independent of the neighbor list algorithm. The $G_{2,3}$ curves are the BP descriptors; b-SO(4) and p-SO(4) curves correspond, respectively, to the bispectrum and powerspectrum functions; G2B4 is the hybrid descriptor. The indexes j refer to dimensionality of the BP descriptors, j_{max} is the angular moment of the spectral descriptors.

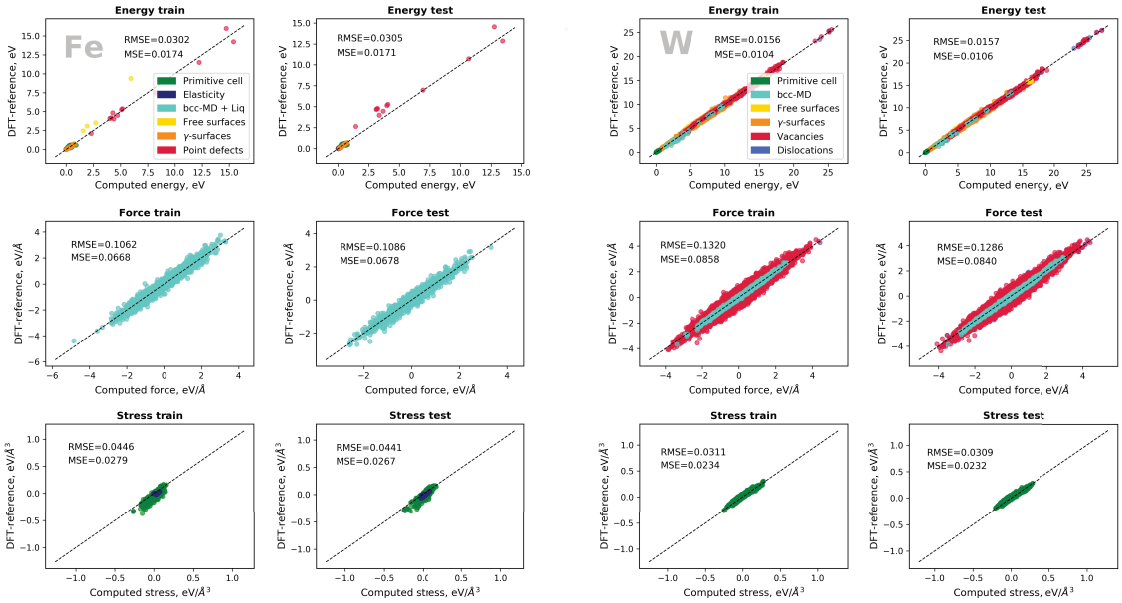


Figure 2: Linear SNAP potential for Fe and W fitted without weights: comparison of the energies, forces and stress with the corresponding DFT values from the database. Atomic systems that carry information on different materials properties (e.g., containing different structural defects) are depicted with different colors. The RMSE and MSE errors of fit are provided on each subplot.

40 meV/Å^3 for energies, forces and stress, respectively. The present linear fit with the train/test strategy yields the same results as the K-cross validation procedure.

The ground state properties of Fe and W from these potentials are illustrated in Figure 3

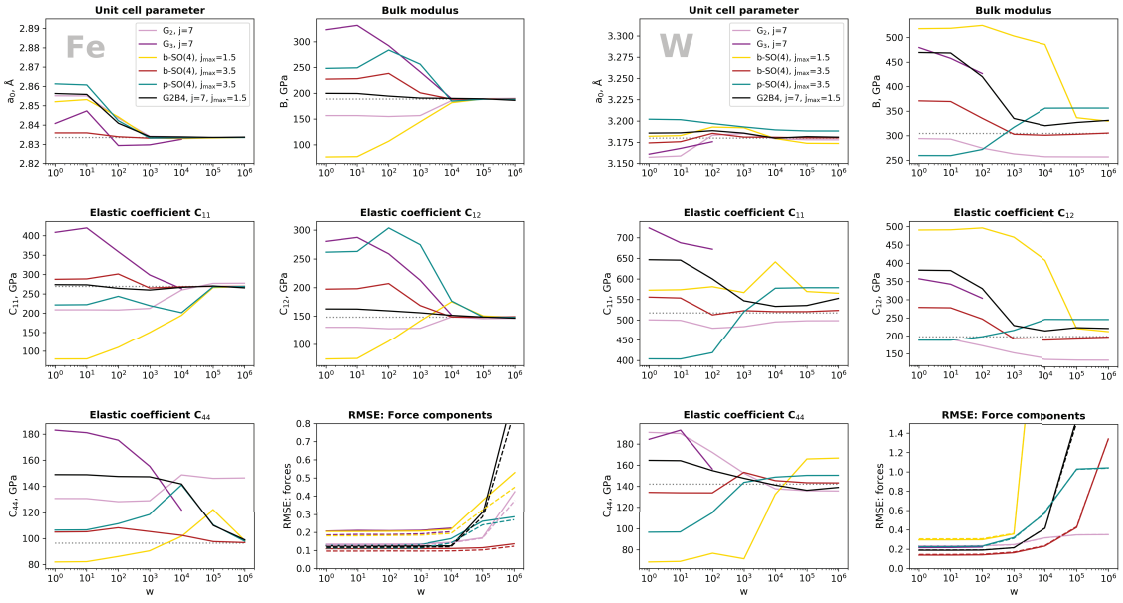


Figure 3: Evolution of the lattice constant a_0, bulk modulus B, elastic constants C_{ij} and of the RMSE error fit for the force components (solid lines- training configurations, dashed lines - test configurations) in bcc-Fe and bcc-W as a function of weight $\mathbf{W}$ applied for fitting elastic properties. The target values of the computed properties (from the DFT database) are indicated with the dashed gray lines. In the legend, the indexes j refer to dimensionality of the BP $G_{2,3}$ descriptors, j_{max} is the angular moment of the spectral descriptors.

(see the red b-SO(4) curves with j_{max}=3.5 at weights $\mathbf{W} = \mathbf{1}$). Without applying regression weights, both potentials provide the unit cell parameter and elastic constants with the 5% −20% difference with respect to the target DFT values from the database (indicated with gray dashed line in Fig. 3). Thus, even combined with a robust descriptor and extensive database, LML method alone is not sufficient to provide a good quality fit for ground state properties. Further, we examine how the quality of fit and the ML potential performance (including computational cost) can be improved.

3.3. Utility of different descriptors and effect of weights

In this section we test the utility of various atomic descriptor functions combined with the LML method. Relying on the chosen databases, we aim to reproduce the essential ground state properties of Fe and W, such as bcc unit cell parameter a_0 (first order derivative of the energy) and elastic properties (second order derivative of the energy) while using different atomic descriptors. In order to reach this objective, we gradually increase the weights on the elastically deformed bcc configurations exclusively. These configurations represent *ca.* 10% and 15% of the training bcc dataset for W and Fe, respectively. From Figures 3 one can see that in order to reach a reasonable convergence of the objective properties, the weights should be increased at least up to 10^3. For the bigger W database, the convergence of elastic constants can be

reached with the weights in the order of 10^4, while for the small Fe dataset, the weights in range of $10^4 - 10^5$ are rather required. As any radial potential (11; 12), LML coupled with the $\mathbf{G}_2$ always provides $C_{12} = C_{44}$. Among the tested descriptors, only bispectrum b-SO(4) with j_{max}=3.5 (this descriptor is conventionally coupled with the SNAP formalism) and the hybrid descriptor G2B4 maintain a relatively low fit error with the weight of 10^4 and 10^5 that ensure the convergence of elastic constants for Fe and W, respectively.

However, when the weights reach the values of $10^4 - 10^5$, transferability of the potentials wanes. Increasing weights on the selected properties results in increase of the train / test fit error for the other configurations where no large weights were set. Figure 3 depicts the evolution of RMSE errors of the force components for training (solid lines) and testing (dashed lines) configurations. While reaching the convergence of elastic properties with $\mathbf{W} = 10^4\mathbf{1}$, we decrease the quality of fit for the other properties and imply an overfit.

For the present LML regression, the values of weights define the confidence limit of the ML potential. This confidence limit is rather similar for the majority of the tested descriptors (except for the powerspectrum p-SO(4) where the confidence zone declines faster). The limit of the confidence zone directly impacts transferability of LML potentials, therefore, it represents one intrinsic potential characteristics.

Within the confidence zone, the convergence of observables can be used as a criterion for the optimum descriptor choice. Considering two atomic descriptors that yield comparable confidence limit when coupled with the LML method, one can choose the one, which is numerically faster. Being much faster (Fig. 1), the hybrid G2B4 can be a promising candidate to substitute of the bispectrum in LML potentials. Moreover, the accuracy of this descriptor can be gradually increased by incorporating more $\mathbf{G}_2$ or $\mathbf{B}$ components. From the trial calculations with a hybrid descriptor that has j_{max}=1.5 of bispectrum $\mathbf{B}$ and $j = 30$ of radial $\mathbf{G}_2$ function, we find the LML fit accuracy similar to that of the regular bispectrum SO(4) with j_{max}=3.5. While increasing the number of components in $\mathbf{G}_2$ from 7 to 30, the computational cost of G2B4 increases only by 20%; whilst the calculations with pure b-SO(4) with j_{max}=3.5 are amost 10 times slower. The flexible character of hybrid descriptors allows for a control over numerical cost and accuracy of the potential.

The user-defined control over the cost/accuracy balance in LML potentials combined with hybrid descriptors enables a hierarchical approach for the materials properties calculations. Here, we provide an example of such hierarchcal calculations for the phonons spectrum of bcc W (Fig. 4). The accuracy of the phonons gradually improves with increasing the fidelity of LML potentials, i.e, with increasing accuracy of the associated descriptor. With this approach, it is

possible to obtain a set of LML potentials that provide a similar level of accuracy for certain materials properties (e.g., for a_0 and elastic constants), whilst the phonons, and, consequently, some finite temperature properties will continuously improve with increasing numerical cost of the potential. Figure 4 illustrates how the increasing the completeness of the descriptor gradually increases the ability of LML potentials to predict the phonons dispersion along the high symmetry direction in first Brillouin zone. The pure b-SO(4) descriptor with j_{max}=4.0 (Fig. 4d) gives the accuracy limit for the calculations (the small differences between the LML curves and experimental points are mainly due to the systematic error of the DFT calculations that contribute to the W database). From the tested set of LML potentails (Fig. 4), the reasonably accurate phonons can be obtained with the hybrid descripors with the bispectrum angular moment $j_{max} \geqslant 2.5$. The hybrid descriptors G2B4 from Fig. 4b and Fig. 4c are, respectively, 8 and 3 times faster than the pure bispectrum from Fig. 4d.

The main difference of the proposed approach from the conventional potentials is in unambiguous control over the accuracy of target properties, which does not trigger a degradation of other already adjusted properties. The proposed hierarchical approach via hybrid descriptors opens many avenues in the field of molecular dynamics calculations that, for instance, use thermodynamical adaptive sampling methods, such as adapting biasing force (54; 55; 56; 57) or potential dynamics (58; 59; 60). The less accurate but fast hybrid descriptors can be used for an exhaustive exploration of the region of interest over the phase space. With a focus on a particular region, one can gradually increase the accuracy of sampling using more accurate but slower ML potentials. As such, the quality of thermodynamical sampling can drastically increase while keeping the numerical effort within reasonable limits.

4. Conclusions and perspectives

In this work we have examined the performance of the LML method combined with different atomic descriptors. In the framework of LML, we have proposed a new type of descriptors, the hybrid descriptors G2B4, built as a direct sum $\mathbf{D} = \mathbf{G}_2 \oplus \mathbf{B}$. The efficiency of the LML method and the newly proposed hybrid descriptors was tested for the bcc metals of primary importance for fusion and fission applications, Fe and W. Below we summarize the main conclusions and perspectives.

- Linear ML methods have a general tendency to overfitting. This generic characteristic is rather independent of descriptors or database. We have shown that even simple weighted fit of elastic constants and lattice parameters can eventually bring the solutions in the overfitting zone. This behavior imposes a limit between the optimal and overfitted set of

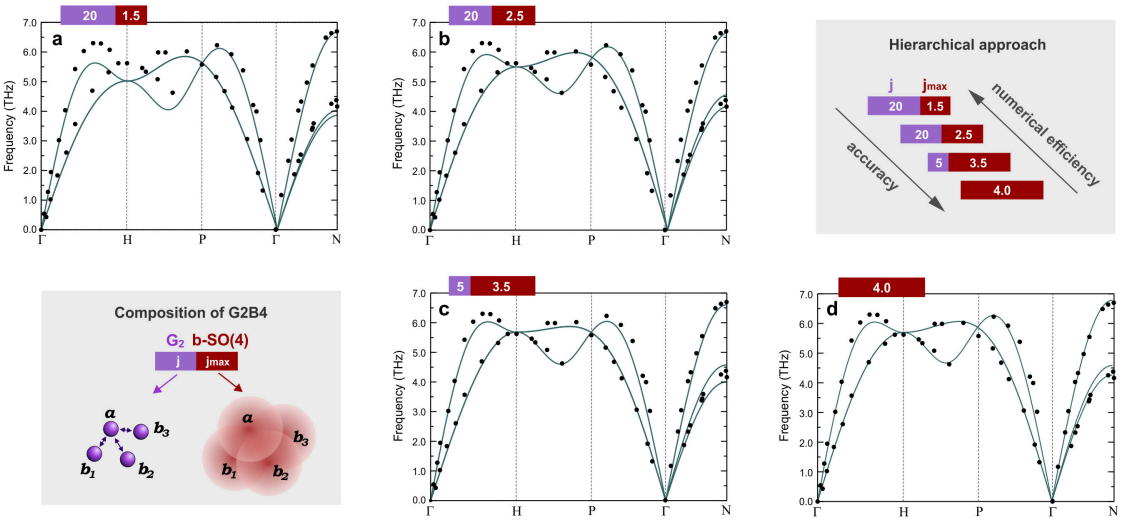


Figure 4: Hierarchical approach for the phonons calculations enabled by the hybrid descriptors G2B4. (a-d) Phonon dispersion curves in W, plotted along high symmetry directions of the first Brillouin zone, computed using four different hybrid descriptors G2B4. Experimental data (61; 62) is depicted with black points; the LML calculations are shown with solid green lines. The inserts with gray background schematically illustrate the structure of the hybrid descriptors (bottom left corner) and their hierarchical performance (top right corner).

potential parameters. However, the accuracy of descriptors strongly impacts the confidence limit of the fit. The overfitting margin is flexible and can be pushed by increasing the dimensions of the descriptor space and/or by changing the type of descriptor.

- In case it is not possible to obtain the target properties without falling into the overfit region, the potential development should propose more complex methods beyond LML regression, for instance, such as a recently proposed quadratic SNAP (50), which represents an intermediate step between the linear and more complex kernel methods.

- Within the confidence margins of the LML fit, hybrid descriptors can effectively boost numerical cost of the potential. The newly proposed G2B4 descriptor, $\mathbf{D} = \mathbf{G}_2 \oplus \mathbf{B}$, can be three to ten times faster than the most commonly used bispectrum descriptor. The mixing proportions between the hybrid descriptor components control the balance between numerical cost and accuracy of the potential. The choice of the hybrid descriptor should be driven by the objective physical problem and by the desirable precision of the fit. We emphasize that our hybrid descriptor is not unique and other combinations can be tested in the future: e.g., the radial $\mathbf{G}_2$ can be replaced by other descriptors like powerspectrum SO(3) or even EAM-like functions.

- Hybrid descriptors enable a hierarchical approach for the the molecular dynamics explo-

ration of complex energetic landscapes. A set of ML potentials based on hybrid descriptors with gradual increasing precision can be used for the free energy calculation methods, such as thermodynamical integration, perturbation or adaptive biasing potential or force dynamics. For example, in the case of thermodynamical integration, a ML potential combined with fast hybrid descriptors can be used as a more robust reference state instead of the commonly used harmonic approximation or even Einstein crystal.

Acknowledgments

This work was financially supported by the Cross-Disciplinary Program on Numerical Simulation of CEA, the French Alternative Energies and Atomic Energy Commission.

Appendix A. Derivative of bispectrum

The derivative of bispectrum coefficients (Eq. 7) with respect to the Cartesian coordinates of the atom b along the direction α becomes:

$$\frac{B^a_{ll_1l_2}}{\partial x_{b\alpha}} = \sum_{m',m=-l}^{l} \sum_{m'_1,m_1=-l_1}^{l_1} \sum_{m'_2,m_2=-l_2}^{l_2} \frac{\partial c^{a*}_{j;m'm}}{\partial x_{b\alpha}} C^{ll_1l_2}_{m'm'_1m'_2} C^{ll_1l_2}_{mm_1m_2} c^{a}_{l_1;m'_1m_1} c^{a}_{l_2;m'_2m_2} + c^{a*}_{j;m'm} C^{ll_1l_2}_{m'm'_1m'_2} C^{ll_1l_2}_{mm_1m_2} \left(\frac{\partial c^{a}_{l_1;m'_1m_1}}{\partial x_{b\alpha}} c^{a}_{l_2;m'_2m_2} + U^{i}_{l_1;m'_1m_1} \frac{\partial c^{a}_{l_2;m'_2m_2}}{\partial x_{b\alpha}} \right) . \quad \text{(A.1)}$$

The derivatives of the coefficients $c_{jmm'}$ can be deduced from Eq. 6. If the distance between the atoms a and b is beyond the cutoff distance, i.e., the atom b is not in the neighborhood of the central atom a, the derivative of $c_{jmm'}$ is zero. If $b = a$, the derivative becomes:

$$\frac{\partial c^{a}_{j;m'm}}{\partial x_{a\alpha}} = \sum_{c\in v(a)} t_{aa}(\mathbf{r}_{ac}) , \quad \text{(A.2)}$$

$$t_{ab}(\mathbf{r}_{ac}) = f'_c(r_{ac}) w_c U_{j,mm'}(\mathbf{r}_{ac}) \frac{\partial r_{ac}}{\partial x_{b\alpha}} + f_c(r_{ac}) w_c \frac{\partial U_{j,mm'}(\mathbf{r}_{ac})}{\partial x_{b\alpha}} . \quad \text{(A.3)}$$

If the atom b is in the neighborhood of the atom a, and $b \neq a$, the derivative of $c_{jmm'}$ can be written as:

$$\frac{\partial c^{a}_{j;m'm}}{\partial x_{b\alpha}} = -t_{ab}(\mathbf{r}_{ab}) . \quad \text{(A.4)}$$

In the Eqs. A.3-A.4, the derivatives of $c_{jmm'}$ coefficients are determined if the functions $U_{j,mm'}(\mathbf{r})$ and their derivatives $\partial U_{j,mm'}(\mathbf{r})/\partial x_\alpha$ are known. Both types of functions can be determined using recurrence relations, which, for the functions $U_{j,mm'}(\mathbf{r})$ with $m' \neq j$, are ((14; 48):

$$U_{j;mm'}(\mathbf{r}) = \left(\frac{j-m}{j-m'}\right)^{1/2} z_+ U_{j-\frac{1}{2};m+\frac{1}{2}m'+\frac{1}{2}}(\mathbf{r}) - \imath \left(\frac{j+m}{j-m'}\right)^{1/2} x_- U_{j-\frac{1}{2};m-\frac{1}{2}m'+\frac{1}{2}}(\mathbf{r}) , \quad \text{(A.5)}$$

While for the $U_{j,mm'}(\mathbf{r})$ with $m' \neq -j$, the following relation can be used:

$$U_{j;mm'}(\mathbf{r}) = \left(\frac{j+m}{j+m'}\right)^{1/2} z_- U_{j-\frac{1}{2};m-\frac{1}{2}m'-\frac{1}{2}}(\mathbf{r}) - \imath \left(\frac{j-m}{j+m'}\right)^{1/2} x_+ U_{j-\frac{1}{2};m+\frac{1}{2}m'-\frac{1}{2}}(\mathbf{r}), \quad \text{(A.6)}$$

with the first terms of the recurrence that define the upper terms:

$$\begin{aligned} U_{0,00}(\mathbf{r}) &= 0, \\ U_{\frac{1}{2};\pm\frac{1}{2}\pm\frac{1}{2}}(\mathbf{r}) &= \frac{1}{2}\frac{z_0 \mp iz}{l_0}, \\ U_{\frac{1}{2};\pm\frac{1}{2}\mp\frac{1}{2}}(\mathbf{r}) &= -\imath\frac{x \pm iy}{l_0}, \end{aligned} \quad \text{(A.7)}$$

where

$$\begin{cases} z_\pm = \cos(\theta_0) \pm \imath \sin(\theta_0)\cos(\theta) = \frac{z_0 \pm \imath z}{l_0}, \\ x_\pm = \sin(\theta_0)\sin(\theta) e^{\pm \imath\phi} = \frac{x \pm \imath y}{l_0}, \end{cases}$$

with $z_0 = r/\tan\theta_0$ and $l_0 = r/\sin\theta_0$. Consequently, the derivative of $U_{j,mm'}$ (Eqs.A.5 and A.6) becomes:

$$\frac{\partial U_{j,mm'}}{\partial x_\alpha} = \begin{cases} \sqrt{\frac{j-m}{j-m'}} \cdot \left(\frac{\partial z_+}{\partial x_\alpha} U^{j-\frac{1}{2}}_{m+\frac{1}{2}m'+\frac{1}{2}} + z_+ \cdot \frac{\partial U^{j-\frac{1}{2}}_{m+\frac{1}{2}m'+\frac{1}{2}}}{\partial x_\alpha} \right) \\ -\imath\sqrt{\frac{j+m}{j-m'}} \cdot \left(\frac{\partial x_-}{\partial x_\alpha} U^{j-\frac{1}{2}}_{m-\frac{1}{2}m'+\frac{1}{2}} + x_- \frac{\partial U^{j-\frac{1}{2}}_{m-\frac{1}{2}m'+\frac{1}{2}}}{\partial x_\alpha} \right) (m' \neq j) \\ \sqrt{\frac{j+m}{j+m'}} \cdot \left(\frac{\partial z_-}{\partial x_\alpha} U^{j-\frac{1}{2}}_{m-\frac{1}{2}m'-\frac{1}{2}} + z_- \frac{\partial U^{j-\frac{1}{2}}_{m-\frac{1}{2}m'-\frac{1}{2}}}{\partial x_\alpha} \right) \\ -\imath\sqrt{\frac{j-m}{j+m'}} \cdot \left(\frac{\partial x_+}{\partial x_\alpha} U^{j-\frac{1}{2}}_{m+\frac{1}{2}m'-\frac{1}{2}} + x_+ \frac{\partial U^{j-\frac{1}{2}}_{m+\frac{1}{2}m'-\frac{1}{2}}}{\partial x_\alpha} \right) (m' \neq -j) \end{cases}$$

The derivatives of $x_\pm$ and $z_\pm$ functions can be computed as:

$$\frac{\partial x_\pm}{\partial x_\alpha} = \begin{cases} (x \pm \imath y) \cdot \frac{\partial(1/l_0)}{\partial x_\alpha} \ (\text{if } \alpha = z) \\ (x \pm \imath y) \cdot \frac{\partial(1/l_0)}{\partial x_\alpha} + \frac{1}{l_0} \ (\text{if } \alpha = x) \\ (x \pm \imath y) \cdot \frac{\partial(1/l_0)}{\partial x_\alpha} \pm \frac{\imath}{l_0} \ (\text{if } \alpha = y) \end{cases}$$

$$\frac{\partial(1/l_0)}{\partial x_\alpha} = \left(\frac{\cos\theta_0}{r_0} - \frac{1}{l_0}\right)\frac{x_\alpha}{r^2},$$

$$\frac{\partial z_\pm}{\partial x_\alpha} = \begin{cases} (z_0 \pm \imath z) \cdot \frac{\partial(1/l_0)}{\partial x_\alpha} + \frac{1}{l_0}\frac{\partial z_0}{\partial x_\alpha} \ (\text{if } \alpha = x, y) \\ (z_0 \pm \imath z) \cdot \frac{\partial(1/l_0)}{\partial x_\alpha} + \frac{1}{l_0}\frac{\partial z_0}{\partial x_\alpha} \pm \frac{\imath}{l_0} \ (\text{if } \alpha = z) \end{cases}$$

$$\frac{\partial z_0}{\partial x_\alpha} = \left(\frac{1}{\tan\theta_0} - \frac{\theta_0}{\sin^2\theta_0} \right) \frac{x_\alpha}{r} \,.$$

References

[1] R. P. Gupta, Lattice relaxation at a metal surface, Phys. Rev. B 23 (1981) 6265–6270. doi:10.1103/PhysRevB.23.6265.

[2] G. J. Ackland, R. Thetford, An improved N-body semi-empirical model for body-centred cubic transition metals, Philos. Mag. A 56 (1) (1987) 15. doi:10.1080/01418618708204464.

[3] V. Rosato, M. Guillope, B. Legrand, Thermodynamical and structural properties of f.c.c. transition metals using a simple tight-binding model, Philos. Mag. A 59 (2) (1989) 321–336. doi:10.1080/01418618908205062.

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

[5] M. S. Daw, S. M. Foiles, M. I. Baskes, The embedded-atom method: a review of theory and applications, Mater. Sci. Reports 9 (7-8) (1993) 251. doi:10.1016/0920-2307(93)90001-U.

[6] M. I. Baskes, Modified embedded-atom potentials for cubic materials and impurities, Phys. Rev. B 46 (5) (1992) 2727. doi:10.1103/PhysRevB.46.2727.

[7] J. S. Luo, B. Legrand, Multilayer relaxation at surfaces of body-centered-cubic transition metals, Phys. Rev. B 38 (1988) 1728–1733. doi:10.1103/PhysRevB.38.1728.

[8] D. G. Pettifor, New many-body potential for the bond order, Phys. Rev. Lett. 63 (22) (1989) 2480. doi:10.1103/PhysRevLett.63.2480.

[9] D. G. Pettifor, Bonding and structure of molecules and solids, Oxford University Press, Oxford, 1996.

[10] A. P. Horsfield, A. M. Bratkovsky, M. Fearn, D. G. Pettifor, M. Aoki, Bond-order potentials: Theory and implementation, Phys. Rev. B 53 (19) (1996) 12694.

[11] M. C. Desjonquères, D. Spanjaard, Concepts in Surface Physics, Springer-Verlag, New York, 1993.

[12] M. W. Finnis, Interatomic forces in condensed matter., Oxford University Press, Oxford, 2003.

[13] J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. 98 (2007) 146401. doi:10.1103/PhysRevLett.98.146401.

[14] A. P. Bartók, Gaussian approximation potential : an interatomic potential derived from first principles quantum mechanics, Ph.D. thesis, University of Cambridge (2009).

[15] A. P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B 87 (2013) 184115. doi:10.1103/PhysRevB.87.184115.

[16] A. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials, Multiscale Model. Sim. 14 (3) (2016) 1153–1173. doi:10.1137/15M1054183.

[17] E. V. Podryabinkin, A. V. Shapeev, Active learning of linearly parametrized interatomic potentials, Comput. Mater. Sci. 140 (2017) 171–180. doi:10.1016/j.commatsci.2017.08.031.

[18] G. Ferré, J.-B. Maillet, G. Stoltz, Permutation-invariant distance between atomic configurations, J. Chem. Phys. 143 (10) (2015) 104114. doi:10.1063/1.4930541.

[19] G. Ferré, T. Haut, K. Barros, Learning molecular energies using localized graph kernels, J. Chem. Phys. 146 (11) (2017) 114107. doi:10.1063/1.4978623.

[20] J. Behler, Perspective: Machine learning potentials for atomistic simulations, J. Chem. Phys. 145 (17) (2016) 170901. doi:10.1063/1.4966192.

[21] E. Cubuk, S. Schoenholz, J. Rieser, B. Malone, J. Rottler, D. Durian, E. Kaxiras, A. Liu, Identifying Structural Flow Defects in Disordered Solids Using Machine-Learning Methods, Phys. Rev. Lett. 114 (10) (2015) 108001. doi:10.1103/PhysRevLett.114.108001.

[22] A. P. Bartók, G. Csányi, Gaussian approximation potentials: A brief tutorial introduction, Int. J. Quantum Chem. 115 (16) (2015) 1051–1057. doi:10.1002/qua.24927.

[23] T. Hofmann, B. Schölkopf, A. J. Smola, Kernel methods in machine learning, Ann. Statist. 36 (3) (2008) 1171–1220. doi:10.1214/009053607000000677.

[24] V. Botu, R. Batra, J. Chapman, R. Ramprasad, Machine Learning Force Fields: Construction, Validation, and Outlook, J. Phys. Chem. C 121 (1) (2017) 511–522. doi:10.1021/acs.jpcc.6b10908.

[25] Z. Li, J. R. Kermode, A. De Vita, Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces, Phys. Rev. Lett. 114 (9) (2015) 096405. doi:10.1103/PhysRevLett.114.096405.

[26] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, M. Ceriotti, Machine learning unifies the modeling of materials and molecules, Sci. Adv. 3 (12) (2017) e1701816. doi:10.1126/sciadv.1701816.

[27] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, Big Data of Materials Science: Critical Role of the Descriptor, Phys. Rev. Lett. 114 (10) (2015) 105503. doi:10.1103/PhysRevLett.114.105503.

[28] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, Phys. Rev. Lett. 104 (2010) 136403. doi:10.1103/PhysRevLett.104.136403.

[29] A. Thompson, L. Swiler, C. Trott, S. Foiles, G. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, J. Comp. Phys. 285 (2015) 316 – 330. doi:https://doi.org/10.1016/j.jcp.2014.12.018.

[30] 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, Phys. Rev. Materials 2 (2018) 013808. doi:10.1103/PhysRevMaterials.2.013808.

[31] W. J. Szlachta, A. P. Bartók, G. Csányi, Accuracy and transferability of gaussian approximation potential models for tungsten, Phys. Rev. B 90 (2014) 104108. doi:10.1103/PhysRevB.90.104108.

[32] M. A. Wood, A. P. Thompson, Quantum-accurate molecular dynamics potential for tungsten, arXiv:1702.07042v1 [physics.comp-ph].

[33] A. P. Thompson, P. A. Schultz, P. S. Crozier, S. G. Moore, L. P. Swiler, J. A. Stephens, C. R. Trott, S. M. Foiles, G. J. Tucker, Sandia report: Automated Algorithms for Quantum-Level Accuracy in Atomistic Simulations: LDRD Final Report, Sandia National Laboratories, 2014.

[34] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (1996) 11169–11186. doi:10.1103/PhysRevB.54.11169.

[35] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson, M. C. Payne, First principles methods using CASTEP, Zeitschrift fur Kristallographie 220 (2005) 567–570. doi:10.1524/zkri.220.5.567.65075.

[36] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. 134 (7) (2011) 074106. doi:10.1063/1.3553717.

[37] V. Botu, R. Ramprasad, Adaptive machine learning framework to accelerate ab initio molecular dynamics, Int. J. Quantum Chem. 115 (16) (2015) 1074–1083. doi:10.1002/qua.24836.

[38] R. Kobayashi, D. Giofré, T. Junge, M. Ceriotti, W. A. Curtin, Neural network potential for Al-Mg-Si alloys, Phys. Rev. Materials 1 (2017) 053604. doi:10.1103/PhysRevMaterials.1.053604.

[39] R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, M. Parrinello, Graphite-diamond phase coexistence study employing a neural-network mapping of the ab initio potential energy surface, Phys. Rev. B 81 (2010) 100103. doi:10.1103/PhysRevB.81.100103.

[40] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, M. Parrinello, Microscopic origins of the anomalous melting behavior of sodium under high pressure, Phys. Rev. Lett. 108 (2012) 115701. doi:10.1103/PhysRevLett.108.115701.

[41] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, M. Parrinello, Ab initio quality neural-network potential for sodium, Phys. Rev. B 81 (2010) 184107. doi:10.1103/PhysRevB.81.184107.

[42] G. C. Sosso, G. Miceli, S. Caravati, J. Behler, M. Bernasconi, Neural network interatomic potential for the phase change material GeTe, Phys. Rev. B 85 (2012) 174103. doi:10.1103/PhysRevB.85.174103.

[43] S. Kondati Natarajan, T. Morawietz, J. Behler, Representing the potential-energy surface of protonated water clusters by high-dimensional neural network potentials, Phys. Chem. Chem. Phys. 17 (2015) 8356–8371. doi:10.1039/C4CP04751F.

[44] T. Morawietz, J. Behler, A density-functional theory-based neural network potential for water clusters including van der waals corrections, The Journal of Physical Chemistry A 117 (32) (2013) 7356–7366. doi:10.1021/jp401225b.

[45] V. Botu, R. Ramprasad, Learning scheme to predict atomic forces and accelerate materials simulations, Phys. Rev. B 92 (9) (2015) 094306. doi:10.1103/PhysRevB.92.094306.

[46] R. Kakarala, Phd Thesis: The bispectrum as a source of phase-sensitive invariants for Fourier descriptors: a group-theoretic approach, Irvine University, 1992.

[47] R. Kondor, A complete set of rotationally and translationally invariant features for images, CoRR abs/cs/0701127. arXiv:cs/0701127.
URL http://arxiv.org/abs/cs/0701127

[48] D. Varshalovich, A. N. Moskalev, V. K. Khersonskii, Quantum Theory of Angular Momentum, Singapore World Scientific, 1988.

[49] A. M. Goryaeva, W. Unn-Toc, M. C. Marinica, MiLaDy - Machine Lerning Dynamics, CEA, Saclay, 2015-2018.

[50] M. A. Wood, A. P. Thompson, Extending the accuracy of the snap interatomic potential form, J. Chem. Phys. 148 (24). doi:10.1063/1.5017641.

[51] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, M. Ceriotti, Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials, J. Chem. Phys. 148 (24) (2018) 241730. doi:10.1063/1.5024611.

[52] G. H. Golub, C. F. van Loan, Matrix Computation, Fourth Edition, The Johns Hopkins University Press, 2013.

[53] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, S. P. Ong, Accurate force field for molybdenum by machine learning large materials data, Phys. Rev. Materials 1 (2017) 043603. doi:10.1103/PhysRevMaterials.1.043603.

[54] E. Darve, A. Pohorille, Calculating free energies using average force, J. Chem. Phys. 115 (20) (2001) 9169–9183. doi:10.1063/1.1410978.

[55] C. Chipot, J. Hénin, Exploring the free-energy landscape of a short peptide using an average force, J. Chem. Phys. 123 (24) (2005) 244906. doi:10.1063/1.2138694.

[56] T. Lelièvre, M. Rousset, G. Stoltz, Computation of free energy profiles with parallel adaptive dynamics, J. Chem. Phys. 126 (13) (2007) 134111. doi:10.1063/1.2711185.

[57] T. Lelièvre, G. Stoltz, M. Rousset, Free energy computations: A mathematical perspective, Imperial College Press, 2010.

[58] A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli, M. Parrinello, Assessing the Accuracy of Metadynamics, The Journal of Physical Chemistry B 109 (14) (2005) 6714–6721. doi:10.1021/jp045424k.

[59] A. Barducci, G. Bussi, M. Parrinello, Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method, Phys. Rev. Lett. 100 (2) (2008) 020603. doi:10.1103/PhysRevLett.100.020603.

[60] L. Bonati, M. Parrinello, Silicon Liquid Structure and Crystal Nucleation from Ab Initio Deep Metadynamics, Phys. Rev. Lett. 121 (26) (2018) 265701. doi:10.1103/PhysRevLett.121.265701.

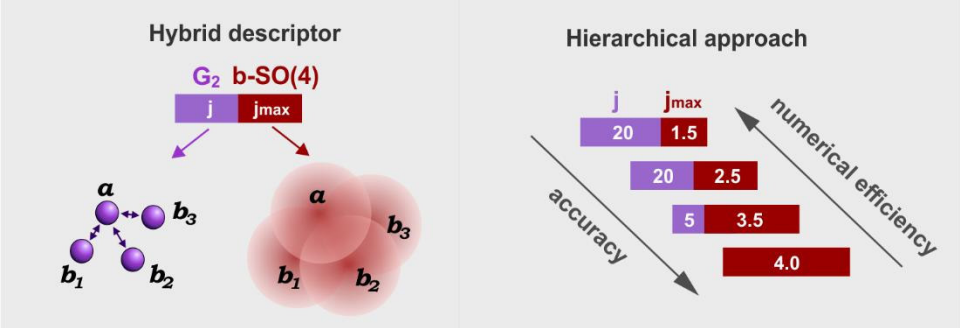
[61] S. H. Chen, B. N. Brockhouse, Lattice vibrations of tungsten, Solid State Commun. 2 (3) (1964) 73–77. doi:10.1016/0038-1098(64)90043-2.

[62] A. Larose, B. N. Brockhouse, Lattice vibrations in tungsten at 22 C studied by neutron scattering, Can. J. Phys. 54 (17) (1976) 1819–1823. doi:10.1139/p76-215.

A.M. Goryaeva: Investigation, Conceptualization, Software, Methodology, Writing - original draft, Writing -review & editing, Visualization

J.-B. Maillet: Conceptualization, Supervision.

M.-C. Marinica: Funding acquisition, Supervision, Conceptualization, Software, Methodology, Writing - original draft, Writing -review & editing


Hybrid descriptor
G2 b-SO(4)
j
jmax
a
b3
b1
b2
a
b3
b1
b2
Hierarchical approach
j
jmax
20
1.5
20
2.5
5
3.5
4.0
numerical efficiency
accuracy