Atomic cluster expansion for accurate and transferable interatomic potentials

Ralf Drautz
ICAMS, Ruhr-Universität Bochum, 44780 Bochum, Germany

(Received 28 September 2018; revised manuscript received 16 December 2018; published 8 January 2019)

The atomic cluster expansion is developed as a complete descriptor of the local atomic environment, including multicomponent materials, and its relation to a number of other descriptors and potentials is discussed. The effort for evaluating the atomic cluster expansion is shown to scale linearly with the number of neighbors, irrespective of the order of the expansion. Application to small Cu clusters demonstrates smooth convergence of the atomic cluster expansion to meV accuracy. By introducing nonlinear functions of the atomic cluster expansion an interatomic potential is obtained that is comparable in accuracy to state-of-the-art machine learning potentials. Because of the efficient convergence of the atomic cluster expansion relevant subspaces can be sampled uniformly and exhaustively. This is demonstrated by testing against a large database of density functional theory calculations for copper.

DOI: 10.1103/PhysRevB.99.014104

I. INTRODUCTION

It is straightforward to write down the energy of a collection of atoms $i = 1, \ldots, N$ in a many-atom expansion

$$E = V_0 + \sum_i V^{(1)}(\boldsymbol{r}_i) + \frac{1}{2}\sum_{ij} V^{(2)}(\boldsymbol{r}_i, \boldsymbol{r}_j) + \frac{1}{3!}\sum_{ijk} V^{(3)}(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k) + \frac{1}{4!}\sum_{ijkl} V^{(4)}(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_l) + \cdots, \quad (1)$$

where $\boldsymbol{r}_i$ corresponds to the position of atom i and the potentials $V^{(2)}, V^{(3)}, \ldots$ are symmetric, uniquely defined, and zero if two or more indices take identical values [1]. V_0 is a constant offset that may be set to zero and $V^{(1)}$ is the chemical potential. A separation of the energy into atomic contributions follows

$$E_i = V^{(1)}(\boldsymbol{r}_i) + \frac{1}{2}\sum_{j} V^{(2)}(\boldsymbol{r}_i, \boldsymbol{r}_j) + \frac{1}{6}\sum_{jk} V^{(3)}(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k) + \frac{1}{24}\sum_{jkl} V^{(4)}(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k, \boldsymbol{r}_l) + \cdots. \quad (2)$$

Apart from a few important examples [2,3] this many-atom expansion has rarely been used in the construction of interatomic potentials for materials. The convergence of the expansion is slow and, for example, for bulk metals potentials $V^{(K)}$ up to $K > 15$ are required [1]. Next, even if a cutoff is introduced in order to take into account only nearby atoms within a cutoff distance r_c, the evaluation of the leading term of order $K + 1$ in the sum Eq. (2) scales as N_c^K, where N_c corresponds to a typical number of neighbors within the cutoff sphere. For accurate potentials one requires cutoffs that in close-packed materials imply $N_c \approx 10^2 \ldots 10^3$, which makes it challenging to sum expansions beyond about $K = 3$ within an acceptable time.

One may take the view that many of the developments in the field of interatomic potentials over the past decades were focused implicitly or explicitly on finding efficient approximate representations of higher-order contributions $V^{(K)}$. Typically these are obtained as nonlinear functions of functions that contain fewer than K atomic positions. For example, the assumption of a constant semi-infinite recursion chain [4] leads to the second-moment potentials [5–7] and justifies the square-root embedding function in the Finnis-Sinclair potential [8],

$$E_i = \sqrt{\rho_i} + \frac{1}{2}\sum_j V(r_{ji}), \quad (3)$$

where $\rho_i = \sum_j \phi(r_{ji})$ is the local density of atomic sites and ϕ and V are pairwise functions of interatomic distance between atoms i and j. The atomic energy is therefore a nonlinear function of coordination and density, an observation [9,10] that has motivated the embedded atom method [11] where the square-root function of the Finnis-Sinclair potential is replaced by a general, quasiconcave embedding function and the site density is regarded as an approximate electron density. Tersoff maintained a nonlinear embedding function for each bond and included an explicit dependence on relative bond angles to be able to describe directional bond formation in semiconductors [12,13], so that the energy is written as a nonlinear function that depends on a three-body contribution $\rho_i = \sum_{jk} \phi(\boldsymbol{r}_{ji}, \boldsymbol{r}_{ki})$. In the modified embedded atom method [14], Baskes introduced angular terms in the originally pairwise density of the embedded atom method. Since then several other potentials that take into account angular dependence and are written as a nonlinear function of a sum over small atomic clusters have been published. These potentials are generally referred to as cluster functionals [15]. The analytic bond-order potentials provide a rigorous derivation of cluster functionals

for semiconductors [16,17] and metals [18,19] from the tight-binding approximation [20,21] and recover the Tersoff and Finnis-Sinclair potential, respectively, at the lowest order of approximation.

As today tens of thousands of density functional theory (DFT) [22,23] calculations can be carried out routinely [24], sufficient data are available for the application of methods from statistical learning that enable the interpolation of high-dimensional data sets. This has led to the adoption of machine learning to the development of interatomic potentials, such as neural networks potentials [25] or Gaussian process regression for the Gaussian approximation potentials [26]. The resulting potentials are generally called machine learning potentials; the field is very active with many recent developments [27–42].

The machine learning potentials employ a descriptor that quantifies the local atomic environment. The atomic energy or other atomic properties are then learned as a nontrivial function of the descriptor by training with a reference data set. The machine learning potentials reproduce DFT reference data sets with excellent accuracy and are currently considered to be significantly more accurate than cluster functionals. As machine learning potentials are not derived or motivated by physical or chemical intuition, the excellent accuracy of the machine learning potentials comes at the cost of interpretability, and the machine learning potentials are generally regarded as black box models.

In this paper I show how to obtain a cluster functional with an accuracy that rivals machine learning potentials but with a simple functional form that is amenable to physical and chemical interpretation. To this end I will first introduce the atomic cluster expansion as a general and complete descriptor of the local atomic environment in Sec. II. I will further show that the expansion scales linearly with the number of neighbors N_c and therefore overcomes the poor N_c^K scaling of general many-atom potentials. Popular descriptors are then discussed in the light of this expansion in Sec. III and it is shown that these descriptors may be cast in the form of the atomic cluster expansion. By combining a physically motivated functional form with a relatively low dimensional atomic cluster expansion, the curse of dimensionality, i.e., the impossibility to sample a high-dimensional space uniformly, that is immanent to general methods from machine/statistical learning, is overcome.

The parametrization of the nonlinear atomic cluster expansion is demonstrated for copper and validated against a comprehensive dataset of DFT calculations in Secs. IV and V. The completeness of the atomic cluster expansion means that in principle it is possible to converge the cluster functional to arbitrary accuracy. In Sec. VI I conclude.

II. ATOMIC CLUSTER EXPANSION

I extend the spin cluster expansion [43] to a complete descriptor of local atomic environments. The spin cluster expansion was obtained as a generalization of the lattice cluster expansion [44] to continuous degrees of freedom, such as the direction of magnetic moments [45]. The cluster expansion is related to Hadamard and multidimensional discrete Fourier and wavelet transforms [46].

Here our interest is in the energy or another property of atom i,

$$E_i(\boldsymbol{\sigma}) = E_i(\boldsymbol{r}_{1i}, \boldsymbol{r}_{2i}, \ldots, \boldsymbol{r}_{Ni}), \tag{4}$$

which is completely characterized by the $N-1$ vectors from atom i to all other atoms, $\boldsymbol{r}_{ji} = \boldsymbol{r}_j - \boldsymbol{r}_i$. The collection of the $N-1$ vectors is abbreviated as the configuration $\boldsymbol{\sigma} = (\boldsymbol{r}_{1i}, \boldsymbol{r}_{2i}, \ldots, \boldsymbol{r}_{Ni})$ of atom i and the order of entries in $\boldsymbol{\sigma}$ is irrelevant. The inner product between two functions $f(\boldsymbol{\sigma})$ and $g(\boldsymbol{\sigma})$ is then defined as

$$\langle f|g\rangle = \int f^*(\boldsymbol{\sigma}) g(\boldsymbol{\sigma})\, d\boldsymbol{\sigma}. \tag{5}$$

Next a set of orthogonal and complete basis functions $\phi_v(\boldsymbol{r})$ with $v = 0, 1, 2, \ldots$ that depend only on a single bond $\boldsymbol{r}$ are introduced,

$$\int \phi_v^*(\boldsymbol{r})\phi_u(\boldsymbol{r})\, d\boldsymbol{r} = \delta_{vu}, \tag{6}$$

$$\sum_v \phi_v^*(\boldsymbol{r})\phi_v(\boldsymbol{r}') = \delta(\boldsymbol{r} - \boldsymbol{r}'). \tag{7}$$

The basis functions for the expansion of the atomic energy Eq. (4) are obtained from the product of single-bond basis functions. By choosing $\phi_0 = 1$ a hierarchical expansion is obtained.

A cluster α with K elements contains K bonds $\alpha = (j_1 i, j_2 i, \ldots, j_K i)$, where the order of entries in α does not matter, and the vector $v = (v_1, v_2, \ldots, v_K)$ contains the list of single-bond basis functions in the cluster. Only single-bond basis functions with $v > 0$ are considered in v. The cluster basis function is given by

$$\Phi_{\alpha v} = \phi_{v_1}(\boldsymbol{r}_{j_1 i})\phi_{v_2}(\boldsymbol{r}_{j_2 i}) \ldots \phi_{v_K}(\boldsymbol{r}_{j_K i}), \tag{8}$$

with $0 \leqslant K \leqslant N-1$. The orthogonality and completeness of the one-bond basis functions transfers to the cluster basis functions

$$\langle \Phi_{\alpha v}|\Phi_{\beta\mu}\rangle = \delta_{\alpha\beta}\delta_{v\mu}, \tag{9}$$

$$1 + \sum_{\gamma\subseteq\alpha}\sum_v \Phi_{\gamma v}^*(\boldsymbol{\sigma})\Phi_{\gamma v}(\boldsymbol{\sigma}') = \delta(\boldsymbol{\sigma} - \boldsymbol{\sigma}'), \tag{10}$$

where α is an arbitrary cluster and the right-hand side of the completeness relation is the product of the relevant right-hand sides of Eq. (7). A kernel may then be obtained as

$$k(\boldsymbol{\sigma}, \boldsymbol{\sigma}') = 1 + \sum_{\gamma v} \Phi_{\gamma v}^*(\boldsymbol{\sigma})\Phi_{\gamma v}(\boldsymbol{\sigma}'), \tag{11}$$

and the expansion of the atomic energy Eq. (4) is written in the form

$$E_i(\boldsymbol{\sigma}) = \langle k(\boldsymbol{\sigma}, \boldsymbol{\sigma}')|E_i(\boldsymbol{\sigma}')\rangle = J_0 + \sum_{\alpha v} J_{\alpha v}\Phi_{\alpha v}(\boldsymbol{\sigma}). \tag{12}$$

The expansion coefficients $J_{\alpha v}$ are obtained by projection

$$J_{\alpha v} = \langle \Phi_{\alpha v}|E_i(\boldsymbol{\sigma})\rangle. \tag{13}$$

The expansion holds for multicomponent systems. In Appendix A this is made explicit by adding an index for the chemical species. As will be discussed in the following, the atomic cluster expansion is conveniently parametrized in a

nonorthogonal basis. The extension of the cluster expansion to a nonorthogonal basis is summarized in Appendix B.

Next I will simplify the expansion by exploiting invariance with respect to permutations of bonds and further show how the expression Eq. (12) may be evaluated in a way that scales linearly with the number of neighbors.

A. Permutation of bonds

If atoms j and k are of the same chemical species then an exchange of the bonds ji and ki leaves the energy or any other atomic observable unchanged. This means that the bonds in a cluster can be grouped by chemical species. If a cluster with K bonds contains $k_A, k_B, \ldots$ bonds to chemical species A, B, …, then by invariance with respect to permutations the expansion coefficient $J_{\alpha\nu}$ is fully characterized by the number of bonds that it contains to a specific chemical species, $J_{\alpha\nu} = J_{k_A k_B \ldots \nu}$. Here I discuss the application of the atomic cluster expansion to elements. In this case only the number of bonds in the cluster is required for the characterization of the expansion coefficient and therefore, for the case of elemental materials, one may write

$$J_{\alpha\nu} = J_\nu^{(K)}. \tag{14}$$

Equation (12), or Eq. (B8) for a nonorthogonal expansion, may then be rewritten as

$$\begin{aligned} E_i(\boldsymbol{\sigma}) = &\sum_j \sum_\nu J_\nu^{(1)} \phi_\nu(\boldsymbol{r}_{ji}) \\ &+ \frac{1}{2} \sum_{j_1 j_2}^{j_1 \neq j_2} \sum_{\nu_1 \nu_2} J_{\nu_1 \nu_2}^{(2)} \phi_{\nu_1}(\boldsymbol{r}_{j_1 i}) \phi_{\nu_2}(\boldsymbol{r}_{j_2 i}) \\ &+ \frac{1}{3!} \sum_{j_1 j_2 j_3}^{j_1 \neq j_2, \cdots} \sum_{\nu_1 \nu_2 \nu_3} J_{\nu_1 \nu_2 \nu_3}^{(3)} \phi_{\nu_1}(\boldsymbol{r}_{j_1 i}) \phi_{\nu_2}(\boldsymbol{r}_{j_2 i}) \phi_{\nu_3}(\boldsymbol{r}_{j_3 i}) \\ &+ \cdots, \end{aligned} \tag{15}$$

and where I set $J_0 = 0$ such that from correspondence to Eq. (2) it is evident that the cluster expansion coefficient $J_\nu^{(K)}$ contributes to the potential $V^{(K+1)}$.

The expansion Eq. (15) may be rewritten in a slightly different way with unrestricted sums and updated expansion coefficients,

$$\begin{aligned} E_i(\boldsymbol{\sigma}) = &\sum_j \sum_\nu c_\nu^{(1)} \phi_\nu(\boldsymbol{r}_{ji}) \\ &+ \frac{1}{2} \sum_{j_1 j_2} \sum_{\nu_1 \nu_2} c_{\nu_1 \nu_2}^{(2)} \phi_{\nu_1}(\boldsymbol{r}_{j_1 i}) \phi_{\nu_2}(\boldsymbol{r}_{j_2 i}) \\ &+ \frac{1}{3!} \sum_{j_1 j_2 j_3} \sum_{\nu_1 \nu_2 \nu_3} c_{\nu_1 \nu_2 \nu_3}^{(3)} \phi_{\nu_1}(\boldsymbol{r}_{j_1 i}) \phi_{\nu_2}(\boldsymbol{r}_{j_2 i}) \phi_{\nu_3}(\boldsymbol{r}_{j_3 i}) \\ &+ \cdots. \end{aligned} \tag{16}$$

The expansion Eq. (16) is identical to Eq. (15), with expansion coefficients $c_\nu^{(K)}$ that are different from the expansion coefficients $J_\nu^{(K)}$ in Eq. (15) and that may be obtained from $J_\nu^{(K)}$ as follows. Due to the unrestricted sums in Eq. (16) products of single-bond basis functions $\phi_{i\nu}$ that contain the same bonds are obtained, i.e., expressions of the form $\phi_{\nu_1}(\boldsymbol{r}_{ji}) \phi_{\nu_2}(\boldsymbol{r}_{ji})$ that are not part of the expansion Eq. (15) and that may be understood as unphysical self-interactions. Because of the completeness of the single-bond basis functions these products may be re-expanded at lower order, for example, $\sum_\nu a_\nu \phi_\nu(\boldsymbol{r}_{ji}) = \phi_{\nu_1}(\boldsymbol{r}_{ji}) \phi_{\nu_2}(\boldsymbol{r}_{ji})$, etc., such that the self-interactions are removed by an appropriate modification of a lower-order expansion coefficient.

B. Atomic base and linear scaling

As argued in Sec. I the summation of the many-atom potentials quickly becomes numerically prohibitively expensive as it scales as N_c^K. For the expansion Eq. (16) this may be avoided by a simple reordering of summations. I define the atomic base as the projection of the basis functions on the atomic density

$$A_{i\nu} = \langle \varrho_i | \phi_\nu \rangle = \sum_j \phi_\nu(\boldsymbol{r}_{ji}), \tag{17}$$

with the atomic density of an elemental material

$$\varrho_i = \sum_j \delta(\boldsymbol{r} - \boldsymbol{r}_{ji}). \tag{18}$$

The atomic energy Eq. (16) then becomes a polynomial in $A_{i\nu}$,

$$\begin{aligned} E_i(\boldsymbol{\sigma}) = &\sum_\nu c_\nu^{(1)} A_{i\nu} + \sum_{\nu_1 \nu_2}^{\nu_1 \geqslant \nu_2} c_{\nu_1 \nu_2}^{(2)} A_{i\nu_1} A_{i\nu_2} \\ &+ \sum_{\nu_1 \nu_2 \nu_3}^{\nu_1 \geqslant \nu_2 \geqslant \nu_3} c_{\nu_1 \nu_2 \nu_3}^{(3)} A_{i\nu_1} A_{i\nu_2} A_{i\nu_3} + \cdots. \end{aligned} \tag{19}$$

As the expansion coefficients $c_\nu^{(K)}$ are fully symmetric with respect to permutations of indices $\nu_1, \nu_2, \nu_3, \ldots$, in Eq. (19) an ordered summation $\nu_1 \geqslant \nu_2 \geqslant \nu_3$ was used.

The numerical effort required for the construction of the atomic base $A_{i\nu}$ is linear with the number of neighbors N_c, while the evaluation of the energy in the form of Eq. (19) is independent of N_c. This means that the time required for the evaluation of the energy scales linearly with the number of neighbors independent of the order of the expansion. The atomic cluster expansion therefore overcomes the poor scaling of the formal many-atom expansion Eq. (2). This is absolutely critical for a fast evaluation of higher-order terms in close-packed materials with hundreds of atoms within the cutoff sphere.

The time for the evaluation of the forces also scales linearly with the number of neighbors; see Sec. II D. The atomic cluster expansion is illustrated graphically in Fig. 1.

C. Basis functions

The second-moment, Finnis-Sinclair, and embedded atom method potentials have shown that volume-dependent terms provide by far the largest contribution to the cohesive energy. Volume contributions are in general significantly larger than angular contributions. Therefore I choose to use the following single-bond basis that separates radial and angular contributions and that has the structure of a linear combination of atomic orbitals basis set that is commonly used in electronic

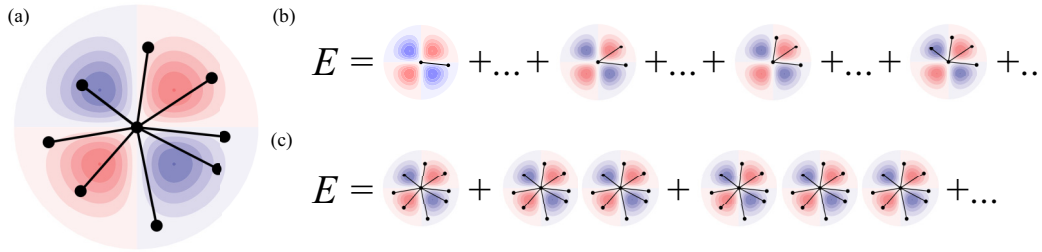

FIG. 1. Illustration of the atomic cluster expansion. (a) Basis function evaluated for bonds to neighbors of the central atom. (b) Atomic cluster expansion as sum of pair, three-body, four-body, etc., contributions, Eq. (15). (c) Atomic cluster expansion as polynomial of the atomic base, Eq. (19), that scales linearly with the number of neighbors irrespective of the order of the expansion.

structure calculations

$$\phi_v(\boldsymbol{r}) = \sqrt{4\pi}\, R_{nl}(r) Y_l^m(\hat{\boldsymbol{r}}), \tag{20}$$

with $v = (nlm)$, the radial functions R_{nl} that depend on distance $r = |\boldsymbol{r}|$, and the spherical harmonics Y_l^m that are functions of the bond direction $\hat{\boldsymbol{r}} = \boldsymbol{r}/r$. In principle any other complete basis set could be used, such as, for example, the hyperspherical harmonics on the unit sphere in four dimensions; see Refs. [31,47].

In the following I will use basis functions that form a complete set but are neither normalized nor orthogonal and therefore require a nonorthogonal atomic cluster expansion Eq. (B8).

1. Radial basis and finite interaction range

Here I adopt the usual assumption that the direct interaction between two atoms vanishes when they are more than the cutoff distance r_c apart. Atoms that move outside or inside the cutoff distance may only change the energy smoothly. I use a radial basis that is adapted to this and built on a complete set of polynomials,

$$g_0 = 1, \tag{21}$$

$$g_1(r) = [1 + \cos(\pi r/r_c)], \tag{22}$$

$$g_k(r) = \tfrac{1}{4}[1 - T_{k-1}(x)][1 + \cos(\pi r/r_c)], \tag{23}$$

for $k = 2, 3, \ldots,$ with the Chebyshev polynomials of the first kind T_k and the scaled distance x that enables an exponentially more dense sampling for smaller distances r,

$$x = 1 - 2\left(\frac{e^{-\lambda(r/r_c - 1)} - 1}{e^{\lambda} - 1}\right). \tag{24}$$

The scaling helps to reproduce steep gradients at small interatomic distances and exponentially smaller gradients at larger distances with few radial functions while the slow decay of the cosine envelope function allows us to estimate the local atomic density and provides a smooth cutoff. The first eight basis functions are shown in Fig. 2. The Chebyshev polynomials and some of their properties are summarized in the Supplemental Material [48]. I also experimented with other radial functions, in particular Gaussians with varying exponents, but found the resulting interpolations to be more wavy and less smooth.

The radial basis functions R_{nl} are then expanded as

$$R_{00}(r) = g_0 = 1, \tag{25}$$

$$R_{0l}(r) = 0 \text{ for } l > 0, \tag{26}$$

$$R_{nl}(r) = \sum_{k=1} c_{nlk} g_k(r) \text{ for } n > 0, \tag{27}$$

with expansion coefficients c_{nlk} and where R_{00} is not used explicitly but corresponds to $\phi_0 = 1$ and is required for the completeness of the single-bond basis functions and the hierarchical structure of the atomic cluster expansion. Radial functions $R_{0l}(r)$ that do not depend on distance but on bond angles cannot contribute to a local expansion and are therefore set to zero.

2. Angular basis and rotational invariance

Without an external field the atomic energy is invariant with respect to rotation of the global coordinate system. This requires that in the expansion Eq. (19) only products of spherical harmonics are taken into account that correspond to a reducible representation of the identity of the rotation group. The reduction of products of spherical harmonics is related closely to the coupling of angular momenta and has been discussed in detail in literature [47,49–51] and is briefly reviewed in Appendix C.

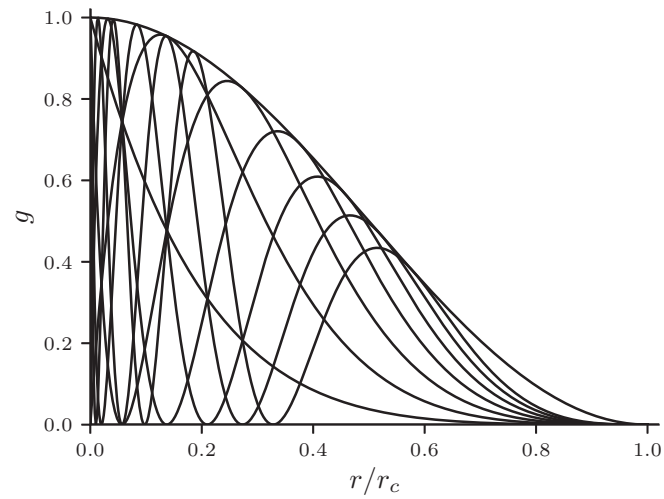

FIG. 2. Radial functions g_1 to g_8 for $\lambda = 5$.

For two-body terms, corresponding to single-bond basis functions, only radial functions can contribute as Y_{00} is the only spherical harmonic that is invariant with respect to rotation. For three-body terms, corresponding to the product of two single-bond basis functions, only contributions with $l_1 = l_2$ and $m_1 = -m_2$ can couple to zero angular momentum. Four-body terms corresponding to the product of three spherical harmonics are conveniently reduced with the help of the Wigner $3j$ symbol or the Clebsch-Gordan coefficients. (The Wigner $3j$ symbol and the Clebsch-Gordan coefficients are summarized in the Supplemental Material [48].) For five-body and six-body terms I obtain analogous couplings of four and five spins, respectively. The necessary equations are summarized in Appendix C.

Invariance with respect to inversion further requires that only products of spherical harmonics may contribute for which the sum of angular momentum quantum numbers adds to an even number.

For rotational invariance products of the atomic base Eq. (17) are reduced as

$$B^{(1)}_{in} = A_{in00}, \tag{28}$$

$$B^{(2)}_{in_1n_2l} = \sum_{m=-l}^{l} (-1)^m A_{in_1lm} A_{in_2l-m}, \tag{29}$$

$$B^{(3)}_{\substack{in_1n_2n_3\\ l_1l_2l_3}} = \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} \sum_{m_3=-l_3}^{l_3} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \times A_{in_1l_1m_1} A_{in_2l_2m_2} A_{in_3l_3m_3}, \tag{30}$$

$$B^{(4)}_{\substack{in_1n_2n_3n_4\\ l_1l_2l_3l_4}} = \sum_{m_1m_2m_3m_4} \begin{bmatrix} l_1 & l_2 & l_3 & l_4 \\ m_1 & m_2 & m_3 & m_4 \end{bmatrix} \times A_{in_1l_1m_1} A_{in_2l_2m_2} A_{in_3l_3m_3} A_{in_4l_4m_4}, \tag{31}$$

$$B^{(5)}_{\substack{in_1n_2n_3n_4n_5\\ l_1l_2l_3l_4l_5}} = \sum_{m_1m_2m_3m_4m_5} \begin{bmatrix} l_1 & l_2 & l_3 & l_4 & l_5 \\ m_1 & m_2 & m_3 & m_4 & m_5 \end{bmatrix} \times A_{in_1l_1m_1} A_{in_2l_2m_2} A_{in_3l_3m_3} A_{in_4l_4m_4} A_{in_5l_5m_5}. \tag{32}$$

The matrix elements of the Wigner $3j$ symbol in Eq. (30) are given in the Supplemental Material [48] and the sum is limited to contributions that fulfill $m_1 + m_2 + m_3 = 0$. Equivalent expressions are provided in Appendix C for the five-body and six-body basis functions $B^{(4)}$ and $B^{(5)}$. Only contributions that fulfill $m_1 + m_2 + m_3 + m_4 = 0$ and $m_1 + m_2 + m_3 + m_4 + m_5 = 0$, respectively, are different from zero. While the spherical harmonics are complex valued, the reduced product basis functions $B^{(K)}$ are real valued.

Therefore, in summary Eq. (19) is reduced to a complete expansion for the atomic energy that is invariant with respect to translation, rotation, and inversion,

$$E_i(\boldsymbol{\sigma}) = \sum_n c^{(1)}_n B^{(1)}_{in} + \sum_{n_1n_2l}{}' c^{(2)}_{n_1n_2l} B^{(2)}_{in_1n_2l} + \sum_{\substack{n_1n_2n_3\\ l_1l_2l_3}}{}' c^{(3)}_{\substack{n_1n_2n_3\\ l_1l_2l_3}} B^{(3)}_{\substack{in_1n_2n_3\\ l_1l_2l_3}} + \sum_{\substack{n_1n_2n_3n_4\\ l_1l_2l_3l_4}}{}' c^{(4)}_{\substack{n_1n_2n_3n_4\\ l_1l_2l_3l_4}} B^{(4)}_{\substack{in_1n_2n_3n_4\\ l_1l_2l_3l_4}} + \sum_{\substack{n_1n_2n_3n_4n_5\\ l_1l_2l_3l_4l_5}}{}' c^{(5)}_{\substack{n_1n_2n_3n_4n_5\\ l_1l_2l_3l_4l_5}} B^{(5)}_{\substack{in_1n_2n_3n_4n_5\\ l_1l_2l_3l_4l_5}} + \cdots. \tag{33}$$

The summations are taken over ordered sets of indices only and for $K \geqslant 3$ only contributions are considered for which the sum of angular momentum quantum numbers adds to an even number.

In a more compact notation that implies summation over basis set indices Eq. (33) may be written as

$$E_i(\boldsymbol{\sigma}) = \sum_{K\boldsymbol{nl}} \boldsymbol{c}^{(K)}_{\boldsymbol{nl}} \boldsymbol{B}^{(K)}_{i\boldsymbol{nl}}. \tag{34}$$

This completes the derivation of the atomic cluster expansion and Eq. (34) is the working equation for the remainder of this paper. Clearly not only the energy but any function of the local atomic environment may be expanded in the form of an atomic cluster expansion and I will use this in Sec. V to expand the density $\rho_i(\boldsymbol{\sigma})$ in the Finnis-Sinclair potential Eq. (3).

D. Forces

Forces may be computed with a similar efficiency as the energy. I derive expressions for the force gradients of the form

$$\boldsymbol{F}_k = -\nabla_k E = -\nabla_k \sum_i F\left(\rho^{(p)}_i\right), \tag{35}$$

where F is differentiable and each of the functions $\rho^{(p)}_i$, $p = 1, 2, 3, \ldots$, is represented by an atomic cluster expansion. From Eq. (34) follows

$$\boldsymbol{F}_k = -\sum_{inlm} \sum_p \sum_{K\boldsymbol{nl}} c^{(K,p)}_{\boldsymbol{nl}} \frac{\partial F}{\partial \rho^{(p)}_i} \frac{\partial B^{(K)}_{i\boldsymbol{nl}}}{\partial A_{inlm}} \nabla_k A_{inlm}. \tag{36}$$

The derivative of the atomic base Eq. (17) is given by

$$\nabla_k A_{inlm} = \nabla_k \phi_{nlm}(\boldsymbol{r}_{ki}) + (-1)^l \sum_j \nabla_k \phi_{nlm}(\boldsymbol{r}_{ij}) \delta_{ik}, \tag{37}$$

where I made use of $\phi_{nlm}(\boldsymbol{r}_{ji}) = (-1)^l \phi_{nlm}(\boldsymbol{r}_{ij})$. I introduce

$$w_{inlm} = \sum_{K\boldsymbol{nl}} \left(\sum_p c^{(K,p)}_{\boldsymbol{nl}} \frac{\partial F}{\partial \rho^{(p)}_i} \right) \frac{\partial B^{(K)}_{i\boldsymbol{nl}}}{\partial A_{inlm}}, \tag{38}$$

such that the forces may be written as

$$\boldsymbol{F}_k = -\sum_{inlm} [w_{inlm} + (-1)^l w_{knlm}] \nabla_k \phi_{nlm}(\boldsymbol{r}_{ki}). \tag{39}$$

Efficient expressions for the gradients of the basis functions $\nabla_k \phi_{nlm}(\boldsymbol{r}_{ki})$ may be obtained. This is briefly outlined in the Supplemental Material [48].

E. Explicit many-atom interactions

The basis functions $B^{(K)}$ contain the many-atom interactions inherent to the atomic cluster expansion. The

single-bond basis function $B^{(1)}_{in}$ depends only on the pairwise distance r_{ji} and therefore corresponds to pairwise contributions,

$$B^{(1)}_{in} = R_{n0}(r_{ji}). \tag{40}$$

The basis functions $B^{(K)}$ with $K > 1$ may be decomposed into explicit many-atom interactions. This is most easily demonstrated for the two-bond basis function $B^{(2)}$ by making use of the addition theorem for spherical harmonics

$$P_l(\hat{\boldsymbol{r}}_i\hat{\boldsymbol{r}}_j) = \frac{4\pi}{2l+1}\sum_{m=-l}^{l}(-1)^m Y_l^m(\hat{\boldsymbol{r}}_i)Y_l^{-m}(\hat{\boldsymbol{r}}_j), \tag{41}$$

with the Legendre polynomials P_l that are summarized in the Supplemental Material [48]. Therefore the three-body contributions are rewritten as

$$B^{(2)}_{in_1n_2l} = \frac{1}{2l+1}\sum_{jk} R_{n_1l}(r_{ji})R_{n_2l}(r_{ki})P_l(\cos\theta_{jik}). \tag{42}$$

Higher-order terms may be decomposed in analogy to the spin cluster expansion [43] as discussed in Refs. [49,52]. For example, $B^{(3)}$ contains contributions of the form

$$R_{n_1l_1}(r_{ji})R_{n_2l_2}(r_{ki})R_{n_3l_3}(r_{li})P_L(\cos\theta_{jik})P_M(\cos\theta_{jil}), \tag{43}$$

etc. The decomposition of the basis functions $B^{(K)}$ leads to a representation of the atomic cluster expansion in the form of the general many-atom expansion Eq. (2). While the explicit decomposition of $B^{(K)}$ is instructive for insight into the many-atom expansion, it destroys the linear scaling of the atomic cluster expansion with the number of neighbors, Sec. II B, and is therefore not of interest for a numerically efficient implementation.

III. RELATION TO OTHER DESCRIPTORS

The atomic cluster expansion in the form of Eq. (34) provides a hierarchical and complete descriptor of the local atomic environment and one can therefore expand other descriptors with respect to the atomic cluster expansion. In the following I will discuss some of the most popular descriptors in the light of the atomic cluster expansion.

A. Steinhardt parameters

The Steinhardt parameters [53] are frequently employed to differentiate between local atomic environments. They have also been used as the basis for the expansion of an interatomic potential [3]. The second-order and third-order bond parameters Q_l and W_l for atom i can be generated from the atomic cluster expansion if only a single radial function is used. This radial function is constant, $R_1(r) = 1$ for $r < r_c$, and zero otherwise. The Steinhardt parameters expressed in the atomic cluster expansion are

$$Q_l = \left(\frac{B^{(2)}_{i11l}}{(2l+1)N_c^2}\right)^{1/2}, \tag{44}$$

$$W_l = \frac{B^{(3)}_{i111lll}}{(4\pi)^{3/2}N_c^3}. \tag{45}$$

B. Correction electron densities

The correction electron densities of the modified embedded atom method, Eqs. (8a)–(8d) in Ref. [14], are written as products of distance dependent functions and cosines of bond angles. They contain up to two bond distances and one bond angle and may be expanded in the atomic cluster expansion using basis functions $B^{(1)}$ and $B^{(2)}$.

C. Behler and Parrinello symmetry functions

The symmetry functions of Behler and Parrinello form the basis for their neural network interatomic potential [25,30,54]. The symmetry functions $G^1_i = \sum_j f_c(r_{ji})$ and $G^2_i = \sum_j \exp(-\eta(r_{ji}-r_s)^2)f_c(r_{ji})$ are two-body functions with a cosine or tanh envelope function f_c that can be expanded in basis functions $B^{(1)}$. The symmetry functions $G^3_i = 2^{1-\zeta}\sum'_{jk}(1+\lambda\cos\theta_{ijk})^\zeta \exp[-\eta(r^2_{ji}+r^2_{kj}+r^2_{ki})]f_c(r_{ji})$ $f_c(r_{kj})f_c(r_{ki})$ and $G^4_i = 2^{1-\zeta}\sum'_{jk}(1+\lambda\cos\theta_{ijk})^\zeta \exp[-\eta$ $(r^2_{ji}+r^2_{ki})]f_c(r_{ji})f_c(r_{ki})$ are three-body functions and closely related to $B^{(2)}$ as represented in Eq. (42) if the radial functions are parametrized by Gaussians with a cosine or tanh envelope function and angular contributions of the type $(1+\lambda\cos\theta)^\zeta$ are expanded in terms of Legendre polynomials.

If instead the angular dependence is expanded in Chebyshev polynomials a descriptor termed angular Fourier series [47] is obtained that may be represented in basis functions $B^{(2)}$ by expanding the Chebyshev polynomials in Legendre polynomials.

D. Bispectrum and smooth overlap of atomic positions

The bispectrum [55] was discussed in Ref. [47] as a generalization of the Steinhardt parameters. The bispectrum is formulated as a product of two or three terms, respectively, where each term is a sum of spherical harmonics that are then contracted with Clebsch-Gordan coefficients. The Clebsch-Gordan coefficients and Wigner $3j$ matrices are closely related [48] and for our considerations here equivalent. In order that the following equations match the expressions given in Ref. [47], $B^{(3)}$ needs to be redefined from Eq. (30) to an entirely equivalent form by replacing the Wigner $3j$ symbol with Eq. (C4).

If in the most general equations for the bispectrum [Eq. (26) of Ref. [47]] the expansion coefficients c_{nlm}, which in Ref. [47] are the projections of the atomic density on basis functions $g_n(r)Y_l^m(\hat{\boldsymbol{r}})$, are rewritten as the atomic base Eq. (17), then the equations given for the bispectrum are equivalent to the basis functions $B^{(2)}_{in_1n_2l}$ and $B^{(3)}_{in_1n_2n_2l_1l_2l_3}$, where in Ref. [47] two of the three radial functions were taken to be identical to keep the number of variables small and the nonorthogonal version of the atomic cluster expansion of Appendix B has to be used.

If the atomic cluster expansion is carried out in the basis of hyperspherical harmonics, which corresponds to a four-dimensional spin cluster expansion [43], then the general expressions for the four-dimensional bispectrum of Ref. [47] may be obtained analogously from $B^{(2)}$ and $B^{(3)}$.

The smooth overlap of atomic positions (SOAP) descriptor is obtained from a kernel based on the bispectrum and expressed in terms of the atomic cluster expansion in the form

$$k(\varrho, \varrho') = \sum_{n_1 n_2 l} B^{(2)}_{n_1 n_2 l}(\varrho) B^{(2)}_{n_1 n_2 l}(\varrho'), \tag{46}$$

for $n = 2$ and

$$k(\varrho, \varrho') = \sum_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}} B^{(3)}_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}}(\varrho) B^{(3)}_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}}(\varrho'), \tag{47}$$

for $n = 3$. These equations are part of the general kernel expansion Eq.(11) limited to a system that is invariant with respect to rotation and inversion and neglecting self-interaction corrections; see the discussion in Sec. II A.

E. Moments of the density of states

The moments of the electronic density of states depend on the atomic structure and the electronic structure of a material. The moments form the basis for several linear scaling implementations of DFT and the analytic bond order potentials [16,18,19] and are closely connected [56] to the recursion method [4], the kernel polynomial method [57–59], and the Fermi-operator expansion method [60,61]. The Nth moment of the local density of states $n_{i\alpha}(\epsilon)$ of orbital α on atom i is defined as

$$\begin{aligned} \mu^{(N)}_{i\alpha} &= \int \epsilon^N n_{i\alpha}(\epsilon)\, d\epsilon \\ &= \sum_{\substack{i_2\alpha_2 i_3\alpha_3 \\ \dots i_{N-1}\alpha_{N-1}}} H_{i\alpha i_2\alpha_2} H_{i_2\alpha_2 i_3\alpha_3} \dots H_{i_{N-1}\alpha_{N-1} i\alpha}, \end{aligned} \tag{48}$$

where $H_{i\alpha j\beta} = \langle i\alpha|\hat{H}|j\beta\rangle$ are Hamiltonian matrix elements, the second identity is the moments theorem [62], and an orthonormal basis was assumed.

A moment of order N comprises a maximum of N different sites; it can therefore be expanded in an atomic cluster expansion of order $K = N - 1$. Furthermore, as the product Eq. (48) contains many paths that visit a particular atom more than once, the atomic cluster expansion can represent parts of moments with $K < N - 1$. The self-returning fourth-moment hopping paths are, for example, important for the structural stability of Si [16,17]. Recent work has used the moments of the density of states explicitly as descriptors [63,64].

F. Spectral neighbor analysis potential

The spectral neighbor analysis potential (SNAP) [31] adds a term linear in the bispectrum to an existing reference potential. As discussed in Sec. III D the bispectrum corresponds to the four-body basis functions $B^{(3)}$ of the atomic cluster expansion. Therefore the linear SNAP may be viewed as part of the linear many-atom expansion Eq. (2). A recent improvement has found improved convergence of the SNAP by adding terms that are quadratic in the bispectrum [32].

G. Moment tensor potentials

The moment tensor potentials [65] are a polynomial expansion of the linear many-atom series Eq. (2). They may be viewed as a generalization of the correction electron densities from the modified embedded atom method [14] to arbitrary order and are closely related to the expansion of the energy in a cosine series [52].

The basis functions of the moment tensor potentials are obtained by contraction of so-called moments tensors, which are schematically defined as

$$M = \sum_j f(r_{ji})\, \boldsymbol{r}_{ji} \otimes \boldsymbol{r}_{ji} \otimes \dots \otimes \boldsymbol{r}_{ji}, \tag{49}$$

with a radial function f that only depends on the distance r_{ji}. If, without loss of generality, I absorb the magnitude of the vectors in a new radial function $\tilde{f}$,

$$M = \sum_j \tilde{f}(r_{ji})\, \hat{\boldsymbol{r}}_{ji} \otimes \hat{\boldsymbol{r}}_{ji} \otimes \dots \otimes \hat{\boldsymbol{r}}_{ji}, \tag{50}$$

it becomes explicit that for rotational invariance the contraction of M must produce reducible representations of the identity representation of the rotation group. Angular terms up to six-body interactions for a cosine expansion were given explicitly in Ref. [52]. These terms may be expanded in the spin-cluster expansion [43], as briefly discussed in Refs. [52] and [49]. The moment tensor potentials may also be cast in the form of an atomic cluster expansion, Eq. (34).

IV. INTERATOMIC POTENTIAL FOR COPPER

The atomic cluster expansion will be demonstrated for Cu. For copper an exhaustive study using a neural network potential [54] is available which enables a comparison to a state-of-the-art machine learning potential. The radial basis introduced in Sec. II C 1 has two parameters that need to be set: the cutoff distance r_c and the exponential decay λ, Eq. (24). The cutoff for Cu was set to $r_c = 7.75$ Å. None of the clusters and periodic structures in the reference data set with all bond lengths larger than or equal to r_c showed interactions larger than 0.3 meV/atom. The exponential decay parameter was set to $\lambda = 5.25$; good fits were obtained for a range of λ from about 4.5 to 5.5, where a large value is more efficient for reproducing steep slopes in the interatomic interaction at short interatomic distances.

A. Loss function and optimization

The target of the optimization was to obtain an error that approximately scales with the energy of the reference data. The loss function was therefore chosen as

$$J = \sum_n w_n \left(E_n - E_n^{\mathrm{ref}}\right)^2 + \alpha_1 \sum_{K\boldsymbol{nl}} \left|c^{(K)}_{\boldsymbol{nl}}\right| + \alpha_2 \sum_{K\boldsymbol{nl}} \left|c^{(K)}_{\boldsymbol{nl}}\right|^2, \tag{51}$$

with weights

$$w_n = \omega / \left[E_n^{\mathrm{ref}} - E^{\mathrm{ref}}_{\mathrm{min}}(1 + \Delta)\right]^2. \tag{52}$$

The sum is taken over all N selected structures n in the reference data set; the energy E_n^{ref} corresponds to the energy per atom in the reference data set and E_n to the predicted

energy. A single atom in vacuum in the reference data set has energy $E = 0$. The regularization weights $\alpha_1 = 10^{-10}$ and $\alpha_2 = 10^{-14}$ were kept at very small values so as to have no significant effect on the root-mean-square error but helped to stabilize the optimization numerically. The normalization factor ω was set to ensure $\sum_n w_n = 1$. The energy $E^{\rm ref}_{\rm min}$ refers to the structure with the smallest energy per atom from the reference data set. Without the parameter Δ the weight would scale inversely quadratic with the difference to the lowest energy structure and diverge for the lowest energy structure. The parameter $\Delta = 0.2$ shifts this divergence to smaller values. The weights w_n allow one to work with heterogeneous reference data, i.e., data with very high energy of hundreds of eV per atom together with structures that are only a few meV away from the ground-state energy.

The RMSE(0) root-mean-square error that I will discuss in the following was obtained from the N_0 structures that had an energy per atom lower than zero and without any further weighting, i.e., $\text{RMSE}(0) = \sqrt{\sum_n (E^{\rm ref}_n - E_n)^2 / N_0}$, the RMSE($\Delta$1) root-mean-square error was obtained from all structures that are within 1 eV/atom from the lowest energy structure. The mean average error is given as $\text{MAE} = \sum_n |E^{\rm ref}_n - E_n| / N$.

The optimization was carried out hierarchically by adding parameters in small steps. First a two-body, from this a three-body, etc., parametrization was obtained. A standard Levenberg-Marquardt algorithm [66,67] as implemented in MINPACK [68] was sufficient for the robust minimization of the loss function.

Only energies were used for the parametrization of the atomic cluster expansion and potentials. For the evaluation of the energy the basis functions $B^{(K)}$ can be pre-computed, therefore the time for the calculation of the energy during parametrization is on the order of 10^{-6} s per atom (of which most is spent for fetching data from memory) while for the evaluation of the forces expressions of the same complexity as $B^{(K)}$ need to be updated in every step (see Sec. II D), which takes about 10^{-4} s per atom.

B. Reference data set

An exhaustive reference data set that consisted of 55 289 clusters and periodic structures was computed using DFT [22,23]. All DFT calculations were carried out with FHI aims [69,70] with "tight" basis settings that include 40 basis functions per Cu atom and strict self-consistency settings. Two identical reference data sets, one for the local-density approximation (LDA) in the parametrization of Refs. [71] and [72] and one employing the generalized gradient approximation in the parametrization of Ref. [73] ([Perdew, Burke, and Ernzerhof (PBE)] were computed, with more than 100 000 DFT calculations in total. The calculations were carried out with a Gaussian smearing of 0.10 eV. All calculations were nonmagnetic. Magnetism may be important for small clusters but was ignored here as the magnetic degrees of freedom introduce new energy hypersurfaces that cannot be captured as a function of atomic positions alone, while for nonmagnetic calculations one may assume that there is only a single Born-Oppenheimer surface.

The reference data were generated to cover the space of local atomic environments as completely as possible. In particular, this comprised small clusters including two, three, and four atoms. The configuration space of these clusters is one, three, and six dimensional, respectively. The configuration space of clusters with up to four atoms is sufficiently small to be sampled uniformly. Clusters with up to four atoms were sampled systematically for interatomic distances from $0.7r_s$ to $3.0r_s$, where the scaling length $r_s = 2.54$ Å is close to the nearest-neighbor distance in fcc Cu. Clusters with two to six atoms were further taken out of fcc, bcc, hcp, and diamond lattices if the shortest bond was not shorter than three times the longest bond in the cluster. The shortest bond length in the clusters was then varied between $0.7r_s$ to $3.0r_s$. Furthermore, clusters with 3 to 15 atoms were randomly generated with the constraint that none of the atoms may be closer than $0.7r_s$ to any other atom and more than $2.5r_s$ distant from any atom in the cluster.

Periodic structures were used with a uniform sampling of interatomic distances such that the smallest interatomic distance in the structures reached from $0.8r_s$ to $3.0r_s$. Only few standard structures (fcc, hcp, bcc, diamond, hexagonal diamond, simple hexagonal, simple cubic) were included. Most of the periodic calculations were carried out for random structures with one or two atoms in the unit cell that homogeneously cover the space of local atomic environments in small bulk structures [63]. The maximum number of neighbors within the cutoff distance for any atom in the reference set was $N_c = 306$. The cohesive energy of fcc copper for LDA and PBE are different by about 1 eV, corresponding to about 25% of the cohesive energy.

Apart from the scaling length r_s the reference structures did not integrate any information of a particular property of copper. The volume, which has by far the largest contribution to the atomic energy, was the only parameter that was explicitly sampled for periodic cells. In particular I did not generate structures using molecular dynamics simulations as the resulting structures are specific to copper and, depending on temperature, correlated for hundreds or thousands of time steps and sample only a small fraction of the space of local atomic environments. Molecular dynamics trajectories therefore could heavily bias the expansion coefficients and artificially lower the fitting errors.

C. Application to small clusters

I will first discuss the application of the atomic cluster expansion to small clusters. If the number of atoms in a cluster is smaller or equal to $K_{\rm max} + 1$ in the atomic cluster expansion, then reference data can be reproduced with the linear expansion Eq. (34). In this Sec. I will focus on two, three, and four-atom clusters for which the phase space of atomic configurations was sampled exhaustively. Because some of the clusters contained small interatomic distances together with large distances, numerical convergence was sometimes difficult and for about 5% of the DFT calculations self-consistency could not be achieved, resulting in 609 (575), 4322 (3809), 20 932 (18 864) converged DFT calculations with PBE for clusters with two, three, and four atoms respectively. The

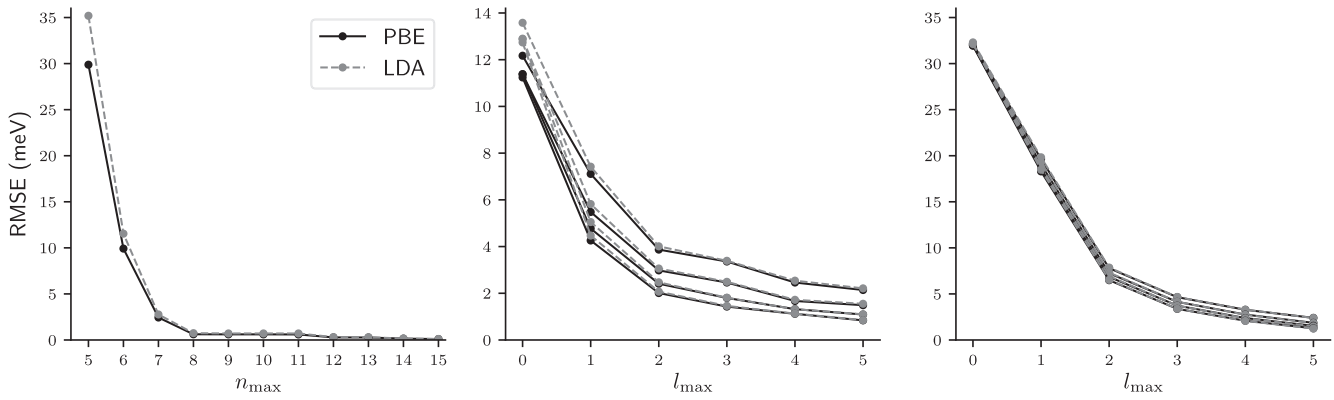

FIG. 3. RMSE(0) for two-atom clusters (left) as a function of $n_{\max}$, for three-atom clusters (middle), and four-atom clusters (right) as a function of $l_{\max}$ for $n_{\max} = 5, \ldots, 8$. A larger value of $n_{\max}$ leads to smaller RMSE(0) for a given $l_{\max}$.

number in brackets gives the size of the subset with an average atomic energy below zero, i.e., lower than the energy of a free, nonmagnetic Cu atom. For LDA the DFT calculations led to 609 (536), 4328 (3799), 20 941 (18 901) converged clusters.

For simplicity I assumed $R_{nl}(r) = g_n(r)$, or in other words $c_{nlk} = \delta_{nk}$ in Eq. (27). For every order K of the atomic cluster expansion I then used two parameters to characterize the expansion, the number of radial functions $n_{\max}$, and the maximum angular momentum $l_{\max}$ used in the expansion. The indices of the basis functions $B^{(1)}$, $B^{(2)}$, ... then took values $1, \ldots, n_{\max}$ and $0, \ldots, l_{\max}$ respectively.

Figure 3 shows the convergence of the atomic cluster expansion for the two-atom test set as a function of the maximum number of radial functions $n_{\max}$. The RMSE quickly converges to below 1 meV. The atomic cluster expansion for three-atom clusters and including the two-atom test set as a function of $l_{\max}$ and for different values of $n_{\max}$ is shown in the middle panel of Fig. 3. The RMSE is computed from all three-atom clusters that have an energy lower than zero. One can estimate the importance of the angular three-body contributions from the RMSE at $l_{\max} = 0$ to be about 12 meV. The RMSE error smoothly converges with $l_{\max}$ and $n_{\max}$ and also for the three-atom clusters values of 1 meV are achieved. The performance for PBE and LDA is very similar, with slightly larger errors for LDA at small values of $l_{\max}$. In the right panel of Fig. 3 the convergence for the four-atom test set, including two- and three-atom clusters, is shown. From the RMSE at $l_{\max} = 0$ it appears that the angular four-body contributions are on the order of 33 meV per atom and therefore significantly larger than the angular three-body contributions. The RMSE converges smoothly with $l_{\max}$ and $n_{\max}$ and values of 1 meV are achieved. The convergence is particularly fast until $l_{\max} = 2$, which is consistent for a d-valent transition metal. Interestingly, the RMSE for PBE and LDA data sets appear to be nearly identical.

As discussed in the Introduction, one expects a slow convergence of a linear atomic cluster expansion with order K as the formation of chemical bonds is a highly nonlinear, cooperative process. In order to achieve a potential that is applicable to small atomic clusters as well as close-packed bulk materials it is therefore efficient to introduce a nonlinear expansion in the next section.

V. NONLINEAR ATOMIC CLUSTER EXPANSION

The atomic cluster expansion Eq. (33) is a linear expansion that has the same slow convergence for bulk structures as the general many-atom expansion discussed in the Introduction. Therefore, the energy of bulk materials cannot be expanded directly. Instead, I focus on the expansion of quantities for which a linear expansion is appropriate while a simple relation to the energy can also be identified. The tight binding (TB) approximation [20,21] is derived from DFT and will be used to motivate a nonlinear cluster functional. In TB one starts from free atoms. The free atoms are placed at the respective positions of atoms in a molecule or crystal; the overlap of the free atom charge densities leads to a repulsive energy between the atoms. As the free atom charge densities are spherical, the repulsive energy is dominated largely by pairwise contributions and to a lesser extent by many-atom contributions, which may originate, for example, from allowing the radial charge distribution of the free atoms to shrink when the atoms are immersed in the bulk [21,74]. The repulsive energy therefore should be suitable for a linear atomic cluster expansion. Next, in the TB approximation the atomic charge density is released, which leads to a rearrangement of the charge distribution and attractive bond formation that comprises hybridization of orbitals and charge transfer. Bond formation is a highly nonlinear process and here I will use very basic assumptions to motivate a functional form for the nonlinear atomic cluster expansion. Bond formation may be approximated as being driven by the band energy. Using the recursion expansion [4] any Hamiltonian may be mapped on a semi-infinite chain. A semi-infinite chain with constant matrix elements is known to have a square-root density of states, which rationalizes the square-root dependence of the atomic energy on the number of neighbors [75]. The Hamiltonian is well described by the two-center approximation [76] with relatively small many-atom contributions, so that the Hamiltonian may be viewed as another quantity suitable for a linear atomic cluster expansion. Clearly, any realistic Hamiltonian may not be mapped on a constant semi-infinite chain and the differences in the first few matrix elements are due to many-atom interactions. Here I assume that to first order a transformation that is required to relate the actual Hamiltonian to the Hamiltonian of a semi-infinite constant chain may be approximated as a linear

atomic cluster expansion, resulting in the following cluster functional:

$$E_i = -\sqrt{\rho_i^{(1)}\left(1+\rho_i^{(2)}\right)} + \rho_i^{(3)}, \tag{53}$$

where an atomic cluster expansion is carried out for $\rho^{(1)}$, $\rho^{(2)}$, and $\rho^{(3)}$ and they are loosely motivated by an environmentally dependent Hamiltonian, the bond order, and the repulsive energy, respectively. The relation to the expression for the Finnis-Sinclair potential [8] Eq. (3) is evident and Eq. (53) may be viewed as an extension of the Finnis-Sinclair potential to include many-atom interactions in ρ and V. Of course one can easily lift the constraint of a square-root embedding function, then extensions of the embedded atom method [11] and modified embedded atom method [14] potentials are obtained and links to the Tersoff potential [12,13] may be established.

The linear term $\rho^{(3)}$ further contains contributions due to nonbonding long-range interactions that are only partly accounted for in the PBE and LDA functionals.

In practice I replace the square-root function in Eq. (53) to allow for negative values of the argument,

$$F(x) = -\operatorname{sgn}(x)((1-y)^2\sqrt{|x|} + y^2|x|), \tag{54}$$

with $y = \exp(-\xi|x|)$ and where a value of $\xi = 20$ was used in the following. This was done purely for numerical stability, as in this way the embedding function varies smoothly when its argument changes sign and the value of ξ has no numerically significant influence on the results of the fit.

It should be emphasized that Eq. (53) is a choice that (1) can be motivated from physical considerations, (2) can be viewed as an extension of available cluster functionals, and (3) reproduces the reference data well, as will be discussed in the next section. Furthermore, it does not comprise any free parameters apart from the expansion coefficients of the atomic cluster expansion for $\rho^{(1)}$, $\rho^{(2)}$, and $\rho^{(3)}$. Other functions $E_i = E_i(\rho_i^{(1)}, \rho_i^{(2)}, \rho_i^{(3)}, \dots)$ should be explored in future work.

In terms of densities Eq. (53) is a three-dimensional function and it is possible to sample this three-dimensional space uniformly. This means that if the densities ρ are expanded as an atomic cluster expansion of moderate order, for example $K_{\max} = 3$, then the complete phase space available to the energy function may be sampled. This is in contrast to machine learning based potentials, where the expression for the energy is a function that has the dimensionality of the descriptor space, i.e., tens or hundreds of dimensions, and therefore cannot be sampled uniformly. Thus the nonlinear atomic cluster expansion avoids the curse of dimensionality that in general machine learning methods suffer from.

A. Parametrization for copper

The data set summarized in Sec. IV B was used for the parametrization of the nonlinear atomic cluster expansion for Cu, with weights and optimization as described in Sec. IV A. The parametrization was limited to $K_{\max} = 3$. Then the two-, three-, and four-atomic clusters largely define the two-, three-, and four-atom cluster expansion coefficients and the function Eq. (53) has to ensure transferability to the bulk. Because of the simple square-root embedding function in Eq. (53) one may expect the following behavior: a smooth convergence for small clusters with up to $K_{\max} + 1$ atoms on the one hand as well as for bulk structures on the other hand. The convergence of the expansion is expected to be worst for clusters with slightly more than $K_{\max} + 1$ atoms. To take this into account the parametrization was carried out for two different reference data sets. The first set contained all clusters with one, two, three, and four atoms as well as all periodic calculations (clu1-4+bulk). The second set further contained random clusters with more than four atoms, about 11 000 additional clusters with 5 to 15 atoms (clu1-15+bulk).

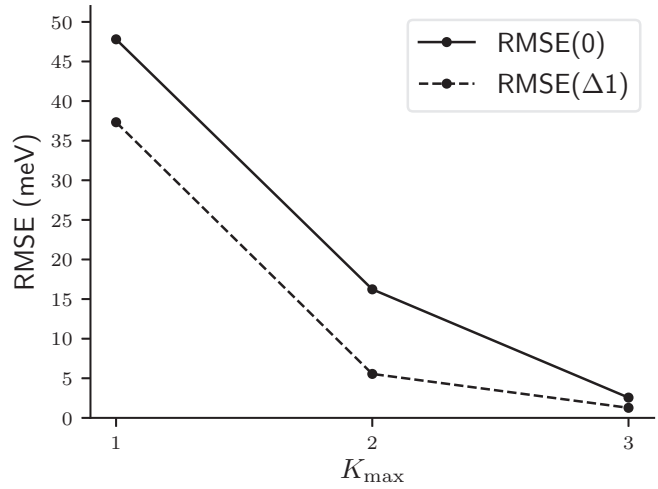

FIG. 4. Convergence of the nonlinear atomic cluster expansion as a function of $K_{\max}$ for the clu1-4+bulk data set. RMSE(0) is obtained from structures with an energy below zero, RMSE(Δ1) from structures within 1 eV/atom from the lowest energy structure.

For the parametrization I simply used radial functions $R_{nl}(r) = g_n(r)$. The following parameters were used in the expansion: two-atom interactions $n_{\max}^{(1)} = 8$, three-body interactions $n_{\max}^{(2)} = 6$ and $l_{\max}^{(2)} = 3$, four-body interactions $n_{\max}^{(3)} = 6$ and $l_{\max}^{(3)} = 3$. Figure 4 shows the convergence of the nonlinear atomic cluster expansion as a function of $K_{\max}$. For the clu1-4+bulk PBE data set RMSE(Δ1) = 1.3 meV and RMSE(0) = 2.6 meV are reached at $K_{\max} = 3$, i.e., when four-atom interactions are explicitly taken into account. For the clu1-15+bulk PBE data set RMSE(Δ1) = 2.0 meV and RMSE(0) = 8.4 meV. The numbers for the LDA data set are very close, although the cohesive energy in LDA is about 25% larger than in PBE. It is worth noting that at $K_{\max} = 2$ RMSE(Δ1) reaches about 5.5 meV for the clu1-4+bulk PBE data set and is therefore smaller than RMSE(0) at $K_{\max} = 3$ for the clu1-15+bulk PBE reference data.

Figure 5 shows the distribution of MAE as a function of the shortest bond in a cluster or periodic structure. The error decreases overall with $K_{\max}$, as expected from Fig. 4. The MAE becomes large when the shortest bond is significantly shorter than the equilibrium bond length in copper, which approximately corresponds to the scaling length $r_s = 2.54$ Å. The MAE of the two data sets are close except for distances smaller than the scaling length r_s. This means that the clu1-15+bulk data set contains more structures with short bond length that have a relatively high energy and therefore

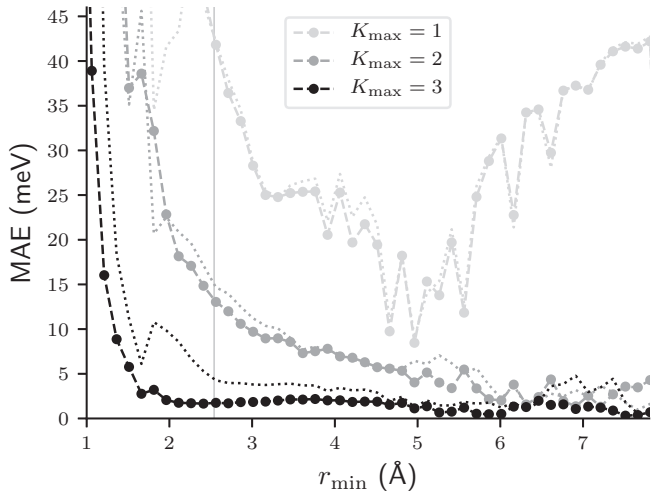

FIG. 5. Distribution of MAE as a function of the shortest distance in a structure or cluster for clu1-4+bulk (with symbols) and clu1-15+bulk (dotted) PBE data sets. The scaling length r_s is indicated with a vertical line.

get little weight in the loss function, Eq. (51), while they still contribute to RMSE(0). Next, as discussed above, in the many-atom interactions only radial basis functions up to $n_{\max} = 6$ were included, which possibly are insufficient for an accurate description of the steep slope in the interatomic interaction at short distances. Finally, also the DFT reference data are most susceptible to noise at steep slopes of the interatomic interaction, for example, due to k-space sampling or numerical integration of radial contributions and might contribute to the increase of the error at shorter distances.

The performance of the potential for the Boltzmann averaged error [77] is shown in Fig. 6. The Boltzmann averaged

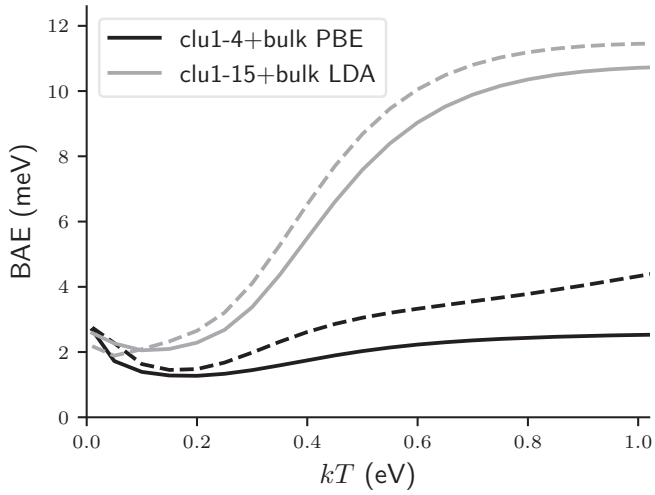

FIG. 6. Boltzmann averaged error BAE for clu1-4+bulk PBE and clu1-15+bulk LDA data sets. Shown are the training (solid) and test (dashed) errors.

error is defined as

$$\mathrm{BAE} = \sqrt{\frac{\sum_n \left(E_n - E_n^{\mathrm{ref}}\right)^2 \exp\left[-\left(E_n^{\mathrm{ref}} - E_{\min}^{\mathrm{ref}}\right)/kT\right]}{\sum_n \exp\left[-\left(E_n^{\mathrm{ref}} - E_{\min}^{\mathrm{ref}}\right)/kT\right]}}. \tag{55}$$

At low temperature kT the error displays the accuracy of the potential close to the ground state, while at high temperature kT the error approaches the RMSE and provides a measure for the transferability of the potential between structures, including high-energy structures. Figure 6 shows that the parametrization strategy that I used leads to small errors for small temperatures. Up to a temperature of about 0.25 eV, which corresponds approximately to 2900 K, the train and test errors are below 3 meV.

Figure 7 compares the reference energies of the clu1-4+bulk PBE data set to the prediction of the nonlinear atomic cluster expansion. The structures are marked with small symbols that overlap, therefore in essence the graph displays the outliers. The data set includes structures with very small interatomic distance that have an energy of several hundreds of eV per atom. These structures are well reproduced. The middle panel of Fig. 7 shows that there are no significant outliers with an energy up to 1 eV. The right panel is a 1000-times magnification of the left panel; the small scatter of the data along the diagonal is visible.

B. Discussion and comparison to Artrith and Behler

I have shown that a nonlinear cluster expansion with $K_{\max} = 3$ achieves an RMSE of 1–3 meV for copper. The expansion is applicable to small clusters as well as bulk materials. I have further illustrated that changes in reference data may lead to a factor of 2 or more in RMSE(0). It is common practice when training machine learning potentials that high-energy structures are discarded, where high energy is relative and sometimes all structures with energies more than 1 eV above the ground-state energy are neglected. As I have shown, this may lead to the conclusion that a three-body expansion performs better than a four-body expansion that is applied to a more complex data set. The differences in reference data and training strategies used by different authors make it hard to compare the performance of different potentials by the RMSE alone and highlight the importance of the details of the reference data set.

In Ref. [54] a neural network potential based on the Behler-Parrinello symmetry functions was developed. As discussed in Sec. III C the descriptors are two- and three-body functions, corresponding to $K = 1$ and $K = 2$. The neural network potential was trained to an extensive reference data set of DFT calculations for copper that included bulk structures, surfaces, and clusters. The reference data were generated with the FHI-aims code and PBE and employed an identical basis set as the present work. The neural network potential achieved a training RMSE of about 4 meV and a test error that was numerically slightly smaller.

As many structures in the reference data set were taken out of molecular dynamics simulations, it should be fair to assume that the data set contained only few structures with small

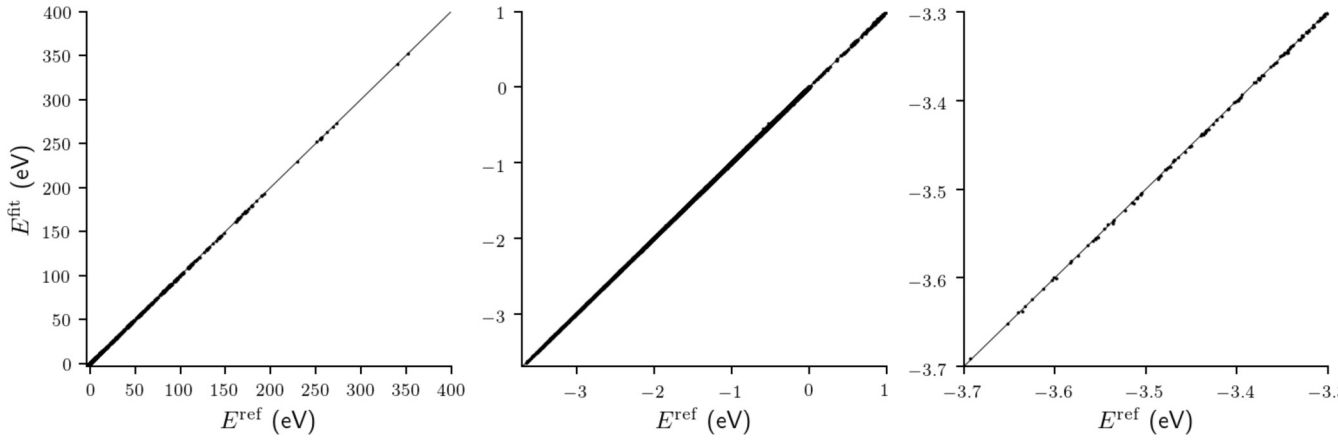

FIG. 7. Comparison of reference and fitted energy for the clu1-4+bulk PBE data set and $K_{\max} = 3$.

interatomic distances. This is also corroborated by Fig. 5 in Ref. [54] that shows only reference data with energies below zero. In other words, if I assume that the reference data set in Ref. [54] is broadly comparable to the clu1-4+bulk set, then the neural network potential with a descriptor up to $K_{\max} = 2$ and the nonlinear cluster expansion with $K_{\max} = 3$ perform approximately equally well.

It would be interesting to combine the atomic cluster expansion including terms $K = 3$ with a neural network or other machine learning method as one would expect that a sophisticated machine learning method should perform better for reproducing the DFT reference data than the simple expression Eq. (53).

VI. CONCLUSIONS

I have introduced the atomic cluster expansion as a general and formally complete descriptor of the local atomic environment and gave expressions for multicomponent systems and nonorthogonal basis functions. The time required for evaluating energy and forces from the atomic cluster expansion scales linearly with the number of neighbors, irrespective of the order of the expansion. The atomic cluster expansion therefore overcomes the problem of poor scaling of formal many-atom expansions. Several well-known and frequently used descriptors may be expanded using the atomic cluster expansion or are part of an atomic cluster expansion.

The atomic cluster expansion was demonstrated numerically for small clusters of copper, parametrized from a comprehensive DFT reference data set. In particular, the test set sampled the space of two-, three-, and four-atomic clusters exhaustively. The atomic cluster expansion was able to reproduce the test set for the small clusters with 1 meV accuracy. A linear expansion is not efficient for bridging from free atoms to bulk and I therefore introduced the nonlinear atomic cluster expansion as an extension of the Finnis-Sinclair and the embedded atom or modified embedded atom potentials. I showed that the nonlinear atomic cluster expansion reaches an RMSE of about 1–3 meV, comparable to that of a neural network potential. I discussed in particular the dependence of the RMSE on the reference data set and illustrated the difficulties that follow for comparing interatomic potentials that are derived from different reference data sets.

ACKNOWLEDGMENTS

I would like to thank J. Jenke for providing random bulk structures, V. Blum for help with the FHI-aims code, and Y. Lysogorskiy for valuable comments on the numerical implementation of the nonlinear atomic cluster expansion. I thank T. Hammerschmidt and M. Mrovec for carefully reading the manuscript.

APPENDIX A: ATOMIC CLUSTER EXPANSION FOR MULTICOMPONENT SYSTEMS

The energy of an atom of species μ_i is given by

$$E_i(\boldsymbol{\sigma}, \boldsymbol{\mu}) = E_i(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \ldots, \boldsymbol{x}_{Ni}), \tag{A1}$$

with $\boldsymbol{x}_{ji} = (\boldsymbol{r}_{ji}, \mu_j)$ and $\boldsymbol{r}_{ji} = \boldsymbol{r}_j - \boldsymbol{r}_i$ and where $\mu = 0, 1, 2, \ldots, M-1$ is a discrete index for the M different chemical species in the system. The collection of the $N-1$ vectors is abbreviated as the configuration $\boldsymbol{\sigma} = (\boldsymbol{r}_{1i}, \boldsymbol{r}_{2i}, \ldots, \boldsymbol{r}_{Ni})$ of atom i and the corresponding vector of chemical species denoted as $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$. The inner product between two functions $f(\boldsymbol{\sigma}, \boldsymbol{\mu})$ and $g(\boldsymbol{\sigma}, \boldsymbol{\mu})$ is then defined as

$$\langle f | g \rangle = \sum_{\boldsymbol{\mu}} \int f^*(\boldsymbol{\sigma}, \boldsymbol{\mu}) g(\boldsymbol{\sigma}, \boldsymbol{\mu}) w(\boldsymbol{\sigma})\, d\boldsymbol{\sigma}, \tag{A2}$$

where the sum over $\boldsymbol{\mu}$ implies summation over all chemical species for all positions. A set of orthogonal and complete basis functions $\phi_{vm}(\boldsymbol{r}, \mu)$ with $v = 0, 1, 2, \ldots$ and $m = 0, 1, 2, \ldots M-1$ that depend only on a single bond $\boldsymbol{r}$ are introduced,

$$\sum_{\mu} \int \phi^*_{vm}(\boldsymbol{x}) \phi_{un}(\boldsymbol{x}) \omega(\boldsymbol{r})\, d\boldsymbol{r} = \delta_{vu}\delta_{mn}, \tag{A3}$$

$$\sum_{vm} \phi^*_{vm}(\boldsymbol{x}_1) \phi_{vm}(\boldsymbol{x}_2) = \frac{\delta(\boldsymbol{r}_1 - \boldsymbol{r}_2)\delta_{\mu_1\mu_2}}{\sqrt{\omega(\boldsymbol{r}_1)\omega(\boldsymbol{r}_2)}}, \tag{A4}$$

with $\boldsymbol{x} = (\boldsymbol{r}, \mu)$. The basis functions for the expansion of the atomic energy Eq. (4) are obtained from the product of $N-1$ single-bond basis functions. By choosing $\phi_{00} = 1$ a hierarchical expansion is obtained. A cluster α with K elements contains K bonds $\alpha = (j_1 i, j_2 i, \ldots, j_K i)$, where the order of entries in α does not matter, and the vector

$\nu = (v_1, m_1; v_2, m_2; \ldots; v_K, m_K)$ with $v_i + m_i > 0$ contains the list of single-bond basis functions of the K bonds in the cluster,

$$\Phi_{\alpha\nu} = \phi_{v_1 m_1}(\boldsymbol{x}_{j_1 i})\phi_{v_2 m_2}(\boldsymbol{x}_{j_2 i}) \ldots \phi_{v_K m_K}(\boldsymbol{x}_{j_K i}), \tag{A5}$$

with $0 \leqslant K \leqslant N-1$. The orthogonality and completeness of the one-bond basis functions transfers to the cluster basis functions,

$$\langle \Phi_{\alpha\nu} | \Phi_{\alpha'\nu'} \rangle = \delta_{\alpha\alpha'}\delta_{\nu\nu'}, \tag{A6}$$

$$1 + \sum_{\gamma \subseteq \alpha} \sum_{\nu} \Phi^*_{\gamma\nu}(\boldsymbol{\sigma}, \boldsymbol{\mu})\Phi_{\gamma\nu}(\boldsymbol{\sigma}', \boldsymbol{\mu}') = \frac{\delta(\boldsymbol{\sigma} - \boldsymbol{\sigma}')\delta_{\boldsymbol{\mu}\boldsymbol{\mu}'}}{\sqrt{w(\boldsymbol{\sigma})w(\boldsymbol{\sigma}')}}, \tag{A7}$$

where α is an arbitrary cluster and the right-hand side of the completeness relation is the product of the relevant right-hand sides of Eq. (A4). A kernel is obtained as

$$k(\boldsymbol{\sigma}, \boldsymbol{\mu}, \boldsymbol{\sigma}', \boldsymbol{\mu}') = 1 + \sum_{\gamma\nu} \Phi^*_{\gamma\nu}(\boldsymbol{\sigma}, \boldsymbol{\mu})\Phi_{\gamma\nu}(\boldsymbol{\sigma}', \boldsymbol{\mu}'), \tag{A8}$$

and an expansion of the atomic energy Eq. (A1) written in the form

$$\begin{aligned} E_i(\boldsymbol{\sigma}, \boldsymbol{\mu}) &= \langle k(\boldsymbol{\sigma}, \boldsymbol{\mu}, \boldsymbol{\sigma}', \boldsymbol{\mu}') | E_i(\boldsymbol{\sigma}', \boldsymbol{\mu}') \rangle \\ &= J_0 + \sum_{\alpha\nu} J_{\alpha\nu}\Phi_{\alpha\nu}(\boldsymbol{\sigma}, \boldsymbol{\mu}), \end{aligned} \tag{A9}$$

with the expansion coefficients $J_{\alpha\nu}$ and where I assume a real-valued expansion. The expansion coefficients are obtained by projection,

$$J_{\alpha\nu} = \langle \Phi_{\alpha\nu} | E_i(\boldsymbol{\sigma}, \boldsymbol{\mu}) \rangle. \tag{A10}$$

APPENDIX B: CLUSTER EXPANSION IN A NONORTHOGONAL BASIS

In practice the atomic cluster expansion Eq. (12) employs nonorthogonal basis sets

$$\int \phi^*_v(\boldsymbol{r})\phi_u(\boldsymbol{r})\omega(\boldsymbol{r})\, d\boldsymbol{r} = S_{vu}, \tag{B1}$$

and $\phi_0 = 1$. It is useful to introduce covariant and contravariant basis functions

$$\phi^v = \sum_u S^{-1}_{vu}\phi_u, \tag{B2}$$

such that the orthogonality and completeness relations Eqs. (6) and (7) become

$$\int [\phi^v(\boldsymbol{r})]^*\phi_u(\boldsymbol{r})\omega(\boldsymbol{r})\, d\boldsymbol{r} = \delta_{vu}, \tag{B3}$$

$$\sum_v [\phi^v(\boldsymbol{r}_1)]^*\phi_v(\boldsymbol{r}_2) = \frac{\delta(\boldsymbol{r}_1 - \boldsymbol{r}_2)}{\sqrt{\omega(\boldsymbol{r}_1)\omega(\boldsymbol{r}_2)}}. \tag{B4}$$

The equivalent relations for the cluster functions follow immediately,

$$\langle \Phi^{\alpha\nu} | \Phi_{\beta\mu} \rangle = \delta_{\alpha\beta}\delta_{\nu\mu}, \tag{B5}$$

$$1 + \sum_{\gamma \subseteq \alpha} \sum_{\nu} [\Phi^{\gamma\nu}(\boldsymbol{\sigma})]^*\Phi_{\gamma\nu}(\boldsymbol{\sigma}') = \frac{\delta(\boldsymbol{\sigma} - \boldsymbol{\sigma}')}{\sqrt{\boldsymbol{\omega}(\boldsymbol{\sigma})\boldsymbol{\omega}(\boldsymbol{\sigma}')}}. \tag{B6}$$

The kernel becomes

$$k(\boldsymbol{\sigma}, \boldsymbol{\sigma}') = 1 + \sum_{\gamma\nu} \Phi^*_{\gamma\nu}(\boldsymbol{\sigma})\Phi^{\gamma\nu}(\boldsymbol{\sigma}'), \tag{B7}$$

and the nonorthogonal expansion may therefore be written as

$$E_i(\boldsymbol{\sigma}) = J_0 + \sum_{\alpha} \sum_{\nu} J^{\alpha\nu}\Phi_{\alpha\nu}(\boldsymbol{\sigma}), \tag{B8}$$

with

$$J^{\alpha\nu} = \langle \Phi^{\alpha\nu} | E_i(\boldsymbol{\sigma}) \rangle. \tag{B9}$$

I continue with only lower indices for ease of notation with the understanding that a nonorthogonal expansion requires mixed upper and lower indices. This is possible because the expansion coefficients $J_{\alpha\nu}$ are obtained numerically and not by explicitly making use of the projection Eq. (B9).

APPENDIX C: REDUCTION OF PRODUCTS OF SPHERICAL HARMONICS

Here the interest is in products of spherical harmonics that are invariant with respect to rotation, i.e., that belong to the identity representation of SO(3) or O(3), respectively. The reduction of products of spherical harmonics is discussed in detail in Ref. [49] and further in textbooks [50,51]. I give expressions for three-, four-, five-, and six-body contributions to the atomic cluster expansion.

1. Three-body contributions

From Eqs. (40), (42), and (43) in the Supplemental Material [48] it is immediately clear that rotational invariance requires that only products

$$\sum_{m=-l}^{l} (-1)^m Y_l^m(\hat{\boldsymbol{r}}_i) Y_l^{-m}(\hat{\boldsymbol{r}}_j) \tag{C1}$$

may contribute to the atomic cluster expansion.

2. Four-body contributions

Contraction of three spherical harmonics with the Wigner $3j$ symbol leads to basis functions for the four-body cluster that are invariant with respect to rotation and inversion,

$$\sum_{m_1 m_2 m_3} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} Y_{l_1}^{m_1}(\hat{\boldsymbol{r}}_i) Y_{l_2}^{m_2}(\hat{\boldsymbol{r}}_j) Y_{l_3}^{m_3}(\hat{\boldsymbol{r}}_k), \tag{C2}$$

and where the sum is limited to terms $m_1 + m_2 + m_3 = 0$. Further, basis functions are different from zero only if $|l_1 - l_2| \leqslant l_3 \leqslant l_1 + l_2$ and, for invariance with respect to inversion, only if $l_1 + l_2 + l_3$ is an even number [48].

Alternatively one could use the Clebsch-Gordan coefficients in combination with the addition theorem for spherical harmonics Eq. (41) to generate invariant basis functions,

$$\sum_{m_1 m_2 m_3} (-1)^{m_1} C^{m_1 m_2 m_3}_{j_1 j_2 j_3} Y_{j_1}^{-m_1}(\hat{\boldsymbol{r}}_i) Y_{j_2}^{m_2}(\hat{\boldsymbol{r}}_j) Y_{j_3}^{m_3}(\hat{\boldsymbol{r}}_k), \tag{C3}$$

and define

$$\begin{bmatrix} j_1 & j_2 & j_3 \\ -m_1 & m_2 & m_3 \end{bmatrix} = (-1)^{m_1} C^{m_1 m_2 m_3}_{j_1 j_2 j_3}. \tag{C4}$$

Both sets of basis functions span the same space. In Ref. [47] basis functions based on the Clebsch-Gordan coefficients were used for the bispectrum and the smooth overlap of atomic positions kernel; see Sec. III D.

3. Five-body contributions

There are different ways to generate invariant products of four spherical harmonics that are equivalent in the sense that the resulting basis functions span the same space. I chose to generate invariant products by contraction of pairs of spherical harmonics with the Clebsch-Gordan coefficients,

$$\sum_{M=-J}^{J} \sum_{\substack{m_1 m_2 \\ m_3 m_4}} (-1)^M C^{M m_1 m_2}_{J j_1 j_2} C^{-M m_3 m_4}_{J j_3 j_4} \times Y^{m_1}_{j_1}(\hat{\boldsymbol{r}}_i) Y^{m_2}_{j_2}(\hat{\boldsymbol{r}}_j) Y^{m_3}_{j_3}(\hat{\boldsymbol{r}}_k) Y^{m_4}_{j_4}(\hat{\boldsymbol{r}}_l). \tag{C5}$$

The symbol in Eq. (31) is therefore defined as

$$\begin{bmatrix} j_1 & j_2 & j_3 & j_4 \\ m_1 & m_2 & m_3 & m_4 \end{bmatrix} = \sum_{J} \sum_{M=-J}^{J} (-1)^M C^{M m_1 m_2}_{J j_1 j_2} C^{-M m_3 m_4}_{J j_3 j_4}. \tag{C6}$$

Many contributions are zero. The Glebsch-Gordan coefficients require $M = m_1 + m_2$ and $-M = m_3 + m_4$, therefore contributions are nonzero only if $m_1 + m_2 + m_3 + m_4 = 0$. Furthermore J only runs over values that range from the larger value of $|j_1 - j_2|$ and $|j_3 - j_4|$ to the smaller value of $j_1 + j_2$ and $j_3 + j_4$.

4. Six-body contributions

I chose to generate invariant products of five spherical harmonics by contraction of three pairs of spherical harmonics with three Clebsch-Gordan coefficients,

$$\sum_{M_1, M_2} \sum_{\substack{m_1 m_2 \\ m_3 m_4 m_5}} (-1)^{M_2} C^{M_1 m_1 m_2}_{J_1 j_1 j_2} C^{M_2 M_1 m_3}_{J_2 J_1 j_3} C^{-M_2 m_4 m_5}_{J_1 j_4 j_5} \times Y^{m_1}_{j_1}(\hat{\boldsymbol{r}}_i) Y^{m_2}_{j_2}(\hat{\boldsymbol{r}}_j) Y^{m_3}_{j_3}(\hat{\boldsymbol{r}}_k) Y^{m_4}_{j_4}(\hat{\boldsymbol{r}}_l) Y^{m_5}_{j_5}(\hat{\boldsymbol{r}}_n). \tag{C7}$$

The symbol in Eq. (32) is therefore defined as

$$\begin{bmatrix} j_1 & j_2 & j_3 & j_4 & j_5 \\ m_1 & m_2 & m_3 & m_4 & m_5 \end{bmatrix} = \sum_{J_1 J_2} \sum_{M_1 M_2} (-1)^{M_2} C^{M_1 m_1 m_2}_{J_1 j_1 j_2} C^{M_2 M_1 m_3}_{J_2 J_1 j_3} C^{-M_2 m_4 m_5}_{J_2 j_4 j_5}. \tag{C8}$$

Also here many contributions are zero. The Glebsch-Gordan coefficients require $M_1 = m_1 + m_2$, $M_2 = m_1 + m_2 + m_3$, and $-M_2 = m_4 + m_5$, therefore contributions are nonzero only if $m_1 + m_2 + m_3 + m_4 + m_5 = 0$. Furthermore J_1 only ranges over values from $|j_1 - j_2|$ to $j_1 + j_2$, while J_2 only takes values from the larger value of $|J_1 - j_3|$ and $|j_4 - j_5|$ to the smaller value of $J_1 + j_3$ and $j_4 + j_5$.

[1] R. Drautz, M. Fähnle, and J. M. Sanchez, J. Phys.: Condens. Matter **16**, 3843 (2004).
[2] F. H. Stillinger and T. A. Weber, Phys. Rev. B **31**, 5262 (1985).
[3] R. Biswas and D. R. Hamann, Phys. Rev. B **36**, 6434 (1987).
[4] R. Haydock, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz, and D. Turnbull, Vol. 35 (Academic Press, New York, 1980), p. 215.
[5] J. Friedel, in *Electrons*, Physics of Metals, edited by J. M. Ziman (Pergamon, London, 1969), Vol. 1.
[6] F. Ducastelle, J. Phys. (Paris) **31**, 1055 (1970).
[7] R. P. Gupta, Phys. Rev. B **23**, 6265 (1981).
[8] M. W. Finnis and J. E. Sinclair, Philos. Mag. A **50**, 45 (1984).
[9] J. K. Nørskov and N. D. Lang, Phys. Rev. B **21**, 2131 (1980).
[10] M. J. Puska, R. M. Nieminen, and M. Manninen, Phys. Rev. B **24**, 3037 (1981).
[11] M. S. Daw and M. I. Baskes, Phys. Rev. B **29**, 6443 (1984).
[12] J. Tersoff, Phys. Rev. Lett. **56**, 632 (1986).
[13] J. Tersoff, Phys. Rev. B **38**, 9902 (1988).
[14] M. I. Baskes, Phys. Rev. B **46**, 2727 (1992).
[15] E. B. Tadmor and R. E. Miller, *Modeling Materials: Continuum, Atomistic and Multiscale Techniques* (Cambridge University Press, Cambridge, UK, 2011).
[16] D. G. Pettifor and I. I. Oleinik, Phys. Rev. Lett. **84**, 4124 (2000).
[17] D. G. Pettifor and I. I. Oleinik, Phys. Rev. B **65**, 172103 (2002).
[18] R. Drautz and D. G. Pettifor, Phys. Rev. B **74**, 174117 (2006).
[19] R. Drautz and D. G. Pettifor, Phys. Rev. B **84**, 214114 (2011).
[20] M. W. Finnis, *Interatomic Forces in Condensed Matter* (Oxford University Press, Oxford, 2003).
[21] R. Drautz, T. Hammerschmidt, M. Cak, and D. G. Pettifor, Modelling Simul. Mater. Sci. Eng. **23**, 074004 (2015).
[22] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[23] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[24] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nat. Mater. **12**, 191 (2013).
[25] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[26] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[27] S. Manzhos and T. Carrington Jr., J. Chem. Phys. **125**, 084109 (2006).
[28] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett **108**, 058301 (2012).
[29] Z. Li, J. R. Kermode, and A. DeVita, Phys. Rev. Lett. **114**, 096405 (2015).
[30] J. Behler, Int. J. Quantum Chem. **115**, 1032 (2015).
[31] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, J. Comput. Phys. **285**, 316 (2015).
[32] M. A. Wood and A. P. Thompson, J. Chem. Phys. **148**, 241721 (2018).
[33] N. Artrith, A. Urban, and G. Ceder, Phys. Rev. B **96**, 014112 (2017).
[34] A. Takahashi, A. Seko, and I. Tanaka, Phys. Rev. Mater. **1**, 063801 (2017).
[35] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, npj Comput. Mater. **3**, 37 (2017).
[36] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017).

[37] T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, J. Chem. Phys. **148**, 241725 (2018).

[38] D. Dragoni, T. D. Daff, G. Csányi, and N. Marzari, Phys. Rev. Mater. **2**, 013808 (2018).

[39] T. Bereau, R. A. DiStasio Jr., A. Tkatchenko, and O. A. von Lilienfeld, J. Chem. Phys. **148**, 241706 (2018).

[40] A. Kamath, R. A. Vargas-Hernández, R. V. Krems, T. Carrington Jr., and S. Manzhos, J. Chem. Phys. **148**, 241702 (2018).

[41] M. Rupp, O. A. von Lilienfeld, and K. Burke, J. Chem. Phys. **148**, 241401 (2018).

[42] F. C. Mocanu, K. Konstantinou, T. H. Lee, N. Bernstein, V. L. Deringer, G. Csányi, and S. R. Elliott, J. Phys. Chem. B **122**, 8998 (2018).

[43] R. Drautz and M. Fähnle, Phys. Rev. B **69**, 104404 (2004).

[44] J. M. Sanchez, F. Ducastelle, and D. Gratias, Physica A **128**, 334 (1984).

[45] R. Singer, F. Dietermann, and M. Fähnle, Phys. Rev. Lett. **107**, 017204 (2011).

[46] J. M. Sanchez, Phys. Rev. B **81**, 224202 (2010).

[47] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).

[48] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevB.99.014104 for details on polynomials and angular functions for the atomic cluster expansion.

[49] R. Singer and M. Fähnle, J. Math. Phys. **47**, 113503 (2006).

[50] A. P. Yutsis, I. B. Levinson, and V. V. Vanagas, *The Theory of Angular Momentum* (Israel Program for Scientific Translations, Jerusalem, 1962).

[51] D. M. Brink and G. R. Satchler, *Angular Momentum* (Clarendon, Oxford, 1968).

[52] R. Drautz and M. Fähnle, Phys. Rev. B **72**, 212405 (2005).

[53] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).

[54] N. Artrith and J. Behler, Phys. Rev. B **85**, 045439 (2012).

[55] R. Kondor, arXiv:cs/0701127.

[56] B. Seiser, D. G. Pettifor, and R. Drautz, Phys. Rev. B **87**, 094105 (2013).

[57] R. N. Silver, H. Röder, A. F. Voter, and J. D. Kress, J. Comput. Phys. **124**, 115 (1996).

[58] A. F. Voter, J. D. Kress, and R. N. Silver, Phys. Rev. B **53**, 12733 (1996).

[59] A. Weiße, G. Wellein, A. Alvermann, and H. Fehske, Rev. Mod. Phys. **78**, 275 (2006).

[60] S. Goedecker and L. Colombo, Phys. Rev. Lett. **73**, 122 (1994).

[61] S. Goedecker and M. Teter, Phys. Rev. B **51**, 9455 (1995).

[62] F. Cyrot-Lackmann, Adv. Phys. **16**, 393 (1967).

[63] J. Jenke, A. P. A. Subramanyam, M. Densow, T. Hammerschmidt, D. G. Pettifor, and R. Drautz, Phys. Rev. B **98**, 144102 (2018).

[64] T. Hammerschmidt, A. N. Ladines, J. Koßmann, and R. Drautz, Crystals **6**, 18 (2016).

[65] A. V. Shapeev, Multiscale Model. Simul. **14**, 1153 (2016).

[66] K. Levenberg, Quart. Appl. Math **2**, 164 (1944).

[67] D. W. Marquardt, J. Soc. Indust. Appl. Math. **11**, 431 (1963).

[68] J. More, B. Garbow, and K. Hillstrom, User Guide for MINPACK-1, Technical Report ANL-80-74, Argonne National Laboratory, 1980.

[69] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comput. Phys. Commun. **180**, 2175 (2009).

[70] V. Havu, V. Blum, P. Havu, and M. Scheffler, J. Comput. Phys. **228**, 8367 (2009).

[71] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[72] J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13244 (1992).

[73] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[74] A. P. Sutton, M. W. Finnis, D. G. Pettifor, and Y. Ohta, J. Phys. C **21**, 35 (1988).

[75] V. Heine, I. J. Robertson, and M. C. Payne, Philos. Trans. R. Soc. London, Ser. A **334**, 393 (1991).

[76] J. C. Slater and G. F. Koster, Phys. Rev. **94**, 1498 (1954).

[77] Y. Lysogorskiy, T. Hammerschmidt, J. Janssen, J. Neugebauer, and R. Drautz (unpublished).