On representing chemical environments

Albert P. Bartók,[1,*] Risi Kondor,[2] and Gábor Csányi[1]
[1]*Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, United Kingdom*
[2]*Department of Computer Science, University of Chicago, 1100 East 58th Street, Chicago, Illinois 60637, USA*
(Received 12 December 2012; published 28 May 2013; corrected 31 May 2013)

We review some recently published methods to represent atomic neighborhood environments, and analyze their relative merits in terms of their faithfulness and suitability for fitting potential energy surfaces. The crucial properties that such representations (sometimes called *descriptors*) must have are differentiability with respect to moving the atoms and invariance to the basic symmetries of physics: rotation, reflection, translation, and permutation of atoms of the same species. We demonstrate that certain widely used descriptors that initially look quite different are specific cases of a general approach, in which a finite set of basis functions with increasing angular wave numbers are used to expand the atomic neighborhood density function. Using the example system of small clusters, we quantitatively show that this expansion needs to be carried to higher and higher wave numbers as the number of neighbors increases in order to obtain a faithful representation, and that variants of the descriptors converge at very different rates. We also propose an altogether different approach, called Smooth Overlap of Atomic Positions, that sidesteps these difficulties by directly defining the similarity between any two neighborhood environments, and show that it is still closely connected to the invariant descriptors. We test the performance of the various representations by fitting models to the potential energy surface of small silicon clusters and the bulk crystal.

DOI: 10.1103/PhysRevB.87.184115 PACS number(s): 07.05.Tp, 36.40.Mr

I. INTRODUCTION

The appropriate representation of atomic environments is a crucial ingredient of algorithms used in modern computational chemistry and condensed matter physics. For example, in structure search applications,[1] each configuration depends numerically on the precise initial conditions and the path of the search, so it is important to be able to identify equivalent structures and detect similarities. In molecular dynamics simulations of phase transitions,[2] one needs good order parameters that are capable of detecting changes in the local order around atoms. Typically, the representation is in terms of a descriptor (also called a fingerprint), a tuple of real valued functions of the atomic positions, e.g., bond lengths, bond angles, etc. "In silico" drug discovery[3,4] and other areas of chemical informatics also rely on characterizing molecules using descriptors. When constructing interatomic potentials and fitting potential energy surfaces (PES),[5–8] the driving application behind this work, the functional forms depend on components of a carefully chosen representation of atomic neighborhoods.

While specifying the position of each atom in a Cartesian coordinate system provides a simple and unequivocal description of atomic configurations, it is not directly suitable for making comparisons between structures: the list of coordinates is ordered arbitrarily and two structures might be mapped to each other by a rotation, reflection, or translation so that two different lists of atomic coordinates can, in fact, represent the same or very similar structures. A good representation is *invariant* with respect to permutational, rotational, reflectional, and translational symmetries, while retaining the faithfulness of the Cartesian representation. In particular, a system of invariant descriptors $q_1, q_2, \ldots, q_M$ is said to be *complete* if it uniquely determines the atomic environment, up to symmetries. It is said to be *overcomplete* if it contains spurious descriptors in the sense that a proper subset of $\{q_1, q_2, \ldots, q_M\}$ is, by itself, complete. If a representation is *complete*, then there is a one-to-one mapping (i.e., a bijection) between the genuinely different atomic environments and the invariant tuples comprising the representation. An *overcomplete* representation assigns potentially many distinct descriptors to a given atomic structure, but guarantees that genuinely different atomic structures will never have identical descriptors associated with them: the function relating representations to atomic structures is a surjection.

Fitting potential energy surfaces (PESs) and electrostatic multipole moment surfaces of small molecules to data generated by first-principles electronic-structure calculations has been a mainstay of computational chemistry for decades.[7,9–27] Typically, when modeling the PES of a small group of atoms, the list of pairwise distances is used or, equivalently, some transformed version of the interatomic distances, e.g. reciprocal[6] or exponential.[7] This description works when the number of atoms is fixed. Even in this case, a seemingly new configuration is obtained by just permuting the order of atoms, i.e., crucial symmetries may be missing in this framework. Braams and Bowman[7] remedied this last shortcoming by using polynomials of pairwise distances, constructed such that each term is invariant to the permutation of identical atoms. Computer code is available that generates the permutationally invariant polynomials automatically[7] (up to 10 atoms), but this approach still does not allow for varying number of atoms in the database of configurations.

In order to generate interatomic potentials for solids or large clusters, capable of describing a wide variety of conditions, the number of neighbors that contribute to the energetics of an atom has to be allowed to vary, with the symmetry-invariant descriptors remaining continuous and differentiable. Even though it is possible to allow the dimensionality M to change with the number of neighbors, for the purpose of function fitting it is more practical that M remains the same. None of the traditional representations fulfill this criterion. Recently, however, a number of new, promising descriptors have been

proposed together with potential energy surfaces based on them.[5,8,24,28,29] Behler and Parrinello's "symmetry functions"[5] were used to generate potentials for silicon,[30] sodium,[31] zinc oxide,[32] and water[33] amongst others; Bartok *et al.* employed the bispectrum[8] to fit a many-body potential for crystalline phases and defects in diamond. Sanville *et al.* used a subset of internal coordinates to fit silicon potentials.[28] Rupp *et al.* used the ordered eigenvalues of the Coulomb matrix to fit atomization energies[29] of a set of over 7000 small organic molecules. At this point, it is not clear which method of representing atomic neighborhoods will prove to be optimal in the long term. We attempt to disentangle this issue from the rather complex details of generating first-principles data and fitting PES, and separately consider the problem of constructing good descriptors.

The most well-known invariants describing atomic neighborhoods are the bond-order parameters originally proposed by Steinhardt *et al.*[34] Here, we show that the bond-order parameters form a subset of a more general set of invariants called the *bispectrum*.[35] The formally infinite array of bispectrum components provides an overcomplete system of invariants, and by truncating it one obtains representations whose sensitivity can be refined at will. We relate the bispectrum to the representation proposed by Behler *et al.*[5,36] and show that, together with another descriptor set described below, their angular parts are all simple polynomials of the same set of canonical invariants.

The paper is organized as follows. In Sec. II, we briefly recall how potential energy surfaces are constructed using invariant descriptors. In Sec. III, we describe a number of descriptors, starting with a simple *distance metric* between atomic configurations which will be used as a reference to assess the faithfulness of all other descriptors but which itself is not differentiable. In Sec. IV, we introduce a *continuous and differentiable* distance metric for constructing potential energy surfaces, called Smooth Overlap of Atomic Positions (SOAP), which has superior properties. In Sec. V, we show numerical tests that help assess the degree of completeness and faithfulness of various descriptors and SOAP, and also show and explore their performance in fitting models for small silicon clusters and the bulk crystal.

II. POTENTIAL ENERGY SURFACE FITTING

The main motivation behind this paper is to define and assess a family of invariant descriptors to be used for fitting interatomic potentials, or potential energy surfaces (PESs). In the construction of potentials for materials applications, the short-range part of the total energy is decomposed into atomic contributions

$$E_{\text{short}} = \sum_n \varepsilon\left(q_1^{(n)}, \ldots, q_M^{(n)}\right),$$

where ε is the contribution of the nth atom, and $\mathbf{q}^{(n)} = (q_1^{(n)}, \ldots, q_M^{(n)})$ is a system of descriptors characterizing the local atomic environment.

Traditionally, such atomic energy functions are defined in closed form. However, recently, there has been a lot of interest in using more flexible, *nonparametric* PESs, derived from computing the total energy and its derivatives at a certain set $\{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(N)}\}$ of "training" configurations using first-principles calculations. A crucial question then is how to fit ε to the computed data points. The simplest approach is to use a linear fit, while Behler and Parrniello advocate using artificial neural networks (NN),[5] and Bartók *et al.* use Gaussian Approximation Potentials (GAP).[8] However, ultimately, each of these procedures result in a PES of the form

$$\varepsilon(\mathbf{q}) = \sum_{k=1}^{N} \alpha_k K(\mathbf{q}, \mathbf{q}^{(k)}), \tag{1}$$

where N is the number of training configurations, which are indexed by k. The coefficient vector $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_N)$ is determined by the fitting procedure, and K is a fixed (nonlinear) function, called the kernel, whose role, intuitively, is to capture the degree of similarity between the atomic environments described by its two arguments. Clearly, then, the choice of descriptors, in particular, their invariance to symmetries, as well as the choice of kernel, are critical ingredients to obtaining good quality PESs.

In general, the kernel K can be interpreted as a covariance function, and therefore it must be symmetric and positive definite [meaning that $K(\mathbf{q}, \mathbf{q}') = K(\mathbf{q}', \mathbf{q})$ and for any nonzero vector $\boldsymbol{\alpha}$ of coefficients, $\sum_k \sum_\ell \alpha_k \alpha_\ell K(\mathbf{q}^{(k)}, \mathbf{q}^{(\ell)}) > 0$). Rasmussen and Williams[37] present a number of such kernels, some of the simplest ones being the following.

The dot-product (DP) kernel is defined as

$$K_{\text{DP}}(\mathbf{q}, \mathbf{q}') = \sum_j q_j q_j', \tag{2}$$

which, when substituted into Eq. (1), results in

$$\begin{aligned}\varepsilon(\mathbf{q}) &= \sum_{k=1}^{N} \alpha_k \sum_j q_j q_j^{(k)} = \sum_j q_j \sum_{k=1}^{N} \alpha_k q_j^{(k)} \\ &= \sum_j q_j \beta_j \equiv \mathbf{q} \cdot \boldsymbol{\beta},\end{aligned}$$

i.e., the linear regression on the descriptor elements with coefficient vector $\boldsymbol{\beta}$.

When using artificial neural networks with N_H hidden units, the atomic energy function ε is given by

$$\varepsilon(\mathbf{q}) = b + \sum_{j=1}^{N_H} v_j h(\mathbf{q}, \mathbf{u}_j), \tag{3}$$

where b is the bias, $\mathbf{v}$ the vector of unit weights, h is the transfer function, and $\{\mathbf{u}_j\}_{j=1}^{N_H}$ the unit parameters.[37] In the limit of an infinite number of hidden units, for specific transfer functions it is possible to reformulate Eq. (3) in the form of Eq. (1) with well-defined covariance functions.[37–39] For example, for $h(\mathbf{q}, \mathbf{u}) = \tanh(u_0 + \sum u_j q_j)$, the corresponding kernel is[39]

$$K_{\text{NN}}(\mathbf{q}, \mathbf{q}') \sim -|\mathbf{q} - \mathbf{q}'|^2 + \text{const.}$$

Finally, the squared exponential (SE) kernel that we have used in the past with the GAP[8] and in some of the examples below is

$$K_{\text{SE}}(\mathbf{q}, \mathbf{q}') = \exp\left(-\sum_j \frac{(q - q')^2}{2\sigma_j^2}\right). \tag{4}$$

III. DESCRIPTORS

Among the applications mentioned in the Introduction, some require representing the geometry of an entire molecule, while for others one needs to describe the neighborhood of an atom perhaps within a finite cutoff distance. While these two cases are closely related, the descriptors for one are not directly suitable for the other, although often the same idea can be used to derive representations for either case. Following, we focus on representing the neighbor environment of a single atom, but for some cases we briefly mention easy generalizations that yield global molecular descriptors.

For N neighboring atomic position vectors $\{\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N\}$ taken relative to a central atom, the symmetric matrix

$$\Sigma = \begin{bmatrix} \mathbf{r}_1\cdot\mathbf{r}_1 & \mathbf{r}_1\cdot\mathbf{r}_2 & \ldots & \mathbf{r}_1\cdot\mathbf{r}_N \\ \mathbf{r}_2\cdot\mathbf{r}_1 & \mathbf{r}_2\cdot\mathbf{r}_2 & \ldots & \mathbf{r}_2\cdot\mathbf{r}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{r}_N\cdot\mathbf{r}_1 & \mathbf{r}_N\cdot\mathbf{r}_2 & \ldots & \mathbf{r}_N\cdot\mathbf{r}_N \end{bmatrix} \tag{5}$$

is, according to Weyl,[40] an overcomplete array of basic invariants with respect to rotation, reflection, and translation. However, Σ is not a suitable descriptor because permutations of atoms change the order of rows and columns. For example, swapping atoms 1 and 2 results in the transformed matrix

$$\begin{bmatrix} \mathbf{r}_2\cdot\mathbf{r}_2 & \mathbf{r}_2\cdot\mathbf{r}_1 & \ldots & \mathbf{r}_2\cdot\mathbf{r}_N \\ \mathbf{r}_1\cdot\mathbf{r}_2 & \mathbf{r}_1\cdot\mathbf{r}_1 & \ldots & \mathbf{r}_1\cdot\mathbf{r}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{r}_N\cdot\mathbf{r}_2 & \mathbf{r}_N\cdot\mathbf{r}_1 & \ldots & \mathbf{r}_N\cdot\mathbf{r}_N \end{bmatrix}. \tag{6}$$

To compare two structures using their Weyl matrices Σ and Σ', we define a *reference* distance metric

$$d_{\text{ref}} = \min_{\mathbf{P}} \|\Sigma - \mathbf{P}\Sigma'\mathbf{P}^T\|, \tag{7}$$

where $\mathbf{P}$ is a permutation matrix and the minimization is over all possible permutations. This metric is not differentiable at locations where the permutation that minimizes (7) changes. It would also be intractable to calculate exactly for large numbers of atoms, but nevertheless we will use this metric to assess the faithfulness of other representations for a small system. Other, differentiable invariants shown later in this paper are, however, closely related to the elements of Σ.

One way to generate permutationally invariant differentiable functions of the Weyl matrix is to compute its eigenvalues; indeed, a very similar descriptor was recently used to fit the atomization energies of a large set of molecules.[29] However, the list of eigenvalues is very far from being complete since there are only N eigenvalues, whereas the dimensionality of the configuration space of N neighbor atoms is $3N-3$ (after the rotational symmetries are removed). It is also unclear how to make the descriptors based on the eigenspectrum continuous and differentiable as the number of neighbors varies.

The Z matrix, or internal coordinates, is a customary set of rotationally invariant descriptors usually used to describe the geometry of entire molecules, but it is not invariant to permutations of atoms. The Z matrix is complete, but in contrast to the Weyl matrix of basic invariants which are based solely on bond lengths and bond angles, it is a minimal set of descriptors that also contains some dihedral angles. It is typically used to construct force fields, in which the total energy is written as a sum over bond, angle, and dihedral terms. Such an expression for the energy is of course permutationally invariant due to the summation over all bonds, angles, and dihedral terms, and all of these terms can indeed be made continuous and differentiable as atoms enter or leave each other's neighborhood. However, we are not considering the Z matrix here because such an ansatz for the atomic energy is itself not general: it manifestly omits all terms that involve more than four atoms. While this has traditionally been accepted for biological systems and organic molecules, the comparative success of embedded atom models for metals and semiconductors shows that fully many-body terms (involving the positions of up to 20 neighbors or more) is likely to be essential for an accurate description of these materials.

Another straightforward way to compare structures is based on pairing the atoms from each and finding the optimal rotation that brings the two structures into as close an alignment as possible. For each pair of structures $\{\mathbf{r}_i\}_{i=1}^N$ and $\{\mathbf{r}'_i\}_{i=1}^N$, it is possible to order the atoms according to their distance from the central atom (or center of mass, in case we want to compare entire molecules) and compute

$$\Delta(\hat{R}) = \sum_j^N |\mathbf{r}_i - \hat{R}\mathbf{r}'_i|^2,$$

where $\hat{R}$ represents an arbitrary rotation (including the possibility of a reflection). We can then define the distance between two configurations as

$$\Delta = \min_{\hat{R}} \Delta(\hat{R}). \tag{8}$$

This distance clearly has all the necessary invariances and completeness properties, but, like d_{ref}, it is not suitable for parametrizing potential energy surfaces because it is again not differentiable: the reordering procedure and the minimization over rotations and reflections introduce cusps.

In the field of molecular informatics, one popular descriptor is based on the histogram of pairwise atomic distances,[41] similar to Valle's crystal fingerprint.[42] We will not consider these here because they are unsuitable for fitting PES, as they are clearly not complete: e.g., from six unordered distance values it is not necessarily possible to construct a unique tetrahedron, even though the number of degrees of freedom is also six. Figure 1 shows two tetrahedra that were constructed such that the edges in each correspond to the same set of six distances. The tetrahedra are manifestly different, which can also be seen by comparing the lists of angles, shown in Table I.

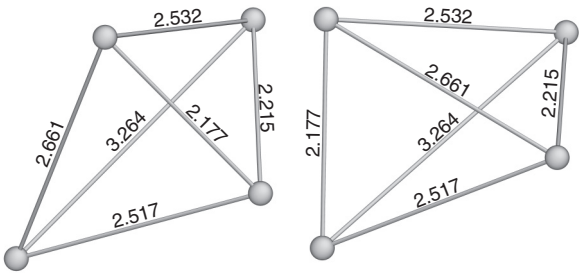

FIG. 1. Two distinct tetrahedra, constructed from the same set of six distances.

TABLE I. Angles of the tetrahedra shown in Fig. 1.

42.66°	49.32°	49.63°	50.36°	52.84°	54.10°	55.50°	61.74°	68.63°	70.40°	77.84°	86.98°
41.78°	42.66°	49.63°	50.36°	50.42°	50.80°	61.74°	61.77°	67.81°	68.63°	86.98°	87.42°

A. Bond-order parameters

As a first step in deriving continuous invariant representations of atomic environments, we define the atomic neighbor density function associated with a given atom as

$$\rho(\mathbf{r}) = \sum_i w_{Z_i} \delta(\mathbf{r} - \mathbf{r}_i), \tag{9}$$

where the index i runs over the neighbors of the atom within some cutoff distance, w_{Z_i} is a unique weight factor assigned according to the atomic species of i, and $\mathbf{r}_i$ is the vector from the central atom to neighbor i. For clarity, we will omit the species weights from now on, and assume a single atomic species, but none of our results rely on this. Determining which neighbors to include in the summation can be done by using a simple binary valued or a smooth real valued cutoff function of the interatomic distance, or via a more sophisticated procedure, e.g., Voronoi analysis.[34] The atomic neighbor density is already invariant to permuting neighbors because changing the order of the atoms in the neighbor list only affects the order of the summation. To simplify the following derivation, for now we omit the information on the radial distance to the neighbors, but will show later how the radial information can be included. The atomic neighbor density function can then be expanded in terms of spherical harmonics:

$$\rho(\hat{\mathbf{r}}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{lm} Y_{lm}(\hat{\mathbf{r}}), \tag{10}$$

where $\hat{\mathbf{r}}$ is the point on the unit sphere corresponding to the direction of the vector $\mathbf{r}$, thus $\rho(\hat{\mathbf{r}})$ is the projection of $\rho(\mathbf{r})$ onto the unit sphere S^2.

The properties of functions defined on the unit sphere are related to the group theory of SO(3), the group of three-dimensional rotations. Spherical harmonics form an orthonormal basis set for $L_2(S^2)$, the class of square integrable functions on the sphere:

$$\langle Y_{lm} | Y_{l'm'} \rangle = \delta_{ll'} \delta_{mm'},$$

where the inner product of functions f and g is defined as

$$\langle f | g \rangle = \int f^*(\hat{\mathbf{r}}) g(\hat{\mathbf{r}})\, d\Omega(\hat{\mathbf{r}}),$$

where the surface element $d\Omega(\hat{\mathbf{r}})$ can be expressed in terms of the polar angles θ and ϕ as

$$d\Omega(\hat{\mathbf{r}}) = \sin\theta\, d\theta\, d\phi,$$

and the coefficients c_{lm} are given by

$$c_{lm} = \langle \rho | Y_{lm} \rangle = \sum_i Y_{lm}(\hat{\mathbf{r}}_i). \tag{11}$$

The quantities Q_{lm} introduced by Steinhardt *et al.*[34] are proportional to the coefficients c_{lm}. Dividing by N, the number of neighbors of the atom (within a finite cutoff distance) provides the atomic order parameters

$$Q_{lm} = \frac{1}{N} \sum_i Y_{lm}(\hat{\mathbf{r}}_i). \tag{12}$$

Furthermore, averaging Eq. (12) over atoms in the entire system gives a set of global order parameters

$$\bar{Q}_{lm} = \frac{1}{N_b} \sum_{ii'} Y_{lm}(\hat{\mathbf{r}}_{ii'}),$$

where N_b is the total number of atom pairs included in the summation, and we wrote $\mathbf{r}_{ii'}$ for the vector connecting atom i to its neighbor i'. Both sets are invariant to permutations of atoms and translations, but still depend on the orientation of the reference frame. However, rotationally invariant combinations can be constructed as

$$Q_l = \left[\frac{4\pi}{2l+1} \sum_{m=-l}^{l} (Q_{lm})^* Q_{lm} \right]^{1/2} \quad \text{and} \tag{13}$$

$$W_l = \sum_{m_1,m_2,m_3=-l}^{l} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} Q_{lm_1} Q_{lm_2} Q_{lm_2} \tag{14}$$

for atomic neighborhoods and

$$\bar{Q}_l = \left[\frac{4\pi}{2l+1} \sum_{m=-l}^{l} \bar{Q}^*_{lm} \bar{Q}_{lm} \right]^{1/2},$$

$$\bar{W}_l = \sum_{m_1,m_2,m_3=-l}^{l} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \bar{Q}_{lm_1} \bar{Q}_{lm_2} \bar{Q}_{lm_2}$$

for the entire structure. The factor in parentheses is the Wigner $3jm$ symbol,[43] which is zero unless $m_1 + m_2 + m_3 = 0$.

The numbers Q_l and W_l are called second-order and third-order bond-order parameters, respectively. It is possible to normalize W_l such that it does not depend strongly on the number of neighbors:

$$\widehat{W}_l = \left[\sum_{m=-l}^{l} (Q_{lm})^* Q_{lm} \right]^{-3/2} W_l.$$

For symmetry reasons, only coefficients with $l \geqslant 4$ have nonzero values in environments with cubic symmetry and $l \geqslant 6$ for environments with icosahedral symmetry. Different values correspond to crystalline materials with different symmetry, while the global order parameters vanish in disordered phases, such as liquids. Bond-order parameters were originally introduced for studying order in liquids and glasses,[34] but were soon adopted for a wide range of applications. They have been used to study the free energy of clusters,[44,45] melting of quantum solids,[46] nucleation,[47] as well as to serve as reaction coordinates in simulations of phase transitions[48,49] and glasses[50] and also to generate interatomic potentials.[51]

B. Power spectrum

Using some basic concepts from representation theory, we now prove that the second-order bond-order parameters are indeed rotationally invariant, then we show a more general set of third-order invariants,[35] of which the Qs and the Ws are a subset. An arbitrary rotation $\hat{R}$ operating on a spherical harmonic function Y_{lm} transforms it into a linear combination of spherical harmonics with the same l index:

$$\hat{R}\, Y_{lm} = \sum_{m'=-l}^{l} D^l_{mm'}(\hat{R}) Y_{lm'},$$

where the $\mathbf{D}^l(\hat{R})$ matrices are known as the Wigner matrices, which form the irreducible representations of the three-dimensional rotation group SO(3). The elements of the Wigner matrices are given by

$$D^l_{mm'}(\hat{R}) = \langle Y_{lm} | \hat{R} | Y_{lm'} \rangle. \tag{15}$$

It follows that the rotation operator $\hat{R}$ acts on the function ρ as

$$\begin{aligned}\hat{R}\rho &= \hat{R} \sum_{l=0} \sum_{m=-l}^{l} c_{lm} Y_{lm} = \sum_{l=0} \sum_{m=-l}^{l} c_{lm} \hat{R} Y_{lm} \\ &= \sum_{l=0} \sum_{m=-l}^{l} \sum_{m'=-l}^{l} c_{lm} D^l_{mm'}(R) Y_{lm'} \\ &\equiv \sum_{l=0} \sum_{m'=-l}^{l} c'_{lm} Y_{lm'},\end{aligned}$$

thus the column vector $\mathbf{c}_l$ of expansion coefficients transforms under rotation as

$$\mathbf{c}_l \to \mathbf{D}^l(\hat{R}) \mathbf{c}_l.$$

Making use of the fact that rotations are unitary operations on $L_2(S^2)$, it is possible to show that the matrices $\mathbf{D}^l$ are unitary,

$$\mathbf{D}^{l\dagger} \mathbf{D}^l = \mathbf{I},$$

and therefore $\mathbf{c}_l^\dagger \mathbf{c}_l$ transforms according to

$$p_l \equiv \mathbf{c}_l^\dagger \mathbf{c}_l \to (\mathbf{c}_l^\dagger \mathbf{D}^{l\dagger})(\mathbf{D}^l \mathbf{c}_l) = \mathbf{c}_l^\dagger \mathbf{c}_l, \tag{16}$$

i.e., is invariant under rotation. We call p_l the rotational power spectrum due to the analogy with the familiar power spectrum of ordinary Fourier analysis.

We also note that the elements of $\mathbf{c}_l$ transform under reflection about the origin as

$$\mathbf{c}_l \to (-1)^l \mathbf{c}_l, \tag{17}$$

thus the power spectrum is also invariant to this symmetry operation. A comparison with Eqs. (11), (12) and (13) shows that the second-order bond-order parameters are related to the power spectrum via the simple scaling

$$Q_l = \left(\frac{4\pi}{2l+1} p_l \right)^{1/2}.$$

The power spectrum is clearly not a complete descriptor for a general function $f(\hat{\mathbf{r}})$ on the sphere, for example, consider the two different functions

$$f_1 = Y_{22} + Y_{2-2} + Y_{33} + Y_{3-3}$$

and

$$f_2 = Y_{21} + Y_{2-1} + Y_{32} + Y_{3-2},$$

which both have the same power spectrum $p_2 = 2$ and $p_3 = 2$ (with all other components equal to zero). However, for the restricted class of functions which are sums of a limited number of delta functions [such as the atomic neighbor density ρ in Eq. (9)], the power spectrum elements turn out to be polynomials of the basic invariants of Weyl. Using numerical experiments we demonstrate in Sec. V that for a fixed number of neighbors a certain set of power spectrum components is likely to be overcomplete.

C. Bispectrum

We generalize the concept of the power spectrum to obtain a larger set of invariants via coupling different angular momentum channels.[35,52] Let us consider the direct product $\mathbf{c}_{l_1} \otimes \mathbf{c}_{l_2}$, which transforms under a rotation as

$$\mathbf{c}_{l_1} \otimes \mathbf{c}_{l_2} \to (\mathbf{D}^{l_1} \otimes \mathbf{D}^{l_2}) \left(\mathbf{c}_{l_1} \otimes \mathbf{c}_{l_2} \right).$$

It follows from the representation theory of compact groups that the direct product of two irreducible representations can be decomposed into a direct sum of irreducible representations.[53] In the case of the SO(3) group, the direct product of two Wigner matrices can be decomposed into a direct sum of Wigner matrices

$$D^{l_1}_{m_1 m'_1} D^{l_2}_{m_2 m'_2} = \sum_{l,m,m'} D^l_{mm'} \left(C^{l\, l_1 l_2}_{m m_1 m_2} \right)^* C^{l\, l_1 l_2}_{m' m'_1 m'_2}, \tag{18}$$

where $C^{l\, l_1 l_2}_{m m_1 m_2}$ denote the Clebsch-Gordan coefficients or, using more compact notation,

$$\mathbf{D}^{l_l} \otimes \mathbf{D}^{l_2} = \left(\mathbf{C}^{l_1 l_2} \right)^\dagger \left[\bigoplus_{l=|l_1-l_2|}^{l_1+l_2} \mathbf{D}^l \right] \mathbf{C}^{l_1 l_2}, \tag{19}$$

where $\mathbf{C}^{l_1 l_2}$ denote the matrices formed of the Clebsch-Gordan coefficients. These are themselves unitary, so the vector $\mathbf{C}^{l_1 l_2}(\mathbf{c}_{l_1} \otimes \mathbf{c}_{l_2})$ transforms as

$$\mathbf{C}^{l_1 l_2} \left(\mathbf{c}_{l_1} \otimes \mathbf{c}_{l_2} \right) \to \left[\bigoplus_{l=|l_1-l_2|}^{l_1+l_2} \mathbf{D}^l \right] \mathbf{C}^{l_1 l_2} \left(\mathbf{c}_{l_1} \otimes \mathbf{c}_{l_2} \right). \tag{20}$$

Writing out the block-diagonal matrix in the square brackets as

$$\left[\bigoplus_{l=|l_1-l_2|}^{l_1+l_2} \mathbf{D}^l \right] \equiv \begin{pmatrix} \boxed{\mathbf{D}^{|l_1-l_2|}} & & & \\ & \boxed{\mathbf{D}^{|l_1-l_2|+1}} & & \\ & & \ddots & \\ & & & \boxed{\mathbf{D}^{l_1+l_2}} \end{pmatrix},$$

we see that each block selects a particular slice of the vector in Eq. (20), which transforms according to a given $\mathbf{D}^l$ matrix. We give a new symbol to these slices, $\mathbf{g}_{l\,l_1l_2}$, so that the original vector is their direct sum

$$\mathbf{C}^{l_1l_2}\left(\mathbf{c}_{l_1}\otimes\mathbf{c}_{l_2}\right)\equiv\bigoplus_{l=|l_1-l_2|}^{l_1+l_2}\mathbf{g}_{l\,l_1l_2},$$

and each $\mathbf{g}_{l\,l_1l_2}$ transforms under rotation as

$$\mathbf{g}_{l\,l_1l_2}\rightarrow\mathbf{D}^l\mathbf{g}_{l\,l_1l_2}.$$

Analogously to the power spectrum, we can now define the bispectrum as the collection of scalars

$$b_{l\,l_1l_2}=\mathbf{c}_l^\dagger\mathbf{g}_{l\,l_1l_2},$$

which are invariant to rotations:

$$b_{l\,l_1l_2}=\mathbf{c}_l^\dagger\mathbf{g}_{l\,l_1l_2}\rightarrow\left(\mathbf{c}_l\mathbf{D}^l\right)^\dagger\mathbf{D}^l\mathbf{g}_{l\,l_1l_2}=\mathbf{c}_l^\dagger\mathbf{g}_{l\,l_1l_2}.$$

It follows from Eq. (17) that those elements of the bispectrum where l_1+l_2+l is odd change sign under reflection about the origin. If invariance to reflection is required, we take the absolute value of these components or omit them from the descriptor.

Rewriting the bispectrum formula as

$$b_{l\,l_1l_2}=\sum_{m=-l}^{l}\sum_{m_1=-l_1}^{l_1}\sum_{m_2=-l_2}^{l_2}c_{lm}^*C^{l\,l_1l_2}_{m\,m_1m_2}c_{l_1m_1}c_{l_2m_2},\qquad(21)$$

the similarity to the third-order bond-order parameters becomes apparent. Indeed, the Wigner $3jm$ symbols are related to the Clebsch-Gordan coefficients through

$$\begin{pmatrix}l_1&l_2&l_3\\m_1&m_2&m_3\end{pmatrix}=\frac{(-1)^{l_1-l_2-m_3}}{\sqrt{2l_3+1}}C^{l_1l_2l_3}_{m_1m_2-m_3},\qquad(22)$$

and by substituting the spherical harmonics identity $Y_{lm}=(-1)^mY^*_{l-m}$ in Eq. (12), it follows that

$$Q_{lm}=(-1)^m(Q_{lm})^*.\qquad(23)$$

Substituting the identities (22) and (23) into the definition (14), we obtain

$$W_l=\frac{1}{\sqrt{2l+1}}\times\sum_{m_1,m_2,m_3=-l}^{l}(-1)^{-m}C^{l\,l\,l}_{m_1m_2-m_3}Q_{lm_1}Q_{lm_2}(-1)^m(Q_{lm_3})^*,$$

thus the third-order parameters W_l are seen to be proportional to the diagonal elements of the bispectrum $b_{l\,l\,l}$. Noting that $Y_{00}\equiv1$, the coefficient c_{00} is the number of neighbors N, and using $C^{l\,0\,l_2}_{m\,0\,m_2}=\delta_{l\,l_2}\delta_{m\,m_2}$, the bispectrum elements $l_1=0, l=l_2$ are identical to the previously introduced power spectrum components:

$$b_{l\,0\,l}=N\sum_{m=-l}^{l}\sum_{m_2=-l}^{l}c_{lm}^*\delta_{m\,m_2}c_{l\,m_2}=N\sum_{m=-l}^{l}c_{lm}^*c_{lm}=Np_l,$$

therefore,

$$Q_l\propto\sqrt{p_l}\propto\sqrt{b_{l\,0\,l}},$$
$$W_l\propto b_{l\,l\,l}.$$

The first few terms of the power spectrum and bispectrum for an atom with three neighbors are shown in the following, where $\theta_{ii'}$ is the angle between the bonds to neighbors i and i', and the sums are over all the neighbors. The fact that dihedral angles do not need to feature here can be seen by considering the Weyl matrix, which is known to be a complete descriptor, and it consists solely of the bond lengths to and angles between the neighbors:

$$p_0=\frac{9}{4\pi},$$
$$p_1=\frac{3}{4\pi}\left(\sum_{ii'}\cos\theta_{ii'}+3\right),$$
$$p_2=\frac{5}{4\pi}\left(\frac{3}{2}\sum_{ii'}\cos^2\theta_{ii'}+6\right),$$
$$p_3=\frac{7}{4\pi}\left(\frac{5}{2}\sum_{ii'}\cos^3\theta_{ii'}-\frac{3}{2}\sum_{ii'}\cos\theta_{ii'}+3\right),$$
$$p_4=\frac{9}{16\pi}\left(\frac{35}{2}\sum_{ii'}\cos^4\theta_{ii'}-15\sum_{ii'}\cos^2\theta_{ii'}+13\right),$$
$$b_{211}=\sqrt{\frac{15}{128\pi^3}}\left[\frac{3}{4}\left(\sum_{ii'}\cos\theta_{ii'}\right)^2+\frac{3}{2}\sum_{ii'}\cos^2\theta_{ii'}+5\sum_{ii'}\cos\theta_{ii'}\right],$$
$$b_{321}=\frac{150}{8}\sqrt{\frac{7}{\pi^3}}\left[\frac{5}{2}\sum_{ii'}\cos^3\theta_{ii'}+\frac{5}{4}\sum_{ii'}\cos^2\theta_{ii'}\sum_{ii'}\cos\theta_{ii'}-\frac{1}{2}\left(\sum_{ii'}\cos\theta_{ii'}\right)^2+4\sum_{ii'}\cos^2\theta_{ii'}-2\sum_{ii'}\cos\theta_{ii'}+18\right].$$

D. Radial basis

Thus far, we neglected the distance of neighboring atoms from the central atom by using the unit-sphere projection of the atomic environment. One way to introduce radial information is to complement the spherical harmonics basis in Eq. (10) with radial basis functions g_n (Ref. 54):

$$\rho(\mathbf{r})=\sum_n\sum_{l=0}\sum_{m=-l}^{l}c_{nlm}g_n(r)Y_{lm}(\hat{\mathbf{r}}).\qquad(24)$$

If the set of radial basis functions is not orthonormal, i.e., $\langle g_n|g_m\rangle=S_{nm}\neq\delta_{nm}$, after obtaining the coefficients

c'_{nlm} with

$$c'_{nlm} = \langle g_n Y_{lm} | \rho \rangle,$$

the elements c_{nlm} are given by

$$c_{nlm} = \sum_{n'} (S^{-1})_{n'n} c'_{n'lm}.$$

In practice, when constructing the bispectrum, either c'_{nlm} or c_{nlm} can be used.

Rotational invariance must only apply globally, and not to each radial basis separately, therefore the angular momentum channels corresponding to different radial basis functions need to be coupled. So, although extending Eqs. (16) and (21) simply as

$$p_{nl} = \sum_{m=-l}^{l} c^*_{nlm} c_{nlm},$$

$$b_{nl_1l_2} = \sum_{m=-l}^{l} \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} c^*_{nlm} C^{l l_1 l_2}_{m m_1 m_2} c_{nl_1m_1} c_{nl_2m_2}$$

provides a set of invariants describing the three-dimensional neighborhood of the atom, this can easily lead to a poor representation if the radial basis functions do not sufficiently overlap. The different radial shells will only be weakly coupled, and the representation will have spurious quasi-invariances to rotating subsets of atoms at approximately the same distance, as illustrated in Fig. 2.

To avoid this, it is necessary to choose basis functions that are sensitive over a wide range of distances, although this may reduce the sensitivity of each radial basis function because they are varying very slowly. The fine tuning of the basis set is rather arbitrary, and there is no guarantee that a choice exists that is optimal or even satisfactory for all systems of interest.

We suggest constructing radial functions from cubic and higher-order polynomials $\phi_\alpha(r) = (r_{\rm cut} - r)^{\alpha+2}/N_\alpha$ for

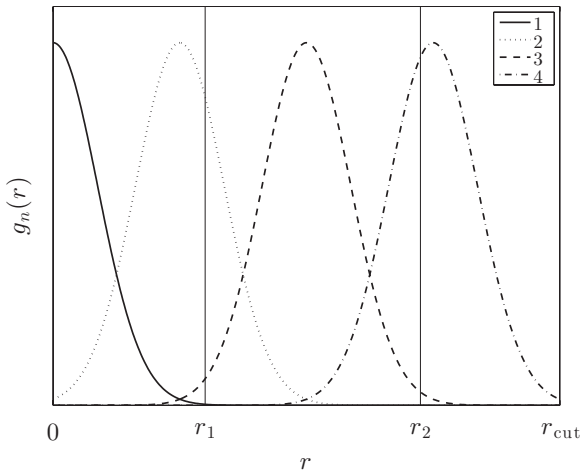

FIG. 2. Example of weakly overlapping radial basis functions $g_n(r)$ [cf. Eq. (24)]. Atoms 1 and 2 at distance r_1 and r_2 from the center become decoupled as their contribution to the power spectrum or bispectrum elements is weighed down by the product $g_n(r_1)g_n(r_2)$, which is rather small for all n.

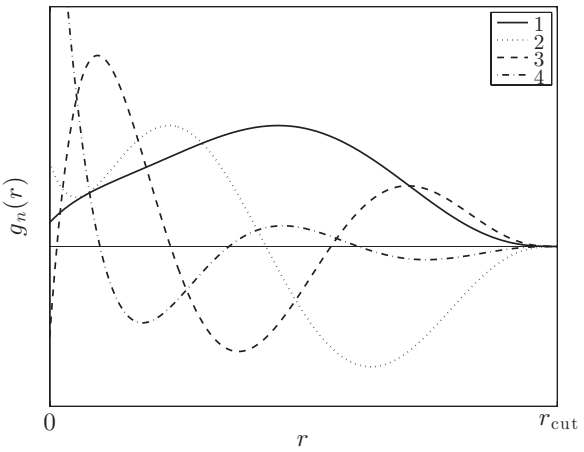

FIG. 3. Example of radial basis functions $g_n(r)$, as defined in Eq. (25) for $n = 1,2,3,4$.

$\alpha = 1,2,\ldots,n_{\rm max}$, normalized on the range $(0,r_{\rm cut})$ using

$$N_\alpha = \sqrt{\int_0^{r_{\rm cut}} (r_{\rm cut} - r)^{2(\alpha+2)} dr} = \sqrt{\frac{r_{\rm cut}^{2\alpha+5}}{2\alpha+5}}.$$

The orthonormalized construction

$$g_n(r) = \sum_{\alpha=1}^{n_{\rm max}} W_{n\alpha} \phi_\alpha(r) \tag{25}$$

guarantees that each radial function returns smoothly to zero at the cutoff with both the first and the second derivatives equal to zero (see Fig. 3). The matrix $\mathbf{W}$ of linear combination coefficients is obtained from the overlap matrix as

$$S_{\alpha\beta} = \int_0^{r_{\rm cut}} \phi_\alpha(r)\phi_\beta(r) dr = \frac{\sqrt{(5+2\alpha)(5+2\beta)}}{5+\alpha+\beta},$$

$$\mathbf{W} = \mathbf{S}^{-1/2}.$$

Another way to avoid radial decoupling is to define the rotational invariants in such a way that they couple different radial channels explicitly, for example, as

$$p_{n_1n_2l} = \sum_{m=-l}^{l} c^*_{n_1lm} c_{n_2lm} \text{ and} \tag{26}$$

$$b_{n_1n_2l\,l_1l_2} = \sum_{m=-l}^{l} \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} c^*_{n_1lm} C^{l l_1 l_2}_{m m_1 m_2} c_{n_2l_1m_1} c_{n_2l_2m_2}.$$

Here, each invariant has contributions from two different radial basis channels, and so we ensure that they can not become decoupled, but at the price of increasing the number of invariants quadratically or even cubically in the number of radial basis functions used.

E. Four-dimensional power spectrum and bispectrum

We now present an alternative to the SO(3) power spectrum and bispectrum that does not need the explicit introduction of a radial basis set, but still represents atomic neighborhoods in three-dimensional space. We start by projecting the atomic neighborhood density within a cutoff radius $r_{\rm cut}$ onto the

surface of the four-dimensional sphere S^3 with radius r_0. The surface of S^3 is defined as the set of points $\mathbf{s} \in \mathbb{R}^4$, where $s_1^2 + s_2^2 + s_3^2 + s_4^2 = r_0^2$, while the polar angles ϕ, θ, and θ_0 of $\mathbf{s}$ are defined so that

$$\begin{aligned} s_1 &= r_0 \cos\theta_0, \\ s_2 &= r_0 \sin\theta_0 \cos\theta, \\ s_3 &= r_0 \sin\theta_0 \sin\theta \cos\phi, \\ s_4 &= r_0 \sin\theta_0 \sin\theta \sin\phi. \end{aligned}$$

We choose to use the projection

$$\mathbf{r} \equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix} \rightarrow \begin{array}{l} \phi = \arctan(y/x) \\ \theta = \arccos(z/|\mathbf{r}|), \\ \theta_0 = \pi |\mathbf{r}|/r_0 \end{array}$$

where $r_0 > r_{\text{cut}}$ is a parameter, thus rotations in three-dimensional space correspond to a subset of rotations in four-dimensional space. This projection is somewhat similar to a Riemann projection, except in that case θ_0 would be defined as

$$\theta_0 = 2 \arctan(|\mathbf{r}|/2r_0),$$

implying

$$\theta_0 \approx |\mathbf{r}|/r_0 \text{ for } |\mathbf{r}| \ll r_0.$$

In contrast to the Riemann projection, our choice of θ_0 allows more sensitive representation of the entire radial range. The limit $r_0 = r_{\text{cut}}$ projects each atom at the cutoff distance to the south pole of the four-dimensional sphere, thus losing all angular information. Too large an r_0 would project all positions onto a small surface area of the sphere around the north pole, requiring a large number of basis functions to represent the atomic environment. In practice, a large range of r_0 values works well, in particular, we used $r_0 = \frac{4}{3} r_{\text{cut}}$.

To illustrate the procedure, Fig. 4 shows the Riemann projections for one and two dimensions, which can be easily drawn.

An arbitrary function ρ defined on the surface of a four-dimensional (4D) sphere can be numerically represented using the hyperspherical harmonic functions $U^j_{m'm}(\phi,\theta,\theta_0)$ (Refs. 43 and 55):

$$\rho = \sum_{j=0}^{\infty} \sum_{m,m'=-j}^{j} c^j_{m'm} U^j_{m'm}, \tag{28}$$

which, in fact, correspond to individual matrix components of the Wigner (i.e., rotational) matrices, as defined in Eq. (15). In this case, the arguments represent a rotation by θ_0 around the vector pointing in the (ϕ,θ) direction, which can be transformed to the conventional Euler angles, and j takes half-integer values.

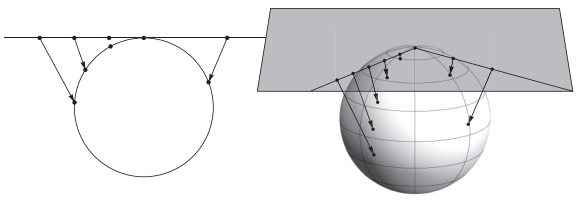

FIG. 4. Two- and three-dimensional Riemann constructions that map a flat space onto the surface of a sphere in one higher dimension.

The hyperspherical harmonics form an orthonormal basis for $L_2(S^3)$, thus the expansion coefficients $c^j_{m'm}$ can be calculated via

$$c^j_{m'm} = \langle U^j_{m'm} | \rho \rangle,$$

where $\langle \ldots | \ldots \rangle$ denotes the inner product defined on the four-dimensional hypersphere

$$\langle f | g \rangle = \int_0^{\pi} d\theta_0 \, \sin^2\theta_0 \int_0^{\pi} d\theta \, \sin\theta \int_0^{2\pi} d\phi \times f^*(\theta_0,\theta,\phi)\, g(\theta_0,\theta,\phi).$$

Although the coefficients $c^j_{m'm}$ have two indices besides j, for each j it is convenient to collect them into a single vector $\mathbf{c}^j$. Similarly to the three-dimensional case, rotations act on the hyperspherical harmonic functions as

$$\hat{R} U^j_{m'_1 m_1} = \sum_{m'_2 m_2} R^j_{m'_1 m_1 m'_2 m_2} U^j_{m'_2 m_2},$$

where the matrix elements $R^j_{m'_1 m_1 m'_2 m_2}$ are given by

$$R^j_{m'_1 m_1 m'_2 m_2} = \langle U^j_{m'_1 m_1} | \hat{R} | U^j_{m'_2 m_2} \rangle.$$

Hence, the rotation $\hat{R}$ acting on ρ transforms the coefficient vectors $\mathbf{c}^j$ according to

$$\mathbf{c}^j \rightarrow \mathbf{R}^j \mathbf{c}^j.$$

The unitary $\mathbf{R}^j$ matrices are the SO(4) analogs of the Wigner matrices $\mathbf{D}^l$ of the SO(3) case above, and it can be shown that the direct product of the four-dimensional rotation matrices decomposes according to

$$\mathbf{R}^{j_1} \otimes \mathbf{R}^{j_2} = \left(\mathbf{H}^{j_1 j_2}\right)^{\dagger} \left[\bigoplus_{j=|j_1-j_2|}^{j_1+j_2} \mathbf{R}^j \right] \mathbf{H}^{j_1 j_2},$$

which is the four-dimensional analog of Eq. (19). The coupling constants $\mathbf{H}^{j_1 j_2}$, or Clebsch-Gordan coefficients of SO(4), are[55,56]

$$H^{jmm'}_{j_1 m_1 m'_1, j_2 m_2 m'_2} \equiv C^{j\, j_1 j_2}_{m m_1 m_2} C^{j\, j_1 j_2}_{m' m'_1 m'_2}.$$

The rest of the derivation continues analogously to the 3D case, and finally we arrive at the expression for the SO(4) bispectrum elements

$$B_{j\, j_1 j_2} = \sum_{m'_1, m_1 = -j_1}^{j_1} c^{j_1}_{m'_1 m_1} \sum_{m'_2, m_2 = -j_2}^{j_2} c^{j_2}_{m'_2 m_2} \times \sum_{m', m = -j}^{j} C^{j\, j_1 j_2}_{m m_1 m_2} C^{j\, j_1 j_2}_{m' m'_1 m'_2} \left(c^j_{m'm}\right)^*,$$

while the SO(4) power spectrum is

$$P_j = \sum_{m', m = -j}^{j} \left(c^j_{m'm}\right)^* c^j_{m'm}.$$

The SO(4) bispectrum is invariant to rotations of four-dimensional space, which include three-dimensional rotations.

TABLE II. Number of components in the full and diagonal bispectrum as a function of the band limit $j_{\max}$.

$j_{\max}$	0	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3	$\frac{7}{2}$	4	$\frac{9}{2}$
$B_{j\,j_1 j_1}$	1	2	5	7	12	15	22	26	35	40
$B_{j\,j_1 j_2}$	1	4	11	23	42	69	106	154	215	290

However, there are additional rotations, associated with the third polar angle θ_0, which, in our case, represents the radial information. In order to eliminate the unphysical invariance with respect to rotations along the third polar angle, we modify the atomic density as

$$\rho(\mathbf{r}) = \delta(\mathbf{0}) + \sum_i \delta(\mathbf{r} - \mathbf{r}_i),$$

i.e., we add the central atom, with the coordinates (0,0,0), as a fixed reference point, anchoring the neighborhood. The resulting invariants $B_{j\,j_1 j_2}$ have only three indices, but contain both radial and angular information, and have the required symmetry properties. There are no adjustable parameters in the definition of these invariants, apart from the projection parameter r_0 discussed above.

The number of components in the truncated representation depends on the band limit $j_{\max}$ in the expansion (28). For symmetry reasons, the bispectrum components with noninteger $j_1 + j_2 + j$ change sign under reflection and, because of this reason, we omitted them. Just as in the 3D case, the representation is probably overcomplete, i.e., most of the bispectrum components are redundant. To reduce the number of redundant elements, we only used the "diagonal" components, i.e., $j_1 = j_2$. Table II shows the number of bispectrum elements for increasing band limit values.

F. Parrinello-Behler descriptor

We include in the tests below the descriptor suggested by Parrinello and Behler[5] using the parameters published recently[57] (and henceforth termed PB) (see Fig. 5). The

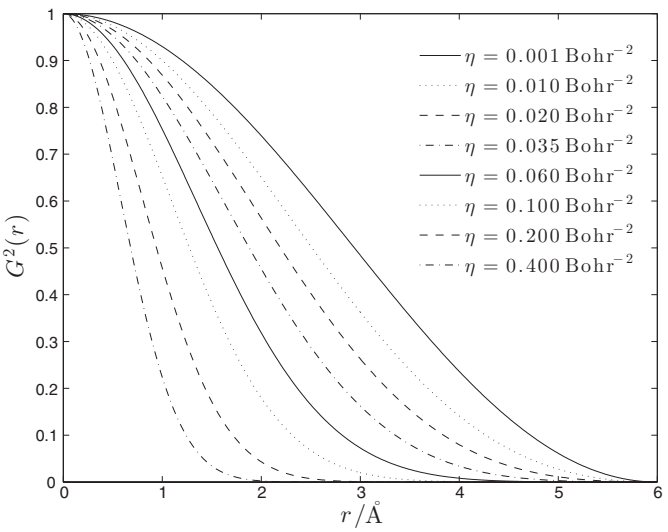

FIG. 5. Radial basis functions G^2 in the Parrinello-Behler (PB) type descriptors (Ref. 57).

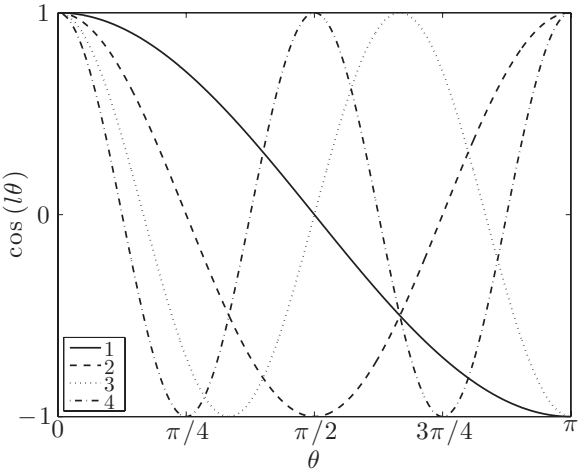

FIG. 6. Examples of the angular basis functions for $l_{\max} = 4$ of the AFS descriptor.

two- and three-body symmetry functions (in their terminology) are

$$G^2_\alpha = \sum_i \exp[-\eta_\alpha (r_i - R_{s\alpha})^2] f_c(r_i)$$

and

$$G^4_\alpha = 2^{1-\zeta_\alpha} \sum_{i,i'} (1 + \lambda_\alpha \cos\theta_{ii'})^{\zeta_\alpha} \times \exp\left[-\eta_\alpha \left(r_i^2 + r_{i'}^2 + r_{ii'}^2\right)\right] f_c(r_i) f_c(r_{i'}) f_c(r_{ii'}),$$

where the cutoff function is defined as

$$f_c(r) = \begin{cases} \left[\cos\left(\frac{\pi r}{r_{\text{cut}}}\right) + 1\right]/2 & \text{for } r \leqslant r_{\text{cut}}, \\ 0 & \text{for } r > r_{\text{cut}}. \end{cases}$$

Different values of the parameters $\eta, R_s, \zeta, \lambda$ can be used to generate an arbitrary number of invariants.

G. Angular Fourier series

Notice that the angular part of the power spectrum, bispectrum (Sec. III C) and the descriptors defined by Parrinello and Behler (Sec. III F) are all simple polynomials of the canonical set $\sum_{ii'} \cos^m \theta_{ii'}$ for integer m, which, in turn, are sums of powers of the basic invariants of Weyl. We include in the tests in the next section a further descriptor set, which we call the angular Fourier series (AFS) descriptor formed by a system of orthogonal polynomials of the basic invariants, conveniently chosen as the Chebyshev polynomials $T_l(x)$, as

$$T_l(\cos\theta) = \cos(l\theta), \quad \text{(see Fig. 6)}$$

and incorporate the radial information using the basis functions defined in Eq. (25), leading to

$$\text{AFS}_{n,l} = \sum_{i,i'} g_n(r_i) g_n(r_{i'}) \cos(l\theta_{ii'}).$$

IV. A SIMILARITY MEASURE BETWEEN ATOMIC ENVIRONMENTS

It is clear from the preceding section that there is a lot of freedom in constructing descriptors, e.g., in the choice of

angular band limit, the radial basis and also which subset of the basis elements are actually used. As we have shown in Sec. II, the key to PES fitting is not the descriptors *per se*, but the similarity measure $K(\mathbf{q},\mathbf{q}')$ that is constructed from the descriptors. This suggests an alternative approach, in which descriptors are bypassed altogether, and a similarity measure between atomic neighborhoods is constructed directly. The criteria for a good similarity measure is not only that it be invariant to symmetry operations of the atoms of each environment and have a well-defined limit when comparing two identical or two very different environments, but also that it change smoothly with the Cartesian atomic coordinates.

We define the similarity of two atomic environments directly as the inner product of two atomic neighbor densities ρ and ρ' [defined in Eq. (9)],

$$S(\rho,\rho') = \int \rho(\mathbf{r})\rho'(\mathbf{r})d\mathbf{r}. \tag{29}$$

This clearly satisfies the permutational invariance criterion. Integrating Eq. (29) over all possible rotations of one of the environments leads to a rotationally invariant similarity kernel

$$\begin{aligned} k(\rho,\rho') &= \int |S(\rho,\hat{R}\rho')|^n d\hat{R} \\ &= \int d\hat{R} \left| \int \rho(\mathbf{r})\rho'(\hat{R}\mathbf{r})d\mathbf{r} \right|^n . \end{aligned} \tag{30}$$

It is easy to see that for $n = 1$, all angular information (the relative orientation of individual atoms) is lost because the order of the two integrations can be exchanged, but for $n \geqslant 2$ the kernel retains the angular information of the original environments. The obvious practical difficulty with this construction is the evaluation of the angular integral, which is addressed next.

A. Analytic evaluation a smooth similarity kernel

Retaining the Dirac-delta functions in the definition of the atomic neighbor density would lead to a discontinuous similarity kernel in that the dissimilarity between two environments with very close but not identical atomic positions would be large. Therefore, instead of the Dirac-delta functions, we construct the atomic neighbor density using Gaussians, expanded in terms of spherical harmonic functions as[58]

$$\begin{aligned} &\exp(-\alpha|\mathbf{r}-\mathbf{r}_i|^2) \\ &\quad = 4\pi \exp\left[-\alpha\left(r^2+r_i^2\right)\right] \sum_{lm} \iota_l(2\alpha r r_i)\, Y_{lm}(\hat{\mathbf{r}})\, Y^*_{lm}(\hat{\mathbf{r}}_i), \end{aligned} \tag{31}$$

where ι_l are the modified spherical Bessel functions of the first kind. The atomic neighbor density function is then defined as a sum of Gaussians with one centered on each neighbor,

$$\rho(\mathbf{r}) = \sum_i \exp(-\alpha|\mathbf{r}-\mathbf{r}_i|^2) = \sum_i \sum_{lm} c^i_{lm}(r) Y_{lm}(\hat{\mathbf{r}}), \tag{32}$$

where

$$c^i_{lm}(r) \equiv 4\pi \exp\left[-\alpha\left(r^2+r_i^2\right)\right] \iota_l(2\alpha r r_i) Y^*_{lm}(\hat{\mathbf{r}}_i).$$

The overlap between an atomic environment (unprimed) and a rotated environment (primed) is

$$\begin{aligned} S(\hat{R}) \equiv S(\rho,\hat{R}\rho') &= \int d\mathbf{r}\, \rho(\mathbf{r})\rho'(\hat{R}\mathbf{r}) \\ &= \sum_{i,i'} \sum_{\substack{l,m \\ l',m',m''}} D^{l'}_{m'm''}(\hat{R}) \int dr\, c^{i*}_{lm}(r) c^{i'}_{l'm'}(r) \int d\hat{\mathbf{r}}\, Y^*_{lm}(\hat{\mathbf{r}}) Y_{l'm''}(\hat{\mathbf{r}}) \\ &= \sum_{i,i'} \sum_{l,m,m'} \tilde{I}^l_{mm'}(\alpha,r_i,r_{i'}) D^l_{mm'}(\hat{R}) = \sum_{l,m,m'} I^l_{mm'} D^l_{mm'}(\hat{R}), \end{aligned}$$

where the integral of the coefficients is

$$\begin{aligned} &\tilde{I}^l_{mm'}(\alpha,r_i,r_{i'}) \\ &\quad = 4\pi \exp\left[-\alpha\left(r_i^2+r_{i'}^2\right)/2\right] \iota_l(\alpha r_i r_{i'})\, Y_{lm}(\hat{\mathbf{r}}_i) Y^*_{lm}(\hat{\mathbf{r}}_{i'}) \end{aligned}$$

and

$$I^l_{mm'} \equiv \sum_{i,i'} \tilde{I}^l_{mm'}(\alpha,r_i,r_{i'}). \tag{33}$$

The rotationally invariant kernel with $n = 2$ then becomes

$$\begin{aligned} k(\rho,\rho') &= \int d\hat{R}\, S^*(\hat{R}) S(\hat{R}) \\ &= \sum_{l,m,m'\lambda,\mu,\mu'} \left(I^l_{mm'}\right)^* I^\lambda_{\mu\mu'} \int d\hat{R}\, D^*(\hat{R})^l_{mm'} D(\hat{R})^\lambda_{\mu\mu'} \\ &= \sum_{l,m,m'} \left(I^l_{mm'}\right)^* I^l_{mm'}, \end{aligned} \tag{34}$$

where we used the orthogonality of the Wigner matrices. Although in practice we always use $n = 2$, it is easy to derive the kernel for any arbitrary order n using the Clebsch-Gordan series in Eq. (18). For $n = 3$, using the fact that S as defined in Eq. (29) is real and positive,

$$k(\rho,\rho') = \int d\hat{R}\; S(\hat{R})^3,$$

which can be shown to be

$$k(\rho,\rho') = \sum I^{l_1}_{m_1 m'_1} I^{l_2}_{m_2 m'_2} I^l_{mm'} C^{lm}_{l_1 m_1 l_2 m_2} C^{lm'}_{l_1 m'_1 l_2 m'_2}. \tag{35}$$

Raising a positive-definite function to a positive integer power yields a function that is similarly positive definite. In our context, raising k to some power $\zeta \geqslant 2$ has the effect of accentuating the sensitivity of the kernel to changing the atomic positions, which we generally found to be advantageous in experiments. Therefore, following normalization by dividing by $\sqrt{k(\rho,\rho)k(\rho',\rho')}$, we define the general form of what we call the SOAP kernel as

$$K(\rho,\rho') = \left(\frac{k(\rho,\rho')}{\sqrt{k(\rho,\rho)k(\rho',\rho')}}\right)^\zeta, \tag{36}$$

where ζ is any positive integer.

B. Radial basis and relation to spectra

Note that $I^l_{mm'}$ needs to be computed for each pair of neighbors, which can become expensive for a large number of neighbors. If we expand Eq. (32) using radial basis functions

$g_n(r)$, the atomic neighbor density function becomes

$$\rho(\mathbf{r}) = \sum_i \exp(-\alpha|\mathbf{r} - \mathbf{r}_i|^2) = \sum_{nlm} c_{nlm} g_n(r) Y_{lm}(\hat{\mathbf{r}}), \quad (37)$$

and similarly, the ρ' environment is expanded using coefficients c'_{nlm}. If the radial basis functions form an orthonormal basis, i.e.,

$$\int dr\, g_n(r) g_{n'}(r) = \delta_{nn'},$$

the sum in Eq. (33) becomes

$$I^l_{mm'} = \sum_n c_{nlm} (c'_{nlm'})^*. \quad (38)$$

The significance of this result becomes apparent when substituting Eq. (38) into (34) to obtain

$$k(\rho,\rho') = \sum_{n,n',l,m,m'} c_{nlm} (c'_{nlm'})^* (c_{nlm})^* c'_{n'lm'}$$
$$\equiv \sum_{n,n',l} p_{nn'l}\, p'_{nn'l} \quad (39)$$

since

$$p_{nn'l} \equiv \sum_m c_{nlm} (c_{n'lm})^* \quad (40)$$

is exactly the power spectrum [cf. Eq. (26)] and, analogously, $p'_{nn'l}$ is the power spectrum of the primed environment. Furthermore, the kernel in Eq. (39) is the dot product of the power spectra [cf. Eq. (2)]. Analogously, the kernel for $n = 3$, defined in Eq. (35), can be expressed as

$$k(\rho,\rho') = \sum_{\substack{n_1,n_2,n \\ l_1,l_2,l}} b_{n_1 n_2 n l\, l_1 l_2} b'_{n_1 n_2 n l\, l_1 l_2}, \quad (41)$$

where

$$b_{n_1 n_2 n\, l_1 l_2 l} \equiv \sum c_{n_1 l_1 m_1} c_{n_2 l_2 m_2} (c_{nlm})^* C^{lm}_{l_1 m_1 l_2 m_2}$$

[cf. Eq. (21)], and $\mathbf{b}'$ is analogously the bispectrum of the primed environment. In Fig. 7, we show the convergence of the similarity kernel (39) with increasing number of angular and radial basis functions in the expansion.

In this section, we started out by taking a different approach to the problem of comparing neighbor environments, defining the SOAP similarity kernel (36) directly, rather than going via a descriptor. Equations (39) and (41), however, reveal the relation between SOAP and the SO(3) power spectrum and bispectrum: SOAP is equivalent to using the SO(3) power or bispectrum descriptor together with Gaussian atomic neighbor density contributions and a dot-product covariance kernel. The advantage of SOAP over the previous descriptors is that it eliminates some of the *ad hoc* choices that were needed before, while retaining control over the smoothness of the similarity measure using α, the width of the Gaussians in defining the atomic neighborhood density in Eq. (32) and its sensitivity using the exponent ζ in Eq. (36).

V. NUMERICAL RESULTS

We have derived methods to transform atomic neighborhoods to descriptors that are invariant under the required symmetry operators, however, their relative merit for fitting potential energy surfaces remains to be seen. The faithfulness of the representation, i.e., that no genuinely different configurations should map onto the same descriptor, is of particular interest. As the inverse transformation from the descriptor to atomic coordinates, apart from the simplest cases, is not available, we describe numerical experiments in which we attempt to reconstruct atomic coordinates from descriptors, up to rotations, reflections, and permutations. Descriptors which severely fail in this test are unlikely to be good for fitting potential energy surfaces because entire manifolds of neighbor environments that are genuinely different with widely varying true energies will be assigned the same descriptor, resulting in fitted PES with many degenerate modes. We compare and test the performance of the various descriptors by generating potential energy surfaces for silicon clusters and the bulk crystal using our GAP framework.[8]

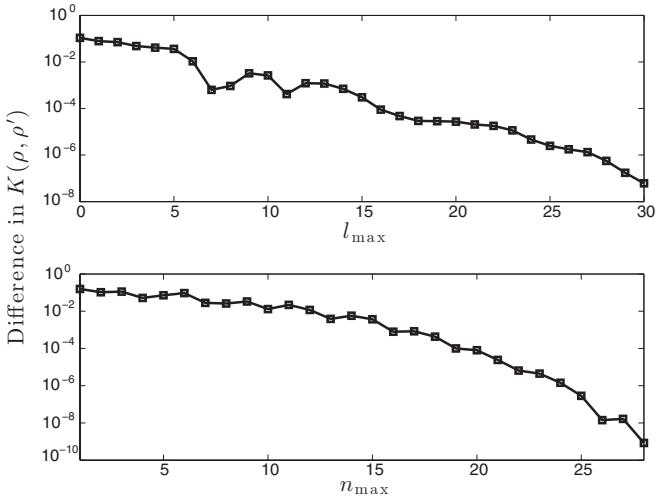

FIG. 7. Convergence of the similarity kernel $K(\rho,\rho')$ of two arbitrary atomic environments with 15 neighbors at different sizes of basis expansion. We used the parameters $\alpha = 0.4$ and $\zeta = 1.0$, and the converged kernel is $K(\rho,\rho') = -0.842\,735$. The top panel shows the convergence of the kernel, evaluated according to Eq. (34), with increasing number of spherical harmonics functions. The bottom panel shows the convergence of the kernel, evaluated according to Eq. (39), with increasing number of radial basis functions, while keeping $l_{\max} = 16$.

A. Reconstruction experiments

Recall that the elements $\mathbf{r}_i \cdot \mathbf{r}_{i'}$ of Σ defined in Eq. (5) are an overcomplete set of basic invariants, which, in the case of atoms scattered on the surface of a unit sphere ($|\mathbf{r}_i| = 1$), are the cosines of the bond angles $\theta_{ii'}$. Thus, the angular parts of all descriptors in Sec. III are permutationally invariant functions of the basic invariants in Σ, and depending on the actual number of descriptor elements used, they may form an incomplete, complete, or overcomplete representation of the atomic environment. In practice, one would like to use as few descriptors as possible, partly due to computational cost, but also because descriptors that use high exponents of the angles are likely to lead to less smooth PESs, as will be shown below.

Given N neighbors, the number of independent degrees of freedom in the neighborhood configuration is $3N - 3$, so we need at most this many *algebraically independent* descriptor elements. But, because the algebraic dependency relationships between the descriptor elements are in general complicated, it is unclear how many descriptor elements are actually needed in order to make the descriptor complete and thus able to uniquely specify an atomic environment of the N neighbors. However, it is possible to conduct numerical experiments in which we compare the descriptors of a fixed target with that of a candidate structure and minimize the difference with respect to the atomic coordinates of the candidate. In this way, we determine if a representation is likely to be complete or not, and in the latter case to characterize the degree of its faithfulness.

1. Descriptor matching procedure

The global minimum of the descriptor difference between the target and the candidate is zero and is always attained on a manifold due to the symmetries built into the descriptors, but for an incomplete descriptor, many inequivalent structures will also appear equivalent, thus enlarging the dimensionality of the global minimum manifold. Furthermore, it can be expected that the descriptor difference function has a number of local minima.

In our experiments, we tried to recover a given target structure after randomizing its atomic coordinates. For each n ($4 \leqslant n \leqslant 19$) we used 10 different Si_n clusters as targets, obtained from a tight-binding[59] molecular dynamics trajectory run at a temperature of 2000 K. For each target cluster, we selected one atom as the origin, randomized the positions of its neighbors by some amount, and then tried to reconstruct the original structure by minimizing the magnitude of the difference between the descriptors of the fixed target and the candidate as the atomic positions of the latter were varied.

In contrast to a general global minimum search problem, we have the advantage of knowing the target value of the objective function at the global minimum. Also, the motivation of our experiments is to find at least one configuration, if it exists, that is genuinely different from the target, but where the descriptors match within a predefined numerical tolerance. Thus, it is sufficient to perform local, gradient-based optimizations starting from random configurations, and reject all local minima (by noting the small gradient of the objective function while the value of the objective function is not small) until we find one where the objective function (the difference in the descriptors) is less than than the specified tolerance. If the configuration thus obtained is genuinely different from the target, the descriptor is shown to be incomplete.

In order to assess the success of the reconstruction procedure (i.e., whether the target and candidate configurations are genuinely different or not) we employed the reference measures defined in Eqs. (7) and (8). However, in some cases it was difficult or impossible to find the right rotation $\hat{R}$ in (8), whereas d_{ref} in (7) proved reliable. For each d_{ref}, an initial $\mathbf{P}$ was generated by ordering the atoms according to their distances from the central atom, then the optimal permutation was found using a simple random search in the space of permutations.

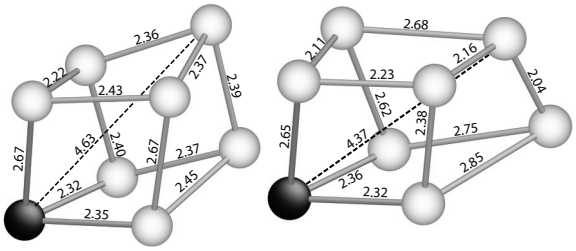

FIG. 8. Two Si_8 clusters that differ by $d_{\mathrm{ref}} = 4.1$ Å^2. The black atoms are taken as the origin in each environment, i.e., the centers of rotations. In terms of Parrinello-Behler–type descriptors, the difference $\sum_\alpha (G_\alpha - G'_\alpha)^2$ between the two atomic environments is 6×10^{-7}. The bond lengths are shown in Ångströms.

We minimized the difference between the target and candidate descriptors in the space of atomic coordinates of the latter using the conjugate gradients algorithm, stopping the minimization when either the gradient or the reference distance d_{ref} became smaller than 10^{-8} Å^2 and 10^{-2} Å^2, respectively. In order to ensure that structures deemed nonequivalent by $d_{\mathrm{ref}} > 10^{-2}$ Å^2 were genuinely different, we cross checked them by noting the value of Δ from Eq. (8) and also employing the atomic fingerprints suggested by Oganov and Valle.[42] To give a sense of the typical magnitude of the d_{ref} measure, the actual difference in terms of atomic distances between two example structures is shown in Fig. 8.

In the first set of reconstruction experiments, in order to provide a fair comparison, the truncation of the formally infinite set of descriptors was chosen in such a way that the finite descriptors had roughly equal numbers of components: 51 in total for the SO(3) bispectrum and PB descriptors and 50 for the AFS and SO(3) power spectrum. This corresponds to a truncation of the SO(4) bispectrum with $2j_{\max} = 5$ (the factor of 2 on account of the half-integer nature of j), the SO(3) bispectrum with $l_{\max} = 4$ and $n_{\max} = 3$, the PB descriptor with its published parameters,[57] and the AFS and SO(3) power spectrum using $l_{\max} = 9$ and $n_{\max} = 5$. We note that in the case of the PB descriptor, the band limit of the angular descriptors (corresponding to our $l_{\max}$ or $j_{\max}$) was $\zeta_{\max} = 16$ and only the values $\zeta = 1,2,4,16$ are used.

Figure 9 shows the quality of reconstruction for different cluster sizes, based on the PB, AFS, SO(3) power spectrum, SO(3) bispectrum, and SO(4) bispectrum as given by the reference distance d_{ref} achieved, averaged over 10 reconstruction trials for each cluster size n. The general trend is the same for all descriptors: as the number of neighbors increases, the average d_{ref} increases, and thus the faithfulness of the reconstruction decreases. Noting that the stopping criterion for the reconstruction process was $d_{\mathrm{ref}} < 10^{-2}$ Å^2, larger randomization of the initial atomic coordinates (bottom panel) reveals the poor representation power for all descriptors using this parameter set for $n > 10$, and the neighbor configuration becomes impossible to determine from the descriptor.

The poor quality of representation is partly attributable to the decrease in sensitivity to the positions of atoms near the cutoff. For example, Fig. 10 shows two Si_8 clusters for which none of the descriptors lead to perfect reconstructions (resulting in the observed peak on Fig. 9). The atom marked A in the figure is within the 6-Å cutoff, but close to it. In

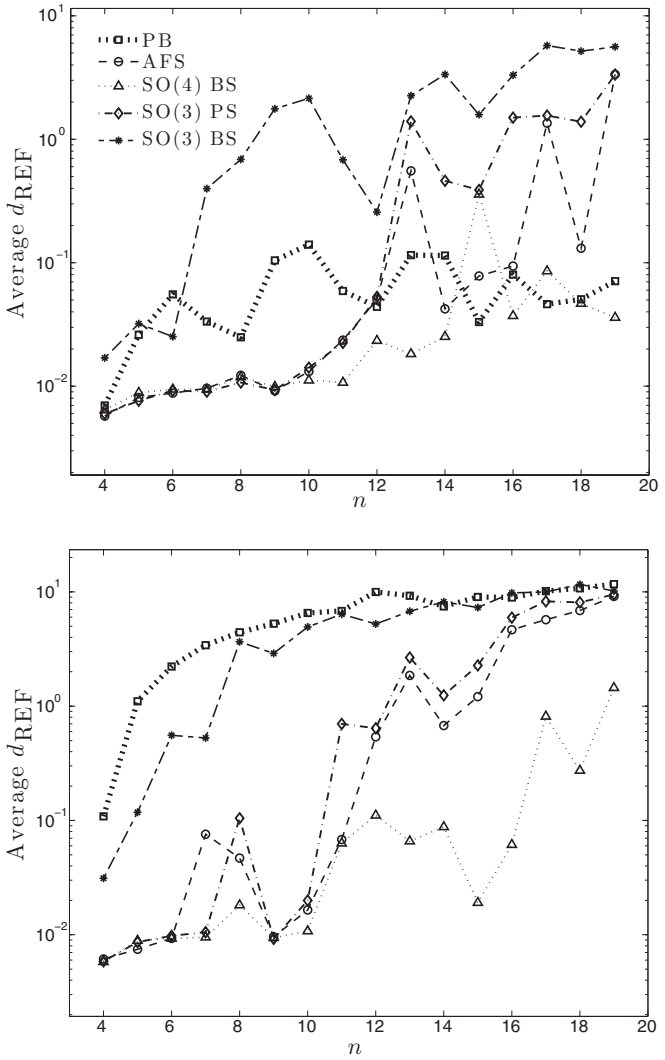

FIG. 9. Difference between the target and reconstructed structures after randomization and minimization for randomizations of 0.2 Å (top) and 1.6 Å (bottom) as a function of the number of atoms in the cluster n averaged over 10 targets for each cluster size. The radial cutoff was 6 Å. Different lines correspond to different descriptors: Parrinello-Behler (PS), angular Fourier series (AFS), bispectrum (BS), and power spectrum (PS). The two versions of the bispectrum differ in the handling of the radial degrees of freedom.

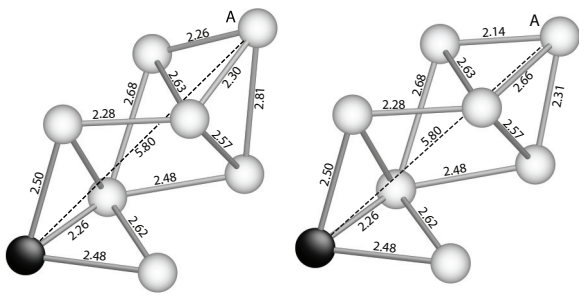

FIG. 10. Two Si_8 clusters that differ by $d_{\mathrm{ref}} = 0.7$ Å^2. The reference atom, i.e., the center of the rotation, is colored black. The only difference between the two clusters is the relative position of the furthest atom A.

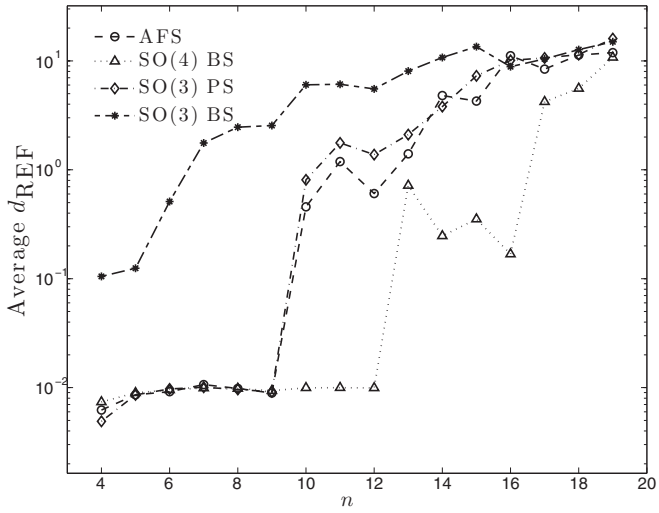

FIG. 11. Difference between the target and reconstructed structures after randomization by 1.6 Å and minimization, as a function of the number of atoms in the cluster n averaged over 10 targets for each cluster size. The cutoff was 9 Å. The line types corresponding to the different descriptors are the same as in Fig. 9.

order to separate out this effect, we repeated the reconstruction experiments with a radial cutoff of 9 Å (omitting the PB descriptor now since there is no published parameter set for this cutoff). The results are shown in Fig. 11 for the larger initial randomization. The peak near $n = 8$ is now absent, and the transition from faithful reconstruction [for $n \leqslant 9$ for the SO(3) power spectrum and AFS, and for $n \leqslant 12$ for the SO(4) bispectrum] to failure for larger n is much clearer.

Since all the descriptors are likely to be overcomplete when the infinite series of the basis set expansion is not truncated, the reconstruction quality is expected to increase with increasing descriptor length. To verify this, Fig. 12 shows the reconstruction quality of the AFS descriptor for varying truncations of the angular part of the basis set. The representation becomes monotonically better for higher angular resolutions. However, this comes at the price of introducing ever more highly oscillating basis functions, which might be less and less suitable for fitting generally smooth potential energy surfaces.

Figure 12 also shows the achieved reference values when using the SOAP similarity measure. In this case, rather than minimizing the difference between descriptors, we optimized the candidate structure until its normalized similarity to the target as given by Eq. (36) was as close to unity as possible. In contrast to the other descriptors, SOAP with the modest band limit of $l_{\max} = 6$ performs perfectly for all structures, without showing any degradation for larger numbers of neighbors.

To verify that the above results are not affected by artifacts of the minimization procedure, e.g., getting stuck, Fig. 13 shows the convergence of the reference measure d_{ref} during a minimization as well as the corresponding convergence of the target function (the difference between the target and candidate descriptors). There was no difficulty in converging the target function to zero (the global minimum) for any of the complete (or overcomplete) descriptors or SOAP, while

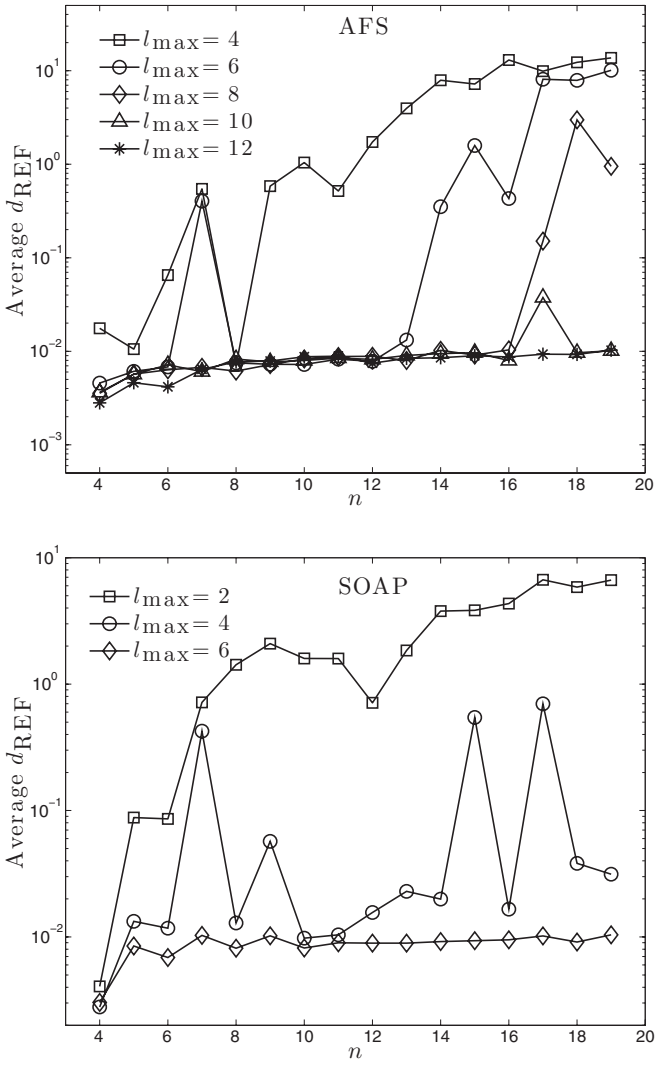

FIG. 12. Difference between target and reconstructed structures after randomization by 1.6 Å and minimization as a function of the number of atoms in the cluster n using the AFS descriptor (top) and SOAP (bottom) with a radial cutoff of 9 Å, averaged over 10 targets for each cluster size. For the case of AFS, the different curves correspond to different numbers of components of the descriptor, achieved by varying the truncation of the angular expansion, while in the case of SOAP, we varied the truncation of the expansion of the atomic neighborhood density, which corresponds to varying the accuracy of the evaluation of the similarity measure.

the reference similarity converged to a nonzero value for incomplete descriptors.

B. Gaussian Approximation Potentials

Our main motivation for assessing different approaches to representing atomic neighbor environments is to determine their efficacy for generating interatomic potentials. Therefore, as a final test, we fitted a series of interatomic potentials for Si_{3-19}, based on different descriptors, using our GAP framework.[8] The training and the testing configurations were obtained from tight-binding[59] molecular dynamics trajectories run at the temperatures between 500 and 2000 K. We used four sets of cluster configurations, containing 2000,

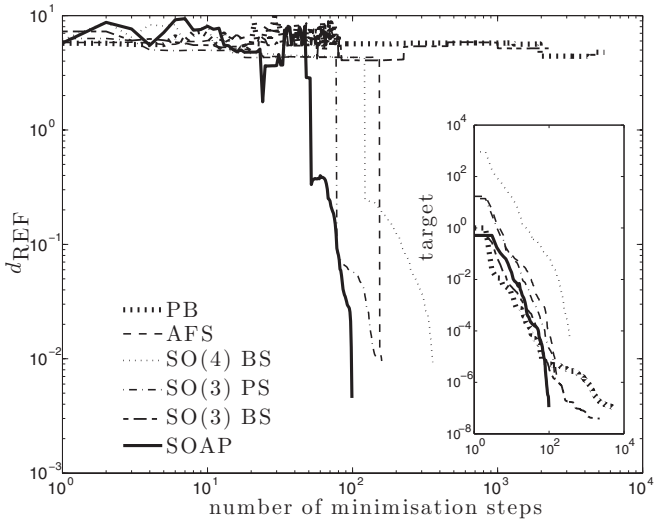

FIG. 13. Convergence of reference distance measure during a typical reconstruction procedure. The inset shows the value of the minimization target approaching zero, i.e., descriptor equivalence, for all of the descriptors.

4000, 6000, and 8000 atomic environments for the training, corresponding to a total of 180, 360, 540, and 720 unique cluster configurations, respectively. The test set contained 12 000 atomic environments, independent from those used in the fitting procedure.

We tested AFS, the SO(3) power spectrum, and the SO(4) bispectrum using the squared exponential covariance kernel (4) as well as SOAP for potential fitting. The accuracy of the resulting potential energy surfaces is shown in Table III as a function of the angular band limit, and in Fig. 14 as a function of the database size. Both demonstrate that SOAP

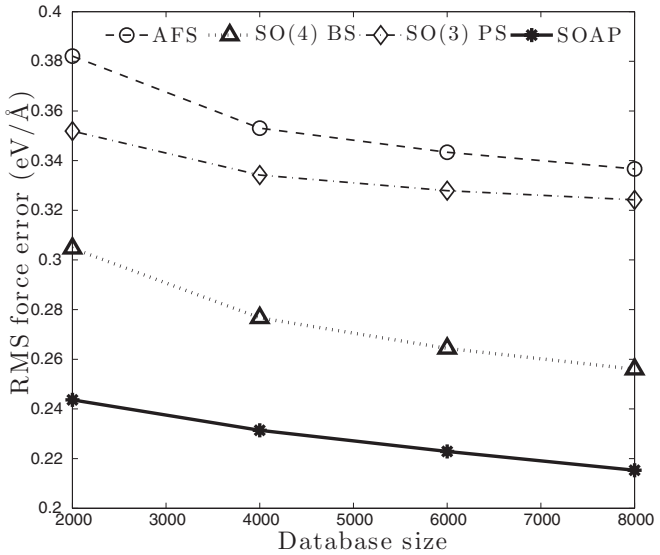

FIG. 14. Quality of GAP potentials constructed using different descriptors, as a function of the size of the database used for the fit, with all configurations drawn from Si_n clusters with $3 \leqslant n \leqslant 19$. The angular band limit was $l_{\max} = 12$ in all cases [equivalent to $2j_{\max} = 12$ for the SO(4) bispectrum].

TABLE III. Quality of the GAP potential energy surface using different descriptors and angular band limits, as measured by the root mean square (RMS) energy and force errors. The fitting database contained 8000 atomic neighborhoods in Si_n clusters with $3 \leqslant n \leqslant 19$. The units of the RMS errors of the energy and force are meV/atom and eV/Å, respectively.

	Angular band limit	RMS (e)	RMS (f)
	$l_{\max}$		
	6	50.0	0.37
AFS	8	47.0	0.35
$n_{\max} = 6$	10	45.7	0.34
	12	44.7	0.34
	$2j_{\max}$		
	6	27.6	0.28
SO(4) BS	8	22.8	0.26
$r_0 = \frac{4}{3} r_{\text{cut}}$	10	20.2	0.25
	12	19.2	0.26
	$l_{\max}$		
	6	41.5	0.36
SO(3) PS	8	37.1	0.34
$n_{\max} = 6$	10	35.7	0.33
	12	35.0	0.32
	$l_{\max}$		
	2	21.4	0.23
SOAP	4	17.6	0.21
$\alpha = 2, \zeta = 4$	6	17.0	0.21
	8	15.3	0.22

outperforms the other descriptors. As can be expected from the reconstruction tests (cf. Fig. 12), the fit gets better with all descriptors when a larger angular resolution is used, with the error not yet saturated for $l_{\max} = 12$. Similarly, increasing the database size makes the fit more accurate, and one can expect improvement if even more than 8000 atomic environments are used (this corresponds to, on average, just 40 configurations for each cluster size).

Perhaps contrary to initial expectations, making all the descriptors more faithful by using a larger angular band limit is not necessarily beneficial. Descriptor components corresponding to high angular momentum channels involve angular basis functions that are highly oscillatory, and can thus degrade the fitted potential energy surface. To demonstrate this, we constructed a GAP model for bulk silicon using a database of configurations with randomly displaced atoms in randomly distorted unit cells, containing two atoms. Figure 15 shows the elastic constants of the GAP fits as a function of the angular band limit for SOAP and the SO(4) bispectrum descriptor (which performed the best in our reconstruction and cluster PES tests compared to the other descriptors). In the case of the bispectrum, the elastic constants of the model improve up to $2j_{\max} = 8$, but then deteriorate dramatically, irrespective of the database size. SOAP does not show this behavior, and leads to reasonable elastic constants using the smaller database, and is already well converged for $l_{\max} = 6$ using the larger database. Given the view of SOAP as an SO(3) power spectrum, it appears that the benefit comes from the combination of building the atomic neighborhood density from smooth Gaussians and using the dot-product kernel,

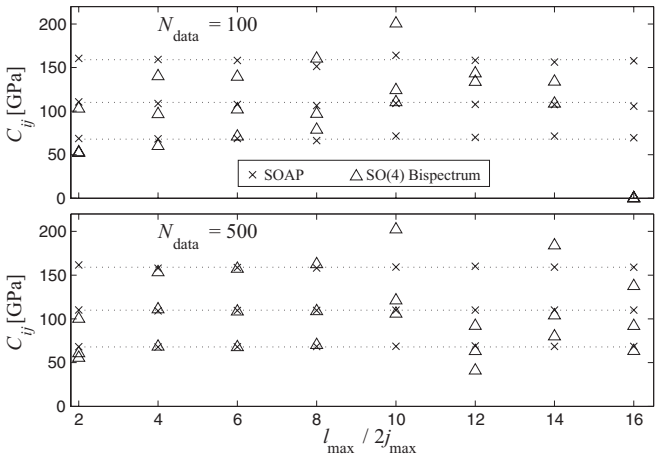

FIG. 15. Elastic constants C_{11}, C_{12}, and C_{44}^0 of GAP models for bulk silicon using the SO(4) bispectrum and SOAP, as a function of the angular band limit $l_{\max}$ (for SOAP) or $j_{\max}$ (for the bispectrum). The top and bottom panels correspond to a database size of 100 and 500 configurations. The dashed line indicates exact values of the tight-binding model which was used to generate the database.

both directly linked to the construction of SOAP as a smooth similarity measure, in contrast to using Dirac-delta functions for the density and a squared exponential kernel for the other descriptors.

VI. CONCLUSION

In this paper, we discussed a number of approaches to representing atomic neighbor environments within a finite cutoff such that the representation is a continuous and differentiable function of the atomic positions and is invariant to global rotations, reflections, and permutations of atoms. We showed that the Steinhardt bond-order parameters are equivalent to certain elements of the SO(3) angular power spectrum and bispectrum. To incorporate radial information, and therefore provide a full description of the atomic neighbor environment, we reviewed the construction of the SO(4) power spectrum and bispectrum as an alternative to introducing explicit radial basis functions. We also demonstrated that all these constructs, as well as the descriptors suggested by Parrinello and Behler, use very similar terms and form part of a general family that is based on the bond angles. In practice, when the expansion is truncated, the faithfulness of the descriptors decreases as the number of neighbors increases, leading to a tunable tradeoff between the size of the descriptor and its faithfulness in terms of its ability to represent the atomic environment uniquely up to symmetries. With typically used parameters, however, the faithfulness of the descriptors is quite different, and all descriptors fail for Si clusters with more than 13 atoms. In order to improve on this, we therefore introduced a similarity measure between atomic neighbor environments called SOAP, which does not suffer from these difficulties and demonstrates excellent faithfulness for any number of neighbors. We also tested the performance of the descriptors for fitting models of small silicon clusters and bulk silicon crystal and found that SOAP leads to a more accurate and much more robust potential energy surface.

ACKNOWLEDGMENTS

A.P.B. gratefully acknowledges funding from Magdalene College, Cambridge. Figures 8 and 10 were generated using AtomEye.[60] This work was partly supported by the European Framework Programme (FP7/2007-2013) under Grant Agreement No. 229205. The authors would like to thank Ferenc Huszár for stimulating discussions on symmetry-invariant kernels.

*ab686@cam.ac.uk

[1] C. J. Pickard and R. J. Needs, Nat. Mater. **7**, 775 (2008).
[2] D. J. Wales, *Energy Landscapes*, Applications to Clusters, Biomolecules and Glasses (Cambridge University Press, Cambridge, UK, 2004).
[3] O. Obrezanova, G. Csányi, J. M. R. Gola, and M. D. Segall, J. Chem. Inf. Model. **47**, 1847 (2007).
[4] M. Segall, J. Comput.-Aided Mol. Des. **26**, 121 (2011).
[5] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[6] L. Raff, R. Komanduri, M. Hagan, and S. Bukkapatnam, *Neural Networks in Chemical Reaction Dynamics* (Oxford University Press, New York, 2012).
[7] B. Braams and J. Bowman, Int. Rev. Phys. Chem. **28**, 577 (2009).
[8] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[9] D. Tromans, Hydrometallurgy **48**, 327 (1998).
[10] J. Ischtwan and M. A. Collins, J. Chem. Phys. **100**, 8080 (1994).
[11] M. A. Collins, Theor. Chem. Acc. **108**, 313 (2002).
[12] T.-S. Ho, T. Hollebeek, H. Rabitz, L. B. Harding, and G. C. Schatz, J. Chem. Phys. **105**, 10472 (1996).
[13] G. G. Maisuradze, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **119**, 10002 (2003).
[14] Y. Guo, A. Kawano, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **121**, 5091 (2004).
[15] X. Huang, B. J. Braams, S. Carter, and J. M. Bowman, J. Am. Chem. Soc. **126**, 5042 (2004).
[16] X. Zhang, B. J. Braams, and J. M. Bowman, J. Chem. Phys. **124**, 021104 (2006).
[17] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, J. Chem. Phys. **103**, 4129 (1995).
[18] H. Gassner, M. Probst, A. Lauenstein, and K. Hermansson, J. Phys. Chem. A **102**, 4596 (1998).
[19] S. Lorenz, A. Groß, and M. Scheffler, Chem. Phys. Lett. **395**, 210 (2004).
[20] A. Brown, B. Braams, K. Christoffel, Z. Jin, and J. Bowman, J. Chem. Phys. **119**, 8790 (2003).
[21] X. Huang, B. J. Braams, and J. M. Bowman, J. Chem. Phys. **122**, 044308 (2005).
[22] S. Manzhos and T. Carrington, J. Chem. Phys. **125**, 194105 (2006).
[23] S. Manzhos, X. Wang, R. Dawes, and T. Carrington, J. Phys. Chem. A **110**, 5295 (2006).
[24] C. Handley and P. Popelier, J. Chem. Theor. Comput. **5**, 1474 (2009).
[25] H. Partridge and D. W. Schwenke, J. Chem. Phys. **106** 4618 (1997).
[26] R. H. Tipping and A. Forbes, J. Mol. Spectrosc. **39**, 65 (1971).
[27] W. Cencek, K. Szalewicz, C. Leforestier, R. v. Harrevelt, and A. v. d. Avoird, Phys. Chem. Chem. Phys. **10**, 4716 (2008).
[28] E. Sanville, A. Bholoa, R. Smith, and S. D. Kenny, J. Phys.: Condens. Matter **20**, 285219 (2008).
[29] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[30] J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, Phys. Rev. Lett. **100**, 185501 (2008).
[31] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, Phys. Rev. Lett. **108**, 115701 (2012).
[32] N. Artrith, T. Morawietz, and J. Behler, Phys. Rev. B **83**, 153101 (2011).
[33] T. Morawietz, V. Sharma, and J. Behler, J. Chem. Phys. **136**, 064103 (2012).
[34] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).
[35] R. Kondor, arXiv:cs/0701127.
[36] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[37] C. E. R. Williams and C. K. I, *Gaussian Processes for Machine Learning* (MIT Press, Boston, 2007).
[38] D. Mackay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003).
[39] R. M. Neal, Ph.D. thesis, Department of Computer Science, University of Toronto, 1995.
[40] H. Weyl, *The Classical Groups*, Their Invariants and Representations (Princeton University Press, Princeton, NJ, 1946).
[41] S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi, Bioinformatics **21**, i359 (2005).
[42] M. Valle and A. Oganov, Acta Crystallogr. A **66**, 507 (2010).
[43] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskii, *Quantum Theory of Angular Momentum* (World Scientific, Singapore, 1987).
[44] J. P. K. Doye, M. A. Miller, and D. J. Wales, J. Chem. Phys. **110**, 6896 (1999).
[45] L. B. Pártay, A. P. Bartók, and G. Csányi, J. Phys. Chem. B **114**, 10502 (2010).
[46] C. Chakravarty and R. M. Lynden-Bell, J. Chem. Phys. **113**, 9239 (2000).
[47] R. D. Mountain and A. C. Brown, J. Chem. Phys. **80**, 2730 (1984).
[48] J. Duijneveldt and D. Frenkel, J. Chem. Phys. **96**, 4655 (1992).
[49] E. R. Hernández and J. Íñiguez, Phys. Rev. Lett. **98**, 055501 (2007).
[50] A. van Blaaderen and P. Wiltzius, Science **270**, 1177 (1995).
[51] R. Biswas and D. R. Hamann, Phys. Rev. B **36**, 6434 (1987).
[52] R. Kakarala, Ph.D. thesis, Department of Mathematics, UC Irvine, 1992.
[53] H. Maschke, Math. Ann. **50**, 492 (1898).
[54] C. D. Taylor, Phys. Rev. B **80**, 024104 (2009).
[55] A. V. Meremianin, J. Phys. A: Math. Gen. **39**, 3099 (2006).
[56] M. A. Caprio, K. D. Sviratcheva, and A. E. McCoy, J. Math. Phys. **51**, 093518 (2010).
[57] N. Artrith and J. Behler, Phys. Rev. B **85**, 045439 (2012).
[58] K. Kaufmann and W. Baumeister, J. Phys. B: At., Mol. Opt. Phys. **22**, 1 (1989).
[59] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, and R. Kaschner, Phys. Rev. B **51**, 12947 (1995).
[60] J. Li, Modell. Simul. Mater. Sci. Eng. **11**, 173 (2003).