LLNL-JRNL-757813

Representing local atomic environment using descriptors based on local correlations

A. Samanta

September 7, 2018

Journal of Chemical Physics

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Representing local atomic environment using descriptors based on local correlations

Amit Samanta*
Physics Division, Lawrence Livermore National Laboratory, Livermore, California 94550, USA

(Dated: December 21, 2018)

Statistical learning of material properties is an emerging topic of research and has been tremendously successful in areas such as representing complex energy landscapes as well as in technologically relevant areas, like identification of better catalysts and electronic materials. However, analysis of large data sets to efficiently learn characteristic features of a complex energy landscape, for example, depends on the ability of descriptors to effectively screen different local atomic environments. Thus, discovering appropriate descriptors of bulk or defect properties and the functional dependence of such properties on these descriptors remains a difficult and tedious process. To this end, we develop a framework to generate descriptors based on many-body correlations that can effectively capture intrinsic geometric features of the local environment of an atom. These descriptors are based on the spectrum of two-body, three-body, four-body and higher order correlations between an atom and its neighbors, and are evaluated by calculating the corresponding two-body, three-body, four-body overlap integrals. They are invariant to global translation, global rotation, reflection and to permutations of atomic indices. By systematically testing the ability to capture local atomic environment, it is shown that the local correlation descriptors are able to successfully reconstruct structures containing 10-25 atoms which was previously not possible.

I. INTRODUCTION

Physical processes, such as phase transformations, crack propagation, microstructural evolution during mechanical deformation or annealing, evolution of dislocation networks in cold worked metals, are inherently multiscale in nature. Atomistic simulations using empirical force fields play an important role in bridging the gap in length scales accessible in quantum mechanical methods (at atomic length scales) and continuum scale methods (at the macroscopic length scales). In the past decade, a paradigm shift in the development of inter-atomic potentials has taken place that has been aided by the advancement of computing power, storage capabilities and the use of machine learning tools. This change was ushered in by a few notable works: (a) artificial neural network based potentials by Behler and Parrinello,[1] and (b) Gaussian approximation potentials by Albert Bartók, Gábor Csýani and Risi Kandor.[2] In these new kinds of potentials, the high-dimensional potential energy landscape of a system containing thousands, or millions of atoms is mapped onto a low-dimensional space of a few descriptors. These descriptors are quantities that can accurately quantify the local neighborhood information of an atom. Subsequently, fitting techniques like least squares regression (with l_1 or l_2 regularization)[3–8], non-linear regression (Gaussian process regression)[2,4,9–12] and artificial neural networks have been used to develop a sparse representation of a potential energy landscape. We refer to these new generation of interatomic potentials developed using machine learning tools as machine learned potentials (MLP).

Even though the computational cost to calculate the force on an atom from these MLP is a few orders of magnitude more than that required in traditional interatomic potentials, this cost is still three or four orders of magnitude cheaper than direct quantum mechanical calculations.[13,14] In addition, the accuracy of these MLP is comparable to that of the training set (which is usually generated from first-principles calculations).[2,3,6,15] This significant improvement in accuracy when compared to classical model potentials[16,17] and the ability to handle multi-component systems easily means that such MLP can potentially be used to study complex processes, such as phase transitions under dynamic loading conditions (like during ramp or shock compression), that have proved difficult to handle using traditional potentials.

Developing a machine learned potential involves generating a large data-set of atomic environments and forces using density functional theory or other quantum chemistry techniques.[18] The size and quality of the database used to obtain the MLP plays an important role in determining its transferability and its ability to capture essential features of an energy landscape.[14,19] To this end, efficient sampling methods like order parameter aided free energy sampling or methods to efficiently sample the configuration space in an unbiased manner can play an important role.[20–23]

The accuracy of a MLP is, however, mostly controlled by the ability of its descriptors to accurately capture the local environment surrounding an atom. In the last few years, many descriptors have been proposed to capture the local environment of an atom. These include, 4-dimensional hyperspherical harmonics,[7,9] SO(4) bi-spectrum components,[9] SO(3) bi-spectrum and power spectrum, eigenspectrum of a Coulomb matrix,[3] Chebyshev polynomials,[9,24,25] Bessel functions,[6] overlap between Gaussian type atomic orbitals.[5,26] Using a different approach, Zhu et al. proposed atom centered overlap matrix to capture the overlap between an atom and its neighbors and the eigenvalues of this matrix are the descriptors of local environment.[26] A set of basis functions that systematically incorporates higher order moments of relative positions of neighbors of an atom was proposed in Ref. [27] to develop moment tensor potentials. This can be considered to be a generalization of the notion of internal vectors proposed by Li et al.[10] On the other hand, in neural network based potentials, the radial environment and the angular dependencies are captured using a variety of two-body and three-body terms.[1,28–30] Recently, Gastegger et al.[31] extended these descriptors to handle a multi-component system. Takahashi et al.[32] on the other hand, have shown that using a second-order polynomial approximation with pairwise and angular dependent descriptors can be used to generate MLP with improved transferability and accuracy. The deep potential molecular dynamics method is another neural network based potential, but it uses descriptors based on the Coulomb matrix and its variants.[33]

Bartok et al. in Ref. [9] tested these descriptors by probing their ability to reconstruct small clusters. Using many descriptors (i.e. three and four-dimensional power spectrum, bispectrum, Parrinello-Behler type descriptors and Chebyshev polynomials), the authors were able to successfully reconstruct clusters containing 4 to 12 Si atoms to a varying level of accuracy. But, for reasons not clear, all of these descriptors were not suitable for reconstruction of clusters containing more than 12 atoms. A majority of these descriptors specifically include two-body and three-body correlations, but their inability to reconstruct larger clusters perhaps suggests that a systematic analysis of the importance of many-body correlations to describe local environments is important. The computational cost of a machine learned potential is determined to a large extent by factors like how many descriptors are required to describe the local environment and the number of operations required to calculate each descriptor. Thus, the development of descriptors that can accurately capture local environments, so that they can be used to reconstruct small clusters (containing less than 10-12 atoms) or large clusters (containing more than 10-12 atoms), is important to improve the accuracy and transferability, and decrease the computational cost of machine learned potentials.

We propose a framework to generate descriptors based on many-body correlations and illustrate that they can effectively capture intrinsic geometric features of the local environment of an atom. These descriptors are based on the spectra of various two-body, three-body, four-body and higher order correlations between an atom and its surroundings. Furthermore, we have explored the relationship between the spectra of these correlation matrices and the spectrum of the adjacency matrix of a regular connected graph. Using these descriptors we were able to successfully reconstruct clusters with sizes ranging from 5 to 25 atoms which was not possible using existing descriptors. This suggests that descriptors that incorporate many-body correlations are important to properly embed neighborhood information in MLP.

The remainder of the paper takes the following form. The importance of local correlations is discussed in Section II, which is followed by a detailed discussion of the procedure used to obtain two-body, three-body, four-body and five-body correlation descriptors in Sections II A-II D. In Section III, the descriptors developed in Section II were tested by performing reconstruction simulations of clusters containing 5 to 25 atoms. This is followed by discussions in Section IV and summary in Section V.

II. DESCRIPTORS BASED ON LOCAL DENSITY CORRELATIONS

To design descriptors that embed local correlations, we first consider how the effective atomic density of an atom is affected by the presence of other atoms. To this end, let us consider a structure S_N at a point $\mathbf{X} \in \mathbf{R}^{3N}$ in the configuration space containing N atoms, located at $\{\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \cdots, \mathbf{r}_N\}$, such that $\mathbf{r}_i \in \mathbf{R}^3$ for i = 1, 2, ... , N. The atomic density of such a system is typically written as

$$\rho\left(\mathbf{r}\right) = \sum_{i=1}^{N} \delta\left(\mathbf{r} - \mathbf{r}_i\right), \tag{1}$$

which is a superposition of the densities of all the N atoms. Here, the effective density of each atom is approximated by a Dirac δ-function. It is important to note that the effective atomic density (which is closely related to the accessible free volume) is affected by the local symmetry, local packing density and the spatial separation between neighbors. Hence, a simple superposition of spatially separated Dirac δ-functions as shown in Eq. 1, completely ignores the correlation between an atom and its immediate neighbors. A simple solution is to replace the δ-functions in Eq. 1 by effective densities that can be calculated by starting with a guess density and by systematically accounting for the overlap between an atom and its neighbors.

To this end, we assume the density of an atom, with index i, when it is isolated from any other atoms is represented by a smooth and differentiable function $\rho_1\left(\mathbf{r}, \mathbf{r}_i\right)$, such that $\int_{-\infty}^{\infty} \rho_1\left(\mathbf{r}, \mathbf{r}_i\right) d\mathbf{r}$ = 1. We represent the effective density of this atom, when placed amongst a distribution of other atoms (placed at $\mathbf{r}_1$, $\mathbf{r}_2$, $\cdots$, $\mathbf{r}_{Ni}$), by

$$\bar{\rho}\left(\mathbf{r}, \mathbf{r}_i | \mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_{Ni}\right) = p_1 - p_2 + p_3 - p_4 + \cdots, \tag{2}$$

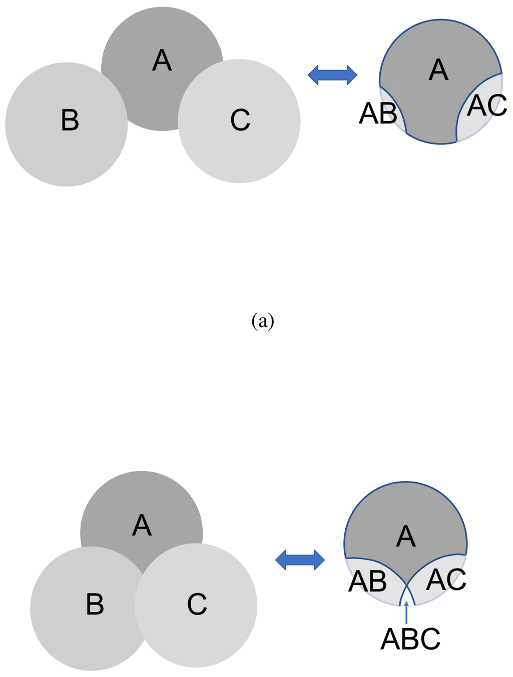
and this effective density can be used to replace the Dirac δ-functions in Eq. 1. Here, $p_1 = \rho_1\left(\mathbf{r}, \mathbf{r}_i\right)$ is the probability of placing the atom at $\mathbf{r}_i$, N_i is the number of neighbors of i and p_k is the overlap between the densities of the atom at $\mathbf{r}_i$ and its k neighbors. For example, two-body correlations can be captured by the overlap between atom i and its neighbors (see Fig. 1)

$$p_2 = \sum_j \rho_1\left(\mathbf{r}, \mathbf{r}_i\right) \rho_1\left(\mathbf{r}, \mathbf{r}_j\right). \tag{3}$$

Thus, when only two-body correlations are present, the effective density of an atom i that captures the overlap between the densities of the atom i and its neighbor j is given by

$$\begin{aligned} \bar{p} &= \sum_j \left[\rho_1\left(\mathbf{r}, \mathbf{r}_i\right) - \rho_1\left(\mathbf{r}, \mathbf{r}_i\right) \rho_1\left(\mathbf{r}, \mathbf{r}_j\right)\right] \\ &= \sum_j \rho_1\left(\mathbf{r}, \mathbf{r}_i\right) \Big[1 - \rho_1\left(\mathbf{r}, \mathbf{r}_j\right)\Big]. \end{aligned} \tag{4}$$

Here, the sum is over all possible neighbors of atom i in the system. Similarly, the overlap between the atom i and two of



(a)

(b)

FIG. 1: The schematic in $1(a)$ illustrates the relevance of two-body correlations: sets A and B, and sets A and C have non-zero overlap between them, while sets B and C do not overlap. Similarly, the schematic in $1(b)$ illustrates the relevance of both two and three-body correlations. There is a finite overlap between any two sets and A∩B∩C is also finite. When sets A, B and C represent guess atomic densities of three atoms, then correlations captured in $1(a)$ and $1(b)$ correspond to p_2 and p_3 in Eqs. 3 and 5, respectively.

its neighbors (with indices j and k) is given by (see Fig. 1)

$$p_3 = \sum_{\langle jk \rangle} \rho_1\left(\mathbf{r}, \mathbf{r}_i\right) \rho_1\left(\mathbf{r}, \mathbf{r}_j\right) \rho_1\left(\mathbf{r}, \mathbf{r}_k\right), \quad (5)$$

and the effective density of the atom i, when both two and three-body correlations are taken into account, is given by

$$\bar{p} = \sum_{\langle jk \rangle} \rho_1\left(\mathbf{r}, \mathbf{r}_i\right) \Big[1 - \rho_1\left(\mathbf{r}, \mathbf{r}_j\right)\Big] \Big[1 - \rho_1\left(\mathbf{r}, \mathbf{r}_k\right)\Big], \quad (6)$$

where the sum is over all possible pairs of neighbors of atom i.

Following the same procedure as mentioned above, we see that the overlap between densities of atom i and three of its neighbors is given by

$$p_4 = \sum_{\langle jkl \rangle} \rho_1\left(\mathbf{r}, \mathbf{r}_i\right) \rho_1\left(\mathbf{r}, \mathbf{r}_j\right) \rho_1\left(\mathbf{r}, \mathbf{r}_k\right) \rho_1\left(\mathbf{r}, \mathbf{r}_l\right), \quad (7)$$

and the effective density of the atom i, when both two, three and four-body correlations are taken into account is given by

$$\begin{aligned} \bar{p} = \sum_{\langle jkl \rangle} \rho_1\left(\mathbf{r}, \mathbf{r}_i\right) \Big[1 - \rho_1\left(\mathbf{r}, \mathbf{r}_j\right)\Big] \Big[1 - \rho_1\left(\mathbf{r}, \mathbf{r}_k\right)\Big] \\ \times \Big[1 - \rho_1\left(\mathbf{r}, \mathbf{r}_l\right)\Big]. \end{aligned} \quad (8)$$

Here j, k and l are indices of three atoms in the neighborhood of atom i. In the following discussion we show that descriptors derived from the spectra of the two-body, three-body, four-body correlations p_2, p_3, p_4. are sensitive to the arrangement of neighbors and can capture the underlying symmetry of the distribution of atoms in S_N. Some of these many-body correlations are illustrated in Figs. 3, 4 and 5.

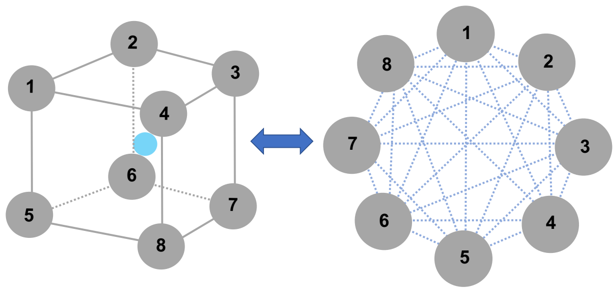


FIG. 2: A schematic representation of the mapping of neighbors, numbered from 1 through 8, of an atom (shown in blue) in a cluster (left) to a graph (right). The nodes of the graph correspond to the neighbors of the central atom, and each edge corresponds to a *bond* between a pair of neighbors. In this case, each node j is connected to 7 other nodes with edges whose weights are given by $\rho^{(2)}_{jk}\left(\sqrt{2}\sigma\right)$, where $k \neq j$.

For a systematic analysis of correlations, the atomic density $\rho_1\left(\mathbf{r}, \mathbf{r}_i\right)$ of an isolated atom is approximated by a normalized Gaussian distribution, i.e.

$$\rho_1\left(\mathbf{r}, \mathbf{r}_i\right) = \frac{1}{\left(2\pi\sigma^2\right)^{3/2}} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_i\right|^2}{2\sigma^2}}. \quad (9)$$

Here, σ determines the spread of the Gaussian. While the width σ is a parameter that can be cross-validated, for the results presented in Section III, σ was set to be equal to the distance corresponding to the local minimum between the first and the second peaks of the pair correlation function. Using the single particle density in Eq. 9, in Sections II A, II B, II C and II D we present in detail the procedure used to obtain two-body, three-body, four-body and five-body correlation descriptors. It is important to note here that using a Gaussian density distribution allows us to analytically evaluate the overlap integrals. We also note that the analysis presented below can be generalized using more generic functions, for example by using Gaussian type orbitals. But, the results presented in Section III show that this simple Gaussian function can capture important geometric features of local neighborhood of an atom. In order to calculate these correlation descriptors, a knowledge of the spatial location of the neighbors of an atom is important. In a typical molecular dynamics simulation, local neighborhood information is preserved in the form of a neighbor-list. Hence, we assume that atom i has N_i neighbors that lie inside a sphere of radius r_{o} centered at $\mathbf{r}_i$.

A. Two-body correlations

We start by describing the procedure used to obtain two-body correlation descriptors. The overlap between the densities of two atoms i and j, located at $\mathbf{r}_i$ and $\mathbf{r}_j$, respectively, is

$$\begin{aligned}\mathcal{C}^{(2)}\left(\mathbf{r}_i,\mathbf{r}_j\right) &= \int_{-\infty}^{\infty} \rho_1\left(\mathbf{r},\mathbf{r}_i\right)\rho_1\left(\mathbf{r},\mathbf{r}_j\right)d\mathbf{r} \\ &= \frac{1}{\left(2\pi\sigma^2\right)^3}\int_{-\infty}^{\infty} e^{\frac{-|\mathbf{r}-\mathbf{r}_i|^2}{2\sigma^2}} e^{\frac{-|\mathbf{r}-\mathbf{r}_j|^2}{2\sigma^2}} d\mathbf{r} \\ &= \frac{1}{8}\frac{1}{\left(\pi\sigma^2\right)^{3/2}}\rho_{ij}^{(2)}\left(\sqrt{2}\sigma\right),\ \ \rho_{ij}^{(2)}\left(\sqrt{2}\sigma\right) = e^{\frac{-|\mathbf{r}_i-\mathbf{r}_j|^2}{4\sigma^2}}\end{aligned} \tag{10}$$

Using this overlap function, we propose two descriptors that can capture local geometric information:

i. Let $\mathbf{u}^{(1)}$ be a vector that quantifies the overlap between i and its neighbors, i.e. $\mathbf{u}^{(1)} = \left[\rho_{i1}^{(2)}\ \rho_{i2}^{(2)}\ \rho_{i3}^{(2)}\ \cdots\rho_{iN_i}^{(2)}\right]$, where $\rho_{i1}^{(2)} = \rho_{i1}^{(2)}\left(\sqrt{2}\sigma\right)$, $\rho_{i2}^{(2)} = \rho_{i2}^{(2)}\left(\sqrt{2}\sigma\right)$, etc. Since, $\rho_{ij}^{(2)}$ depends on the norm of the distance between the atom i and its neighbor j, i.e. $|\mathbf{r}_i - \mathbf{r}_j|^2$, it is invariant to global translation, and other unitary transformations such as global rotation, i.e. if $\mathbf{U}$ is an unitary matrix and if $\mathbf{r}_i \to \mathbf{U}\mathbf{r}_i$ and $\mathbf{r}_j \to \mathbf{U}\mathbf{r}_j$, then $|\mathbf{U}\mathbf{r}_i - \mathbf{U}\mathbf{r}_j|^2 = |\mathbf{r}_i - \mathbf{r}_j|^2$. The descriptor $\mathbf{u}^{(1)}$ is, however, not invariant to permutation of two or more atomic indices. A simple way to make $\mathbf{u}^{(1)}$ invariant to permutations of atomic indices is to sort the two-body correlations such that $\rho_{i1}^{(2)} \geq \rho_{i2}^{(2)} \geq \cdots \rho_{iN_i}^{(2)}$.

ii. Another descriptor that captures two-body correlations is motivated by the use of discrete diffusion kernels and tools from spectral graph theory for the analysis of large data sets.[34–36] We consider the N_i neighbors of an atom i to be located at the vertices of a graph G (see Fig. 2) and each neighbor j of i is connected to another vertex k (which is also a neighbor of vertex i) by an edge with weights given by $w_{jk} = w_{kj} = \rho_{jk}^{(2)}\left(\sqrt{2}\sigma\right)(1-\delta_{jk}) = e^{\frac{-|\mathbf{r}_j-\mathbf{r}_k|^2}{4\sigma^2}}(1-\delta_{jk})$. These weights, obtained from a scalar-valued Gaussian kernel, are symmetric with respect to the interchange of positions of atoms j and k and provide a local measure of the proximity between two atoms. If we assume that all neighbors of an atom i interact with each other via such weights, then each vertex in the graph G is connected to all other vertices in G. Thus, there exists a path connecting any two vertices j and k, meaning that G is a connected regular graph.[37] The adjacency matrix of G, denoted by $\mathbf{K}^{(2)}$, is a $N_i \times N_i$ real symmetric matrix with entries $K_{jk}^{(2)} = w_{jk}$.

The adjacency matrix $\mathbf{K}^{(2)}$ has many interesting properties: Since it is a symmetric matrix, its eigenvalues are real. Further, the diagonal elements of $\mathbf{K}^{(2)}$ are zero meaning that the trace of $\mathbf{K}^{(2)}$, and the sum of its eigenvalues is zero. Since, the two-body overlap function, $\rho_{jk}^{(2)}\left(\sqrt{2}\sigma\right)$ (see Eq. 10), depends on the L_2-norm of the distance between two vertices, it is invariant to global translation and rotation. From the Perron-Frobenius theorem, we know that the largest eigenvalue, $\lambda_{\max}^{(2)}$, of $\mathbf{K}^{(2)}$ lies between the average degree and the maximum degree of the graph G and if G is bipartite then the smallest eigenvalue of $\mathbf{K}^{(2)}$ is related to the largest eigenvalue: $\lambda_{\min}^{(2)} = -\lambda_{\max}^{(2)}$.[37] The adjacency matrix (after proper normalization) is related to the graph Laplacian that is commonly used in dimension reduction techniques like Laplacian eigenmaps, diffusion maps, and clustering techniques, like spectral clustering.[35,36,38–40] Coifman et al. have shown that the diffusion distance (captured by the two-body overlap function $\rho^{(2)}$) is an important geometric quantity that links the spectral theory of the Markov process to the geometry, density and distribution of the data.[36] Further, it was shown that a small subset of the eigenfunctions of the graph Laplacian can be used to construct a low-dimensional geometric embedding of the high-dimensional data set.[35,36,40] Thus, the set of N_i eigenvalues (denoted by the N_i-dimensional vector $\mathbf{u}^{(2)}$) of $\mathbf{K}^{(2)}$ is a descriptor that captures the essential geometric features of the local environment of atom i. These eigenvalues are invariant to global translation, rotation and to permutations of atomic indices of the neighbors.

We note here that graph kernels, which are actively used to compare graphs, encode intrinsic geometric features and are also an attractive route to develop accurate machine learning potentials.[41–43] In non-linear regression techniques like Gaussian process regression, the development of kernels that can effectively distinguish between two different atomic environments has also received attention. Typically, Gaussian kernels (i.e. $\exp\left(-|\mathbf{X}_i - \mathbf{X}_j|^2/2\sigma^2\right)$, where $\mathbf{X}_i$ and $\mathbf{X}_j$ are the positions of two structures and σ is the width) are used to measure the proximity between two different environments and construct the covariance matrix for Gaussian process regression based potential fitting.[44] Duvenaud et al. have shown that the ability to capture intrinsic structural features can be significantly improved by including correlations between intrinsic degrees of freedom.[45] Along this line, based on the overlap integral between two configurations, Glielmo et al. proposed a kernel that includes the correlation between all the atoms in structures being compared.[11] These authors further proposed using covariant kernels that naturally incorporate rotation and reflection symmetries. This idea was later generalized by Grisafi and co-workers to design kernels that explicitly include tensorial properties of arbitrary order.[12] Similarly, the Coulomb kernel[3] and the symmetric overlap between atomic positions (SOAP) kernel[9] have also been successfully used to predict energies of different structures with high accuracy.[43,46]

B. Three-body correlations

Next, we consider descriptors based on three-body correlations. Two important three-body correlations that arise due to the overlap between densities of an atom with index i and two of its neighbors, with indices j and k, are:

iii. the correlation due to the overlap between the densities of atoms i and j, and between atoms i and k. This is denoted by $\mathcal{C}^{(31)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right)$.

iv. the correlation due to a non-zero overlap between the densities of all the three atoms, i.e. atom i and its two neighbors j and k. This is denoted by $\mathcal{C}^{(32)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right)$.

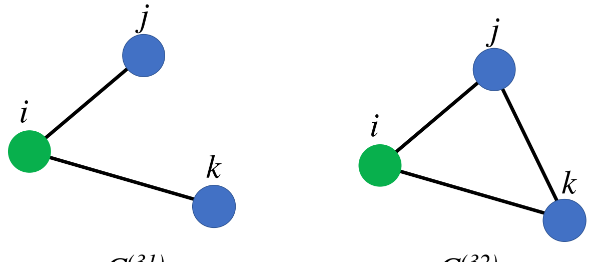


FIG. 3: A schematic representation of correlations in a cluster containing 3 atoms. Each edge represents an overlap between atomic densities of two atoms.

These correlations are illustrated graphically in Fig. 3. The correlation $\mathcal{C}^{(31)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right)$ is similar to a star graph while the correlation $\mathcal{C}^{(32)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right)$ tracks closed paths. To calculate these correlations we consider the following overlap integral:

$$\begin{aligned}\mathcal{C}^{(31)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right) &= \frac{1}{\left(2\pi\sigma^2\right)^6}\left[\int_{-\infty}^{\infty} e^{\frac{-|\mathbf{r}-\mathbf{r}_i|^2}{2\sigma^2}} e^{\frac{-|\mathbf{r}-\mathbf{r}_j|^2}{2\sigma^2}} d\mathbf{r}\right] \\ &\times \left[\int_{-\infty}^{\infty} e^{\frac{-|\mathbf{r}'-\mathbf{r}_i|^2}{2\sigma^2}} e^{\frac{-|\mathbf{r}'-\mathbf{r}_k|^2}{2\sigma^2}} d\mathbf{r}'\right] \\ &= \frac{1}{64}\frac{1}{\left(\pi\sigma^2\right)^3}\,\rho_{ij}^{(2)}\left(\sqrt{2}\sigma\right)\rho_{ik}^{(2)}\left(\sqrt{2}\sigma\right)\end{aligned} \tag{11}$$

For each atom i, the extent of the overlap between densities of two neighbors j and k is denoted by the matrix $\mathbf{K}^{(31)}$, where $K_{jk}^{(31)} = \rho_{ij}^{(2)}\left(\sqrt{2}\sigma\right)\rho_{ik}^{(2)}\left(\sqrt{2}\sigma\right)\left(1-\delta_{jk}\right)$ and $j, k = 1, 2, 3, \cdots, N_i$. Let $\lambda_1^{(31)} \geq \lambda_2^{(31)} \geq \cdots \geq \lambda_{N_i}^{(31)}$ be the eigenvalues of the symmetric $N_i \times N_i$ square matrix $\mathbf{K}^{(31)}$. Thus, the vector $\mathbf{u}^{(31)} = [\lambda_1^{(31)}\ \lambda_2^{(31)}\ \lambda_3^{(31)}\ \cdots\ \lambda_{N_i}^{(31)}]$ is a descriptor that incorporates three body correlations described by $\mathcal{C}^{(31)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right)$.

The other three-body correlation, $\mathcal{C}^{(32)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right)$, corresponds to the following overlap integral:

$$\begin{aligned}&\mathcal{C}^{(32)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right) \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{9/2}}\int_{-\infty}^{\infty} e^{\frac{-|\mathbf{r}-\mathbf{r}_i|^2}{2\sigma^2}} e^{\frac{-|\mathbf{r}-\mathbf{r}_j|^2}{2\sigma^2}} e^{\frac{-|\mathbf{r}-\mathbf{r}_k|^2}{2\sigma^2}} d\mathbf{r} \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{9/2}}\, e^{\frac{-1}{3}\frac{1}{2\sigma^2}\left(|\mathbf{r}_i-\mathbf{r}_j|^2+|\mathbf{r}_j-\mathbf{r}_k|^2+|\mathbf{r}_k-\mathbf{r}_i|^2\right)} \\ &\quad\left[\int_{-\infty}^{\infty} e^{\frac{-|\mathbf{r}-(\mathbf{r}_i+\mathbf{r}_j+\mathbf{r}_k)/3|^2}{2(\sigma/\sqrt{3})^2}} d\mathbf{r}\right] \\ &= \frac{1}{3\sqrt{3}\ \left(2\pi\sigma^2\right)^3}\,\rho_{ij}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{jk}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{ki}^{(2)}\left(\sqrt{3}\sigma\right)\end{aligned} \tag{12}$$

These correlations are captured by the square symmetric matrix $\mathbf{K}^{(32)}$ with entries given by $K_{jk}^{(32)} = \rho_{ij}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{jk}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{ki}^{(2)}\left(\sqrt{3}\sigma\right)\left(1-\delta_{jk}\right)$, where $j, k \in \{1, 2, 3, \cdots, N_i\}$. The vector $\mathbf{u}^{(32)}$ containing the N_i eigenvalues (sorted) of the symmetric $N_i \times N_i$ matrix $\mathbf{K}^{(32)}$ is a descriptor that incorporates the three-body correlations described by $\mathcal{C}^{(32)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right)$.

It is easy to see that matrices $\mathbf{K}^{(31)}$ and $\mathbf{K}^{(32)}$ are related to the two-body correlations described in Section II A. For example, $\mathbf{K}^{(31)}$ is related to $\mathbf{u}^{(1)}$ via an outer product, i.e. $\mathbf{K}^{(31)} = \left(\mathbf{u}^{(1)} \otimes \mathbf{u}^{(1)} - \mathbf{I}\right)$, where $\mathbf{I}$ is a $N_i \times N_i$ matrix with entries $I_{ij} = \delta_{ij}$. The three-body correlation involving the overlap between the densities of three atoms seems to be related to $\mathbf{K}^{(31)}$ and $\mathbf{K}^{(2)}$ via a Hadamard product: $\bar{\mathbf{K}}^{(32)} = \mathbf{K}^{(31)} \circ \mathbf{K}^{(2)} = \left(\mathbf{u}^{(1)} \otimes \mathbf{u}^{(1)} - \mathbf{I}\right) \circ \mathbf{K}^{(2)}$. But, the width of the two-body ($\sqrt{2}\sigma$) and three-body ($\sqrt{3}\sigma$) Gaussian overlap integrals are different meaning that correlations embedded in the matrices $\bar{\mathbf{K}}^{(32)}$ and $\mathbf{K}^{(32)}$ are different. For the results presented in Section III, we found no difference if

$\mathbf{u}^{(32)}$ was obtained from $\bar{\mathbf{K}}^{(32)}$ instead of $\mathbf{K}^{(32)}$.

Similar to the discussion in Section II A, the matrix $\mathbf{K}^{(32)}$ has properties of an adjacency matrix of a weighted graph G such that the vertices of G are the N_i neighbors and there is an edge joining each atom in the neighborhood of i to all the other neighbors of i (as shown in Fig. 2). Thus, G is a regular connected graph. The three-body correlation matrix $\mathbf{K}^{(32)}$ is similar to the adjacency matrix of G, except that the weight associated with the edge between two vertices with indices j, k in this case are given by $K_{jk}^{(32)} = \rho_{ij}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{ik}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{jk}^{(2)}\left(\sqrt{3}\sigma\right)(1-\delta_{jk})$. Since the two-body correlations $\rho_{ij}^{(2)}\left(\sqrt{2}\sigma\right)$, $\rho_{ik}^{(2)}\left(\sqrt{2}\sigma\right)$ and $\rho_{jk}^{(2)}\left(\sqrt{2}\sigma\right)$. depend on the L_2-norm of the distance between two atoms, these correlations are thus invariant to global translation and rotation operations. Similarly, the eigenvalues of $\mathbf{K}^{(31)}$ and $\mathbf{K}^{(32)}$ are invariant to the permutation of atomic indices of the neighbors.

We note here that the three-body correlation $\mathcal{C}^{(31)}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k)$ (illustrated in Fig. 3) corresponds to a star graph. The adjacency matrix of the three-node star graph in Fig. 3 (a) is

$$\mathbf{K}_{\text{adj}} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \tag{13}$$

with eigenvalues of $\{-\sqrt{2}, 0, \sqrt{2}\}$. In general, the $k \times k$ adjacency matrix of a star graph (with one node connected to $(k-1)$ other nodes) has $(k-2)$ zero eigenvalues and the two non-zero eigenvalues are $\pm\sqrt{k-1}$. Thus, the eigenvalues of the symmetric matrix $\mathbf{K}^{(31)}$ that embeds three-body correlations are different from the adjacency matrix of a star graph.

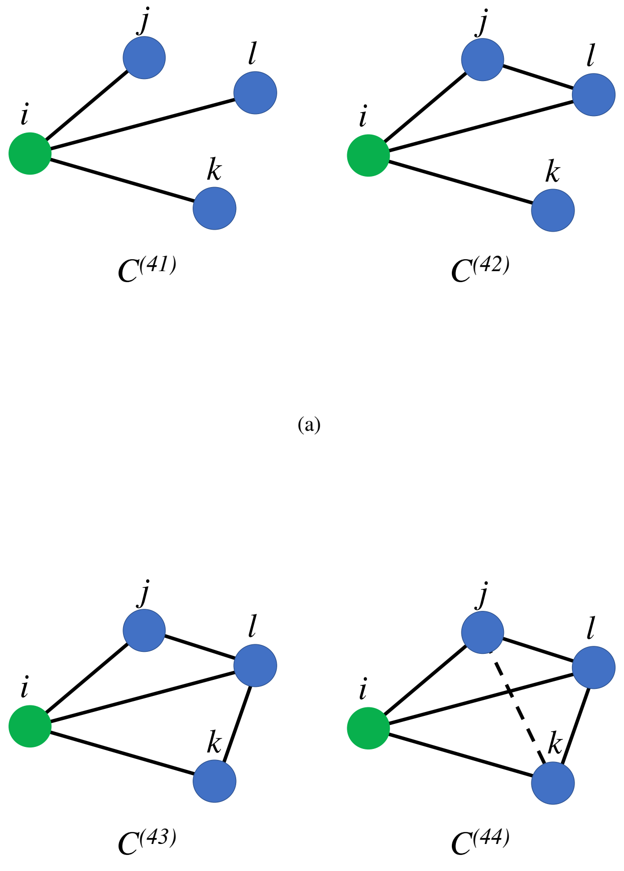
C. Four-body correlations

To obtain descriptors derived from four-body correlations, we consider four different contributions due to the overlap between local densities of three atoms in the neighbor-list of atom i. These contributions are detailed below:

v. The correlation due to the overlap between densities of pairs (i, j), (i, k), and (i, l) is given by

$$\begin{aligned}&\mathcal{C}^{(41)}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) \\ &= \frac{1}{8^3(\pi\sigma^2)^{9/2}}\rho_{ij}^{(2)}\left(\sqrt{2}\sigma\right)\rho_{ik}^{(2)}\left(\sqrt{2}\sigma\right)\rho_{il}^{(2)}\left(\sqrt{2}\sigma\right)\end{aligned} \tag{14}$$

Geometrically, this is similar to a star graph as shown in Fig. $4(a)$. To evaluate this four-body correlation, for each pair of atom (i, j) we define a $(N_i-1)\times(N_i-1)$ square symmetric matrix $\mathbf{K}^{(41)}$, such that $K_{kl}^{(41)} = \rho_{ik}^{(2)}\left(\sqrt{2}\sigma\right)\rho_{il}^{(2)}\left(\sqrt{2}\sigma\right)(1-\delta_{kl})(1-\delta_{jl})(1-\delta_{jk})$, where k, l = 1, 2, 3, ..., (N_i-1). Let $\mathbf{v}^{(41)}(k, l|i, j) = [\lambda_1^{(41)}\ \lambda_2^{(41)}\ \cdots\lambda_{(N_i-1)}^{(41)}]$ be a vector that contains the sorted eigenvalues of $\mathbf{K}^{(41)}$.



(a)

(b)

FIG. 4: Correlations due to the overlap between the atomic densities in a cluster containing 4 atoms.

The vector $\mathbf{u}^{(41)}$ defined by

$$\mathbf{u}^{(41)} = \sum_{j=1}^{N_i}\rho_{ij}^{(2)}\left(\sqrt{2}\sigma\right)\mathbf{v}^{(41)}(k, l|i, j) \tag{15}$$

is a descriptor that incorporates the four-body correlations in $\mathcal{C}^{(41)}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l)$.

vi. The correlation due to the presence of an atom k when densities of three other atoms i, j and l already have a finite non-zero overlap is given by

$$
\begin{aligned}
&\mathcal{C}^{(42)}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k,\mathbf{r}_l\right)\\
&=\frac{1}{\left(2\pi\sigma^2\right)^{15/2}}\left[\int_{-\infty}^{\infty} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_i\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_j\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_l\right|^2}{2\sigma^2}} d\mathbf{r}\right]\\
&\left[\int_{-\infty}^{\infty} e^{\frac{-\left|\mathbf{r}'-\mathbf{r}_i\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}'-\mathbf{r}_k\right|^2}{2\sigma^2}} d\mathbf{r}'\right]+\text{permutations} \qquad (16)\\
&=\frac{1}{192\sqrt{3}\,\left(\pi\sigma^2\right)^{9/2}}\,\rho_{ij}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{il}^{(2)}\left(\sqrt{3}\sigma\right)\\
&\times\ \rho_{jl}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{ik}^{(2)}\left(\sqrt{2}\sigma\right)+\cdots
\end{aligned}
$$

Geometrically, this correlation accounts for an extra edge between the two neighbors j and l of vertex i as shown in Fig. $4(a)$. Following a procedure that is similar to one described above in (v), for each pair of atom $(i,\ j)$ we define a $(N_i-1)\times(N_i-1)$ square symmetric matrix $\mathbf{K}^{(42)}$, such that $K_{kl}^{(42)} = \rho_{il}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{jl}^{(2)}\left(\sqrt{3}\sigma\right)\rho_{ik}^{(2)}\left(\sqrt{2}\sigma\right)(1-\delta_{kl})(1-\delta_{jl})(1-\delta_{jk})$, where $k,\ l$ = 1, 2, 3, ..., (N_i-1). Let $\mathbf{v}^{(42)}(k,l|i,j)=[\lambda_1^{(42)}\ \lambda_2^{(42)}\ \cdots\lambda_{(N_i-1)}^{(42)}]$ be a vector that contains the sorted eigenvalues of $\mathbf{K}^{(42)}$. The vector $\mathbf{u}^{(42)}$ defined by

$$
\mathbf{u}^{(42)}=\sum_{j=1}^{N_i}\rho_{ij}^{(2)}\left(\sqrt{3}\sigma\right)\mathbf{v}^{(42)}(k,l|i,j) \qquad (17)
$$

is a descriptor that incorporates the four-body correlations in $\mathcal{C}^{(42)}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k,\mathbf{r}_l\right)$.

vii. The correlation $\mathcal{C}^{(43)}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k,\mathbf{r}_l\right)$ arises when two pairs of neighbors of vertex i are connected. For example, in Fig. $4(b)$, two pairs of neighbors $(j,\ l)$ and $(k,\ l)$ are connected by extra edges that lead to two different three-body overlap integrals. Hence, this correlation is given by

$$
\begin{aligned}
&\mathcal{C}^{(43)}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k,\mathbf{r}_l\right)\\
&=\frac{1}{\left(2\pi\sigma^2\right)^{9}}\left[\int_{-\infty}^{\infty} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_i\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_k\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_l\right|^2}{2\sigma^2}} d\mathbf{r}\right]\\
&\times\left[\int_{-\infty}^{\infty} e^{\frac{-\left|\mathbf{r}'-\mathbf{r}_i\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}'-\mathbf{r}_j\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}'-\mathbf{r}_l\right|^2}{2\sigma^2}} d\mathbf{r}'\right] \qquad (18)\\
&\quad+\text{permutations}
\end{aligned}
$$

Following (vi) and (vii), we defined a vector $\mathbf{u}^{(43)}$ that captures this correlation.

viii. The correlation, $\mathcal{C}^{(44)}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k,\mathbf{r}_l\right)$, arising due to the overlap between the densities of all four atoms i, j, k, l, as shown in Fig. $4(b)$, is given by

$$
\begin{aligned}
&\mathcal{C}^{(44)}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k,\mathbf{r}_l\right)\\
&=\frac{1}{\left(2\pi\sigma^2\right)^{6}}\int_{-\infty}^{\infty} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_i\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_j\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_k\right|^2}{2\sigma^2}} e^{\frac{-\left|\mathbf{r}-\mathbf{r}_l\right|^2}{2\sigma^2}} d\mathbf{r}\\
&=\frac{1}{8\left(2\pi\sigma^2\right)^{9/2}}\,\rho_{ij}^{(2)}(2\sigma)\,\rho_{ik}^{(2)}(2\sigma)\,\rho_{il}^{(2)}(2\sigma)\\
&\times\ \rho_{jk}^{(2)}(2\sigma)\,\rho_{jl}^{(2)}(2\sigma)\,\rho_{kl}^{(2)}(2\sigma)
\end{aligned}
\qquad (19)
$$

Similar to $\mathcal{C}^{(32)}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k,\mathbf{r}_l\right)$, this correlation captures closed paths and corresponds to a connected regular graph. For each pair of atom $(i,\ j)$, let $\mathbf{v}^{(44)}(k,l|i,j) = [\lambda_1^{(44)}\ \lambda_2^{(44)}\ \cdots\lambda_{(N_i-1)}^{(44)}]$ be a vector that contains the sorted eigenvalues of the $(N_i-1)\times(N_i-1)$ symmetric matrix $\mathbf{K}^{(44)}$, such that $K_{kl}^{(44)} = \rho_{ik}^{(2)}(2\sigma)\,\rho_{il}^{(2)}(2\sigma)\,\rho_{jk}^{(2)}(2\sigma)\,\rho_{jl}^{(2)}(2\sigma)\,\rho_{kl}^{(2)}(2\sigma)\,(1-\delta_{kl})(1-\delta_{jl})(1-\delta_{jk})$, where $k,\ l$ = 1, 2, 3, ..., (N_i-1). The vector $\mathbf{u}^{(44)}$ defined by

$$
\mathbf{u}^{(44)}=\sum_{j=1}^{N_i}\rho_{ij}^{(2)}(2\sigma)\,\mathbf{v}^{(44)}(k,l|i,j) \qquad (20)
$$

is a descriptor that incorporates the four-body correlations in $\mathcal{C}^{(44)}\left(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k,\mathbf{r}_l\right)$.

D. Higher order correlations

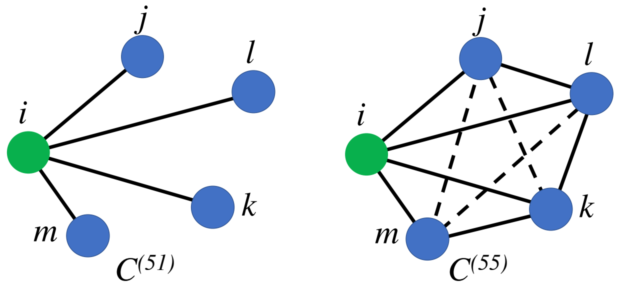


FIG. 5: Correlations due to the overlap between the atomic densities in a cluster containing 5 atoms.

In principle, the procedure outlined above can be extended to calculate five-body and even higher order density correlations. Five-body correlations account for interactions of an atom with four neighbors some of which are shown in Fig. 5. However, for the examples studied in Section III, correlations that arise due to two-, three- and four-body density overlaps are sufficient to capture intrinsic details of the local geometry. Thus, the contributions of five-body or higher order correlations are negligible for the systems considered in this study.

III. NUMERICAL RESULTS: RECONSTRUCTION OF SMALL CLUSTERS

A. Simulation setup and computational details

The ability of the many-body correlation descriptors to capture intrinsic geometric features of local atomic environments was tested by performing reconstruction simulations. To this end, we used 8 clusters containing $\{10, 14, 20, 25\}$ Ge atoms and $\{10, 15, 20, 25\}$ Cu atoms. These structures were generated from bulk supercells containing 64 Ge and 256 Cu atoms by multiple surface cuts. For reconstruction simulations, atomic positions in these clusters were perturbed by using random numbers uniformly distributed in the interval (-1, 1) Å and sets containing 200 randomly perturbed structures were generated. These perturbed structures were then used as starting configurations for reconstruction simulations. The goal was to generate the unperturbed configuration (denoted by $\mathbf{X}$) using the set of vectors $\mathbf{u} = \{\, \mathbf{u}^{(1)},\, \mathbf{u}^{(2)},\, \mathbf{u}^{(31)},\, \mathbf{u}^{(32)},\, \mathbf{u}^{(41)},\, \mathbf{u}^{(42)},\, \mathbf{u}^{(43)},\, \mathbf{u}^{(44)},\, \mathbf{u}^{(51)}\, \}$.

We minimized the cost function, Ω, to recover the target structures, where

$$\begin{aligned}
&\Omega\left(\mathbf{X}, \mathbf{X}'\right) \\
&= \sum_{i=1}^{N} \left\{ \left[1 - \left(\hat{\mathbf{u}}^{(1)}_{i_{\mathbf{X}}},\, \hat{\mathbf{u}}^{(1)}_{i_{\mathbf{X}'}}\right)^{m}\right] + \left[1 - \left(\hat{\mathbf{u}}^{(2)}_{i_{\mathbf{X}}},\, \hat{\mathbf{u}}^{(2)}_{i_{\mathbf{X}'}}\right)^{m}\right] \right. \\
&\quad + \left[1 - \left(\hat{\mathbf{u}}^{(31)}_{i_{\mathbf{X}}},\, \hat{\mathbf{u}}^{(31)}_{i_{\mathbf{X}'}}\right)^{m}\right] + \left[1 - \left(\hat{\mathbf{u}}^{(32)}_{i_{\mathbf{X}}},\, \hat{\mathbf{u}}^{(32)}_{i_{\mathbf{X}'}}\right)^{m}\right] \\
&\quad \left. + \left[1 - \left(\hat{\mathbf{u}}^{(41)}_{i_{\mathbf{X}}},\, \hat{\mathbf{u}}^{(41)}_{i_{\mathbf{X}'}}\right)^{m}\right] + \cdots \right\}.
\end{aligned} \tag{21}$$

Here, $\mathbf{X}$ is the reference or the target structure, $\mathbf{X}'$ is the reconstructed structure, $\hat{\mathbf{u}}^{(\cdot)} = \mathbf{u}^{(\cdot)}/\left|\mathbf{u}^{(\cdot)}\right|$, m = 8, N is the total number of atoms in $\mathbf{X}$, $\mathbf{X}'$, the vectors $\mathbf{u}^{(2)}_{i_{\mathbf{X}}}$, $\mathbf{u}^{(2)}_{i_{\mathbf{X}'}}$, $\mathbf{u}^{(31)}_{i_{\mathbf{X}}}$, $\mathbf{u}^{(31)}_{i_{\mathbf{X}'}}$, etc. are descriptors that encode different correlations. The sum in the above equation is over all atoms in the two structures, $\mathbf{X}$ and $\mathbf{X}'$, and $(\mathbf{u}, \mathbf{v})$ denotes a scalar product. The cost function was optimized using the Nelder-Mead simplex search method proposed by Lagarias et al.[47] as implemented in Matlab. The width, σ, of the single particle density in Eq. 9 was set to 3, 3.5 Å (midway between the first and the second peaks of the pair correlation function) for Cu and Ge clusters, respectively. While the cost function can be further minimized (using cross-validation) with respect to this hyper-parameter, we were able to obtain satisfactory results with these values of σ.

Following Ref. [9], reconstructed structures (denoted by $\mathbf{X}'$) were compared to the target structure, $\mathbf{X}$, using the metric $d^{W}_{\mathbf{X},\mathbf{X}'}$ which compares Weyl matrices of these two structures. The Weyl matrix, Σ, used here is invariant to global translation, rotation and reflection, and takes the following form:

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \mathbf{r}_1 \cdot \mathbf{r}_1 & \mathbf{r}_1 \cdot \mathbf{r}_2 & \mathbf{r}_1 \cdot \mathbf{r}_3 & \cdots & \mathbf{r}_1 \cdot \mathbf{r}_N \\ \mathbf{r}_2 \cdot \mathbf{r}_1 & \mathbf{r}_2 \cdot \mathbf{r}_2 & \mathbf{r}_2 \cdot \mathbf{r}_3 & \cdots & \mathbf{r}_2 \cdot \mathbf{r}_N \\ \mathbf{r}_3 \cdot \mathbf{r}_1 & \mathbf{r}_3 \cdot \mathbf{r}_2 & \mathbf{r}_3 \cdot \mathbf{r}_3 & \cdots & \mathbf{r}_3 \cdot \mathbf{r}_N \\ & & \cdots & & \\ \mathbf{r}_N \cdot \mathbf{r}_1 & \mathbf{r}_N \cdot \mathbf{r}_2 & \mathbf{r}_N \cdot \mathbf{r}_3 & \cdots & \mathbf{r}_N \cdot \mathbf{r}_N \end{bmatrix} \tag{22}$$

The metric $d^{W}_{\mathbf{X},\mathbf{X}'}$ which measures the distance between two Weyl matrices was calculated from

$$d^{W}_{\mathbf{X},\mathbf{X}'} = \left[\sum_{i=1}^{N} \sum_{j=1}^{N} \left(\left(\Sigma_{\mathbf{X}}\right)_{ij} - \left(\Sigma_{\mathbf{X}'}\right)_{ij} \right)^{8} \right]^{1/8} . \tag{23}$$

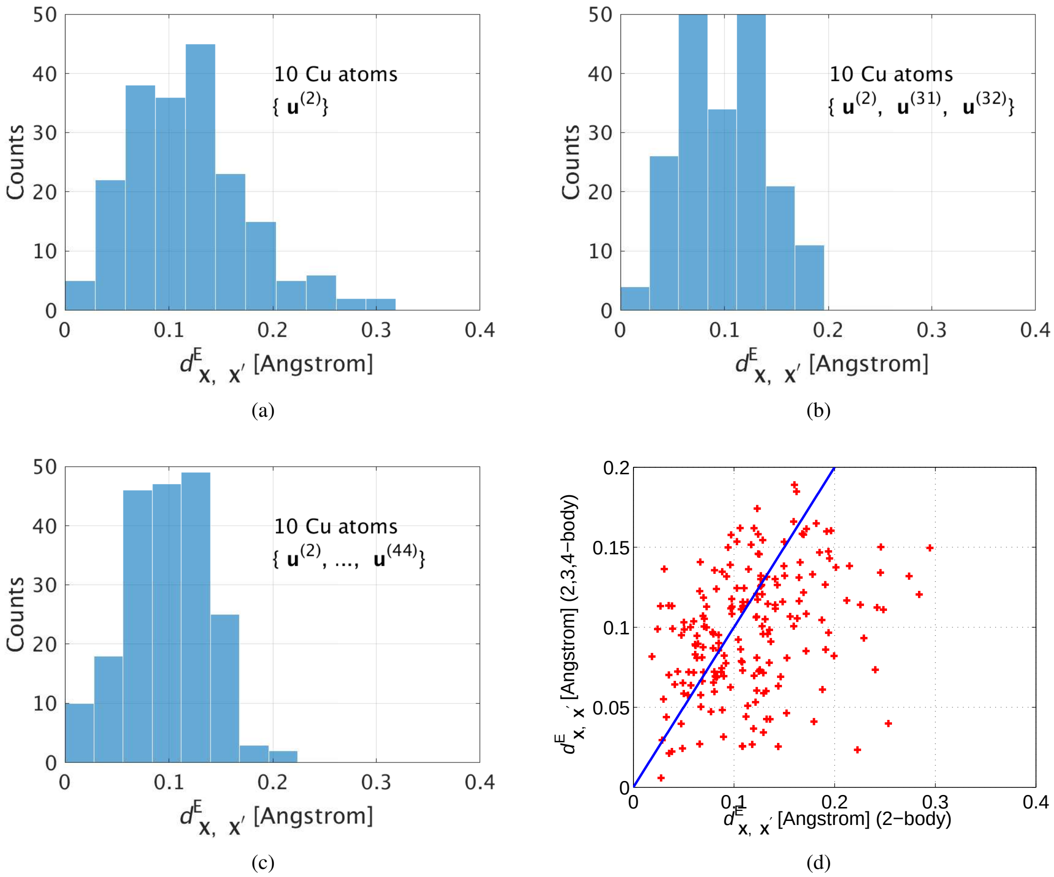
Reconstruction simulations were terminated when $d^{W}_{\mathbf{X},\mathbf{X}'} < 1 \times 10^{-3}$ Å^2. The quality of reconstructed structures was also monitored using the root mean squared deviation of each atomic degree of freedom from the target, $\mathbf{X}$:

$$d^{E}_{\mathbf{X},\mathbf{X}'} = \left[\frac{1}{3N} \sum_{i=1}^{N} \sum_{j=1}^{3} \left(r^{j}_{i_{\mathbf{X}}} - r^{j}_{i_{\mathbf{X}'}} \right)^{2} \right]^{1/2} . \tag{24}$$

For clusters containing less than 20 atoms, reconstructions simulations typically required 15000-20000 function evaluations to converge. On the other hand, structures containing 20 and 25 Cu or Ge atoms typically required 40000-50000 function evaluations to converge.

B. Results

Figure 6 shows the effect of two-body, three-body and four-body correlations on the quality of reconstructed structures containing 10 Cu atoms. The distribution of $d^{E}_{\mathbf{X},\mathbf{X}'}$ in Fig. 6(a) lies between 0 and 0.3 Å (mean and standard deviations are 0.118 and 0.057 Å, respectively) when only $\mathbf{u}^{(2)}$ descriptor was used for the 200 reconstruction simulations. The distributions of $d^{E}_{\mathbf{X},\mathbf{X}'}$ in Figs. 6(b) and 6(c) correspond to reconstruction simulations performed using $\{\mathbf{u}^{(2)}, \mathbf{u}^{(31)}, \mathbf{u}^{(32)}\}$ and $\{\mathbf{u}^{(2)}, \mathbf{u}^{(31)}, \mathbf{u}^{(32)}, \mathbf{u}^{(41)}, \mathbf{u}^{(42)}, \mathbf{u}^{(43)}, \mathbf{u}^{(44)}\}$, respectively. Using both two-body and three-body correlations decreased the mean and standard deviations of $d^{E}_{\mathbf{X},\mathbf{X}'}$ to 0.100 and 0.039 Å, respectively. Similarly, using two-body, three-body and four-body correlations pushed the distribution further left, i.e. the mean of $d^{E}_{\mathbf{X},\mathbf{X}'}$ from all the 200 reconstructed structures was 0.098 Å while the standard deviation remained unchanged. This is evident from the fact that the maximum value of $d^{E}_{\mathbf{X},\mathbf{X}'}$ decreased from ~0.30Å (when only $\mathbf{u}^{(2)}$ was used) to 0.20 Å when three-body and four-body descriptors were also used (see Fig. 6(b) and 6(c)). This is also evident from Fig. 6(d) which compares the values of $d^{E}_{\mathbf{X},\mathbf{X}'}$ when reconstruction simulations were performed using



(a)

(b)

(c)

(d)

FIG. 6: The quality of reconstruction improved when descriptors corresponding to three-body and four-body correlations were included. Shown here is the distribution of the metric $\mathbf{d}^E_{\mathbf{X},\mathbf{X}'}$ for the 200 reconstructed structures containing 10 Cu atoms. The results in Fig. $6(a)$ correspond to reconstructions performed using only $\mathbf{u}^{(2)}$, while results in Figs. $6(b)$ and $6(c)$ correspond to reconstruction simulations performed using $\mathbf{u} = \{\mathbf{u}^{(2)}, \mathbf{u}^{(31)}, \mathbf{u}^{(32)}\}$, $\mathbf{u} = \{\mathbf{u}^{(2)}, \mathbf{u}^{(31)}, \mathbf{u}^{(32)}, \mathbf{u}^{(41)}, \mathbf{u}^{(42)}, \mathbf{u}^{(43)}, \mathbf{u}^{(44)}\}$, respectively. Fig. $6(d)$ compares the values of $\mathbf{d}^E_{\mathbf{X},\mathbf{X}'}$ when reconstruction simulations were performed using only $\mathbf{u}^{(2)}$ to those obtained using two, three and four-body correlations starting from the same set of configurations.

only $\mathbf{u}^{(2)}$ to those obtained using all two-body, three-body and four-body correlations. Thus, the quality of the reconstructed structures improved when three-body and four-body correlation descriptors were included for the reconstruction simulations. Similar conclusions were obtained from reconstruction simulations performed for structures containing 10 Ge. But, Figs. $7(a)$ and $7(b)$ show that the mean and standard deviation of $d^E_{\mathbf{X},\mathbf{X}'}$, for 200 reconstructed structures, each containing 14 Ge atoms, increased slightly from (0.102, 0.043) Å for $\mathbf{u}^{(2)}$ to (0.110, 0.046) Å for $\{\mathbf{u}^{(2)}, \mathbf{u}^{(31)}, \mathbf{u}^{(32)}\}$. This suggests that the relevance of different many-body correlations depends on the local structure and the system size. Fig. $7(c)$ shows that the quality of the reconstructed structures improved when $\mathbf{u}^{(1)}$ was also included along with the two and three-body correlations (the mean and standard deviations of $d^E_{\mathbf{X},\mathbf{X}'}$ were 0.096 and 0.037 Å, respectively). The improvement in quality when three-body correlations were used along with $\mathbf{u}^{(2)}$ for the reconstruction of clusters containing 15 Cu atoms is also evident in Figs. $8(a)$ and $8(b)$.

We note here that changes in the mean and standard deviation of $d^E_{\mathbf{X},\mathbf{X}'}$ were negligible when the five-body correlation descriptor $\mathbf{u}^{(51)}$ was also used along with two-body, three-body and four-body descriptors. This is consistent with the observation of only marginal decrease in the standard deviation of $d^E_{\mathbf{X},\mathbf{X}'}$ when four-body correlations were incorporated in addition to the two- and three-body correlations as is evident from Figs. $6(b)$ and $6(c)$ ($d^E_{\mathbf{X},\mathbf{X}'}$ decreased from 0.100 to 0.098). This suggests that two-body and three-body correlations are vital to obtain good quality reconstructions, while higher order correlations only marginally effect the quality of reconstruction.

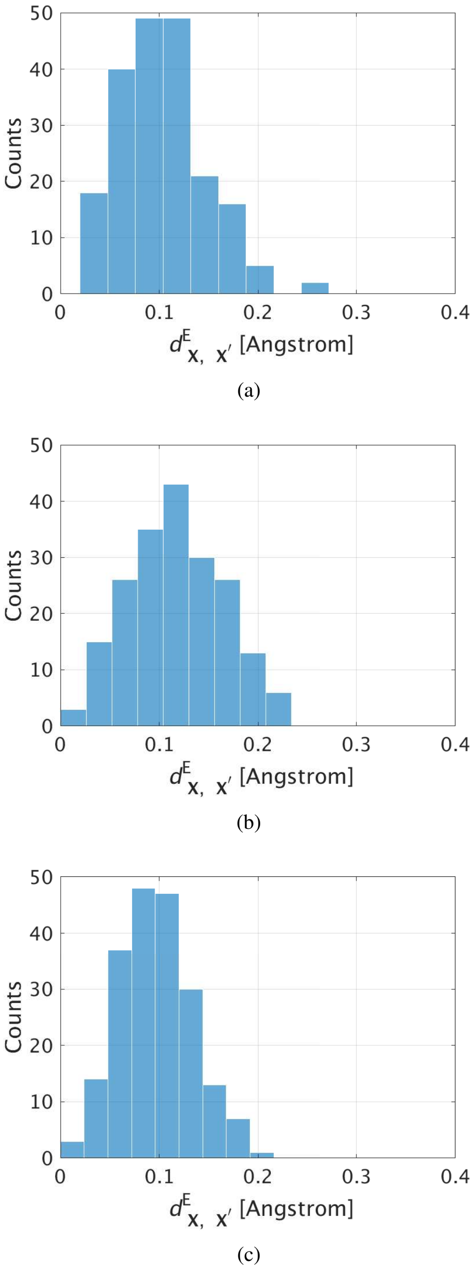
Eigenvalues of graph Laplacians, which are related to eigenvalues of adjacency matrices (which are similar in form to $\mathbf{K}^{(2)}$) via a normalization, are routinely used to identify intrinsic dimensionality. Since these methods are so popular, we analyzed the effect of system size on the quality of reconstruction by using only $\mathbf{u}^{(2)}$. With increase in system size, the number of nearest neighbors also increases meaning that information about the second, and third nearest-neighbors are now incorporated into the descriptors. Thus, as the size of a cluster increases, the number of entries in each of the descriptors $\mathbf{u}^{(2)}$, $\mathbf{u}^{(31)}$, $\mathbf{u}^{(32)}$ also increases, and the ensuing improvement in the quality of reconstruction is evident from Fig. 8. The mean and the standard deviation of the metric $d^E_{\mathbf{X},\mathbf{X}'}$ for the reconstructed structures decreased from (0.104, 0.047) Å for 15 atom clusters to (0.083, 0.038) Å for 20 atom clusters, and to (0.080, 0.034) Å for clusters containing 25 Cu atoms. This means that in bulk systems when the cutoff radius is large, $\mathbf{u}^{(2)}$ alone can perhaps capture the important structural features.

We also analyzed the effect of $\mathbf{u}^{(1)}$ on the quality of the reconstructed structures. We found that a large fraction of the reconstruction simulations performed using only $\mathbf{u}^{(1)}$ converged to metastable configurations that were different from the global minimum structure (i.e. the target structure). This means that the landscape of corresponding cost function,

$$\Omega\left(\mathbf{X},\mathbf{X}'\right) = \sum_{i=1}^{N}\left[1-\left(\hat{\mathbf{u}}^{(1)}_{i\mathbf{X}},\ \hat{\mathbf{u}}^{(1)}_{i\mathbf{X}'}\right)^{m}\right], \qquad (25)$$

is rugged. Interestingly, when combined $\mathbf{u}^{(1)}$ was combined with the descriptor $\mathbf{u}^{(2)}$, reconstruction simulations typically needed fewer function evaluations than required when only $\mathbf{u}^{(2)}$ was used. Similarly, when $\mathbf{u}^{(1)}$ was used together with $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(31)}$ the quality of reconstruction showed marginal improvement (compared to using only $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(31)}$, in Fig. 7) for both Cu and Ge clusters, but the rate of convergence improved significantly. We note here that characteristics of descriptors $\mathbf{u}^{(2)}$, $\mathbf{u}^{(31)}$, $\mathbf{u}^{(32)}$ obtained from correlation matrices are different from the behavior of $\mathbf{u}^{(1)}$: $\mathbf{u}^{(1)}$ simply contains the overlap between an atom and its neighbors. Thus, $\mathbf{u}^{(1)}$ is very sensitive to small changes in bond lengths and the sum of all entries in $\mathbf{u}^{(1)}$ or the norm of $\mathbf{u}^{(1)}$ can easily distinguish an atom in the bulk from an atom on a surface, or an atom at the core of a dislocation. On the other hand, $\mathbf{u}^{(2)}$, $\mathbf{u}^{(31)}$, $\mathbf{u}^{(32)}$ capture collective behavior and contain important geometric information (such as the degree and connectivity, closed paths) about the distribution of neighbors of an atom. Thus, the improved rate of convergence when $\mathbf{u}^{(1)}$ was included with both the two-body and the three-body descriptors is a manifestation of inclusion of this local as well as non-local (collective) structural information.

The effect of the two different three-body correlations on the quality of reconstruction was analyzed by performing two sets of reconstructions using $\mathbf{u} = \{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(31)}\}$ and $\mathbf{u} = \{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(31)}, \mathbf{u}^{(32)}\}$ as descriptors. Figure 3 shows that the three-body correlation (i.e. $\mathcal{C}^{(32)}$ and $\mathbf{u}^{(32)}$) corresponding to closed paths are already included in $\mathbf{u}^{(31)}$, meaning that when two neighbors (of atom i) with indices j and k are close, they both contribute to $\mathbf{u}^{(31)}$ and $\mathbf{u}^{(32)}$. If the atoms are far apart, they contribute to $\mathbf{u}^{(31)}$ but not to $\mathbf{u}^{(32)}$. To analyze this, we used clusters containing 20 and 25 Cu atoms and performed reconstruction simulations using $\{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(31)}\}$ and $\{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(31)}, \mathbf{u}^{(32)}\}$. The results remained inconclusive for clusters with 20 Cu atoms (i.e. the mean and standard deviations were close) meaning that the inclusion of $\mathbf{u}^{(32)}$ did not result in improved in accuracy. However, for clusters containing 25 Cu atoms, the quality of the reconstructed structures showed a marginal improvement when both $\mathbf{u}^{(31)}$ and $\mathbf{u}^{(32)}$ were used along with the two-body correlation descriptor. This suggests that both the three-body correlations are important. These results also strengthen our ar-



(a)

(b)

(c)

FIG. 7: Shown here is the distribution of the metric $\mathbf{d}^E_{\mathbf{X},\mathbf{X}'}$ for the 200 reconstructed structures containing 14 Ge atoms. The results in Figs. $7(a)$ and $7(b)$ correspond to reconstructions performed using $\mathbf{u}^{(2)}$ and $\{\mathbf{u}^{(2)}, \mathbf{u}^{(31)}, \mathbf{u}^{(32)}\}$, respectively. Fig. $7(c)$ shows that the quality of the reconstructed structures improved when $\mathbf{u}^{1}$ was also included along with the two and three-body correlations.

gument that two and three-body correlations are sufficient to obtain high quality of reconstruction for Cu clusters.

Finally, we verified that the spectra of only two-body and three-body overlap matrices are sufficient to reconstruct clusters containing 20 and 25 Ge atoms. To this end, we used the descriptors $\mathbf{u} = \{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(31)}, \mathbf{u}^{(32)}\}$. Figures $9(b)$ and $9(d)$ show the distribution of $d^E_{\mathbf{X},\mathbf{X}'}$ of the final reconstructed structures clusters containing 20 and 25 Ge atoms, respectively. The distribution of initial structures used for these reconstructed are shown in the two-dimensional space of $d^W_{\mathbf{X},\mathbf{X}'}$-$d^E_{\mathbf{X},\mathbf{X}'}$ in Figs. $9(a)$ and $9(c)$. The fact that the reconstructed structures satisfy the convergence condition $d^W_{\mathbf{X},\mathbf{X}'} < 1 \times 10^{-3}$ Å^2 clearly shows that higher order correlations are not required for this system.

IV. DISCUSSIONS

A. Uniqueness of reconstruction

It is known that the spectrum of the adjacency matrix of a graph is not unique,[48] meaning that the spectrum of the adjacency matrix cannot be used to uniquely reconstruct the original graph from a perturbed one.[48] However, we were able to successfully reconstruct structures of clusters because we used descriptors for each atom in the cluster. In addition, we note here that neighborhood information of atoms in these clusters had considerable overlap. For example, for the cluster shown in Fig. 2, the central atom (shown in blue) has eight neighbors within a distance of $\sqrt{3}a/2$, where a is the length of the cube. The atom marked 1 is a neighbor of the central atom and its nearest neighbors are the central atom and atoms marked 2, 4 and 5 (distance from atom 1 is a). Thus, our reconstruction results suggest that spectra of adjacency matrices of such overlapping sub-graphs of a graph can be used to uniquely reconstruct the graph.

B. Reconstruction with a truncated spectrum

The many-body correlation descriptors presented here can capture the overlap between the densities of neighboring atoms, yet by incorporating local proximity information between different atoms (or nodes when the atoms are considered to be located at the vertices of a regular connected graph), they provide a global description of a high dimensional system. Such a procedure is routinely used to find meaningful geometric description in clustering and dimension reduction techniques. In this context, the top few eigenvalues of the adjacency matrix of a graph embed important information about the degree and connectivity of the graph and such properties are relevant for reconstruction of clusters. Thus, it is intuitive to ask if all the eigenvalues are required for successful structural reconstruction simulations. To investigate this, we used two clusters containing 14 and 25 Ge atoms. Two sets, each containing 100 structures, were obtained by perturbing the atomic positions in these clusters. For the reconstructions reported in Section III B, for a cluster with N = 14 Ge atoms, information about the neighborhood of each atom was captured using $(N - 1)$ = 13 eigenvalues of $\mathbf{K}^{(2)}$, $\mathbf{K}^{(31)}$ and $\mathbf{K}^{(32)}$. We were, however, able to successfully reconstruct the structures by removing as many as 40% of the small eigenvalues of the correlation matrices. This suggests that structural reconstruction can be performed using only the top eigenvalues of the two-body and three-body correlation descriptors.

C. Comparison with existing models

The results in Section III show that an accurate representation of the local environment can be achieved by systematically incorporating higher order correlations. Machine learned potentials within the Gaussian approximation potential (GAP) framework incorporated two and three-body terms using the power-spectrum and bi-spectrum descriptors. For example, the local environment of an atom i in GAP is characterized by the atomic density defined as

$$\begin{aligned}\rho_i(\mathbf{r}) &= \sum_j e^{-|\mathbf{r}-\mathbf{r}_{ij}|^2/2\sigma^2} \\ &= \sum_{n,l,m} c_{nlm} g_n(|\mathbf{r}|) Y_{lm}(\mathbf{r})\end{aligned} \tag{26}$$

and the descriptor vector is given by $\mathbf{q} = \sum_m c^*_{nlm} c_{n'lm}$. Since, g_n and Y_{lm} are orthogonal basis sets, the product $c^*_{nlm} c_{n'lm}$ includes two-body correlations like $g_n(|\mathbf{r}_{ij}|)\, g_{n'}(|\mathbf{r}_{ik}|)$, where j and k are indices of two neighbors of atom i. Similarly, bi-spectrum components include three-body correlations like $g_n(|\mathbf{r}_{ij}|)\, g_{n'}(|\mathbf{r}_{ik}|)\, g_n(|\mathbf{r}_{il}|)$. In this study, instead of the sum of all many-body correlations, different two-body, three-body and four-body correlation information were used to obtain correlation matrices. The spectra of such correlation matrices were shown to embed structural information about the local neighborhood of an atom.

The spectral neighbor analysis potential (SNAP)[7] framework, is another MLP in which the energy of an atom i is expressed as

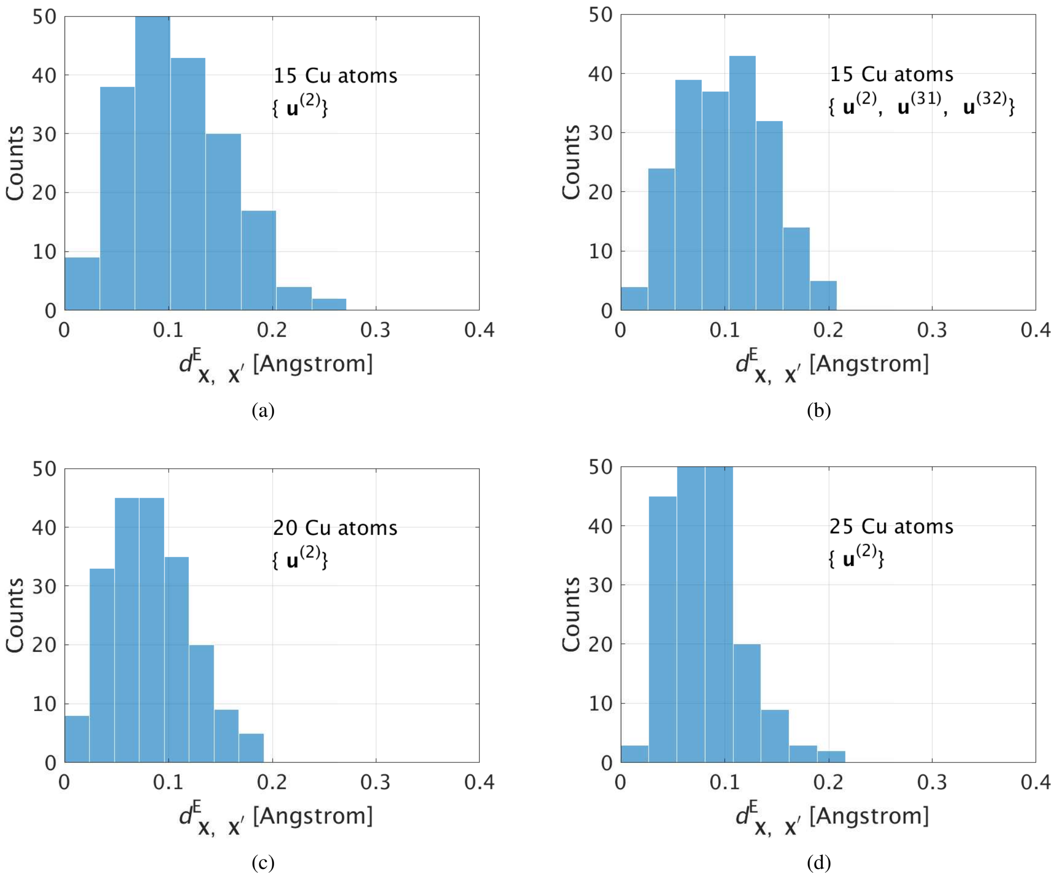
$$E_i = \beta_0 + \sum_p \beta_p B^i_p, \tag{27}$$

where B^i_p are the bi-spectrum components and β_p are the coefficients of this expansion. Recently, it was shown that the accuracy of this MLP can be increased (an order of magnitude decrease in training error was reported) by incorporating a second order term,[8] i.e.

$$E_i = \beta_0 + \sum_p \beta_p B^i_p + \sum_{p,q} \alpha_{pq} B^i_p B^i_q. \tag{28}$$

Here α_{pq} are the coefficients. This improvement in accuracy due to the presence of a second order term is in agreement with our conclusion that incorporating higher order correlations results in better learning of the local neighborhood.

Another machine learning based interatomic potential that



(a) (b) (c) (d)

FIG. 8: Quality of reconstruction, when using only $\mathbf{u}^{(2)}$, improved when the system size was increased. Shown here is the distribution of the metric $\mathbf{d}^{E}_{\mathbf{X},\mathbf{X}'}$ for the 200 reconstructed structures containing 15 (Figs. $8(a)$, $8(b)$), 20 (Fig. $8(c)$) and 25 (Fig. $8(d)$) Cu atoms. A systematic improvement in reconstruction of clusters containing 15 Cu atoms was also observed when both two-body and three-body density correlations were used.

incorporates local correlations was proposed by Lindsey, et al.[24,25] In this model, the energy of each atom was expanded using a basis of Chebyshev functions:

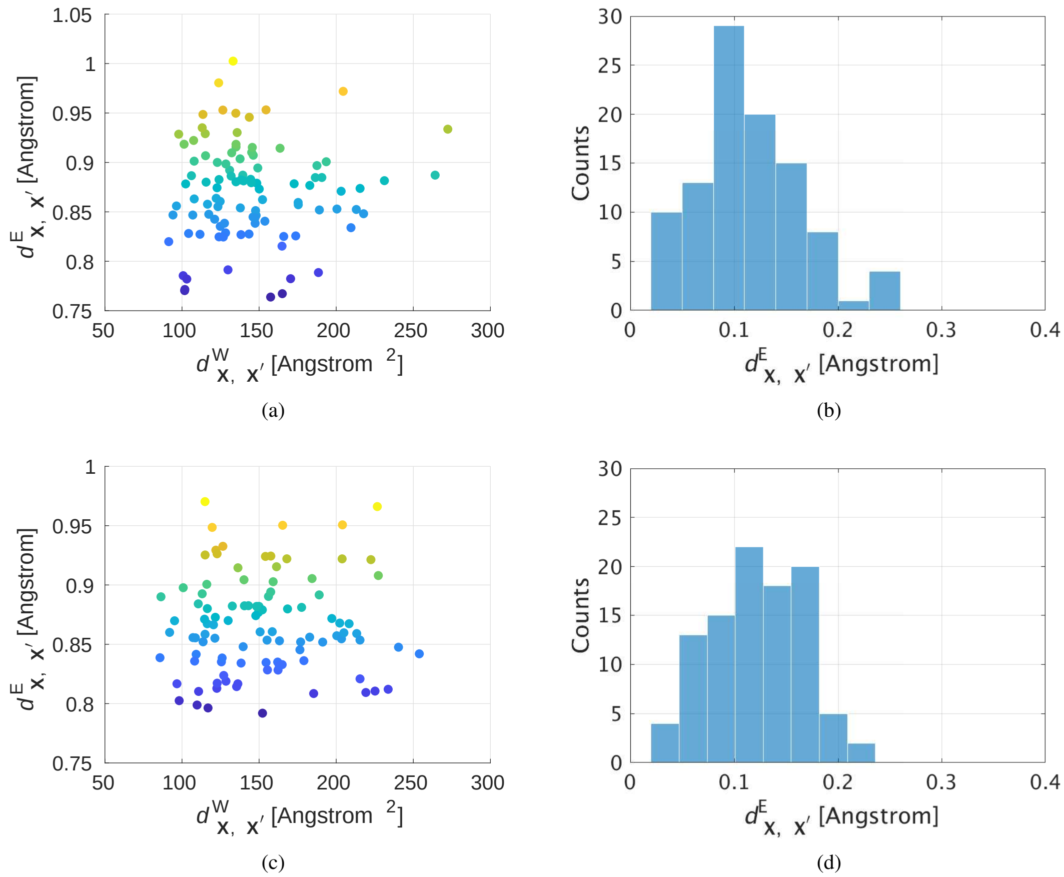
$$
\begin{aligned}
E_i &= \sum_j E_{ij} + \sum_{j,k} E_{ijk}, \text{ where} \\
E_{ij} &= \sum_p c_p T_p\left(s_{ij}\right), \text{ and} \\
E_{ijk} &= \sum_p \sum_q \sum_r c_{pqr} T_r\left(s_{ij}\right) T_q\left(s_{ik}\right) T_r\left(s_{jk}\right).
\end{aligned}
\tag{29}
$$

In the above summation, $T_n\left(s_{ij}\right)$ is a Chebyshev polynomial (of order n) of the first kind, $s_{ij} \in [-1,1]$ is a transformed nearest neighbor distance, and c_p, c_{pqr} are coefficients of the two and three-body terms. While E_{ij} incorporates two-body correlations, E_{ijk} incorporates correlations described by $\mathcal{C}^{(32)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right)$. Based on our results (see Sections II B), we hypothesize that the accuracy and the transferability of this MLP can be improved by including a term that incorporates correlations described by $\mathcal{C}^{(31)}\left(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\right)$:

$$
\begin{aligned}
E_i &= \sum_j E_{ij} + \sum_{j,k} \left(\bar{E}_{ijk} + E_{ijk}\right), \text{ where} \\
\bar{E}_{ijk} &= \sum_p \sum_q c_{pq} T_p\left(s_{ij}\right) T_q\left(s_{ik}\right).
\end{aligned}
\tag{30}
$$

Here, c_{pq} are coefficients of expansion, and E_{ij}, E_{ijk} are the two and three-body terms described in Eq. 29. Similarly, higher order correlations can be incorporated by following the strategy discussed in Section II C. In general, the Chebyshev basis functions can be replaced by any other basis functions, such as Bessel functions, Neumann functions, Morlet wavelets, Slater or Gaussian type orbitals, that have been used recently in the literature to generate MLP.[5,6]

Machine learned potentials based on neural networks that have been proposed in the literature have used symmetry functions that included two-body and three-body correlation.[1,28–30,49] Even though such correlation functions were not based on an expansion using a basis set, we hypothesize that the accuracy and transferability can be enhanced



(a)

(b)

(c)

(d)

FIG. 9: Figs. $9(a)$ and $9(c)$ show the distribution of 100 initial structures, containing 20 and 25 Ge atoms, respectively, obtained by perturbing the target structure (using uniformly random numbers in the range of [-1, 1]) in the two-dimensional space of $d^W_{\mathbf{X},\mathbf{X}'}$ and $d^E_{\mathbf{X},\mathbf{X}'}$. The values of $d^E_{\mathbf{X},\mathbf{X}'}$ for the final reconstructed structures are shown in Figs. $9(b)$ (clusters containing 20 Ge atoms) and $9(d)$ (clusters containing 25 Ge atoms).

by incorporating higher order correlations similar to those described in Eqs. 28 and 30.

V. SUMMARY

We have developed a framework to systematically generate descriptors based on many-body correlations that can effectively capture intrinsic geometric features of the local environment of an atom. These descriptors were obtained from the eigenvalues of two-body, three-body, four-body and higher order correlations between an atom and its neighbors. These many-body correlations correspond to two-body, three-body, four-body, etc. overlap integrals which were solved analytically by representing the density of an isolated atom by a Gaussian function. These descriptors are invariant to global translation, global rotation, reflection as well as permutations of atomic indices.

The ability of these descriptors to describe local environments was systematically tested by reconstructing structures of clusters containing 10 to 25 atoms. Our results suggest that two and three-body correlations, i.e. $\mathbf{u}^{(2)}$, $\mathbf{u}^{(31)}$ and $\mathbf{u}^{(32)}$, are important for successful reconstruction of the original structure from a perturbed one. These descriptors capture the collective behavior of constituent atoms and contain important geometric information (such as the degree and connectivity, closed paths) about the distribution of neighbors of an atom.

The two-body descriptor, $\mathbf{u}^{(1)}$, is sensitive to small changes in bond length and was found to increase the rate of convergence of reconstruction simulations. This improved rate of convergence when $\mathbf{u}^{(1)}$ was included with both the two-body and the three-body descriptors is a manifestation of inclusion of local as well as non-local (collective) structural information.

We also found that both the three-body descriptors, i.e. $\mathbf{u}^{(31)}$ and $\mathbf{u}^{(32)}$, are important. In general, the quality of reconstruction can be further improved by including four-body correlations, but for the systems considered in this study this lead to only a marginal improvement in the quality of the reconstructed structures. Thus, higher order correlations are not so important as the two and three-body correlations. But, this framework allows us to systematically select the number of descriptors (i.e. the type of correlations used) based on the trade off between accuracy, efficiency and computational cost.

The ability of these descriptors to capture intrinsic neighborhood information means that these descriptors can be used to generate machine learned potentials and expedite first principles molecular dynamics or Monte Carlo simulations within the *learn on the fly* paradigm.[10]

Acknowledgments

The author wishes to thank John Klepeis, Eric Schwegler, Jonathan Belof, Edgar Landinez, Xavier Andrade, Arthur Tamm, ShinYoung Kang and Leora Dresselhaus-Cooper for numerous stimulating discussions, and Albert Bartók for sharing the details of reconstruction of Si-clusters using bispectrum and power spectrum descriptors. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Computing support for this work came from the Lawrence Livermore National Laboratory (LLNL) Institutional Computing Grand Challenge program. (IM release number: LLNL-JRNL-757813)

[*] Electronic address: `samanta1@llnl.gov`

[1] J. Behler and M. Parrinello, Physical Review Letters **98**, 146401 (2007).

[2] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Physical Review Letters **104**, 136403 (2010).

[3] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Physical Review Letters **108**, 058301 (2012).

[4] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, Journal of Chemical Theory and Computation **9**, 3404 (2013).

[5] A. Seko, A. Takahashi, and I. Tanaka, Physical Review B **90**, 024101 (2014).

[6] A. Seko, A. Takahashi, and I. Tanaka, Physical Review B **92**, 054113 (2015).

[7] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, Journal of Computational Physics **285**, 316 (2015).

[8] M. A. Wood and A. P. Thompson, Journal of Chemical Physics **148**, 241721 (2018).

[9] A. P. Bartók, R. Kondor, and G. Csányi, Physical Review B **87**, 184115 (2013).

[10] Z. Li, J. R. Kermode, and A. De Vita, Physical Review Letters **114**, 096405 (2015).

[11] A. Glielmo, P. Sollich, and A. De Vita, Physical Review B **95**, 214302 (2017).

[12] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, Physical Review Letters **120**, 036002 (2018).

[13] S. J. Plimpton and A. P. Thompson, MRS Bulletin **37**, 513 (2012).

[14] V. Botu and R. Ramprasad, International Journal of Quantum Chemistry **115**, 1074 (2015).

[15] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, The Journal of Physical Chemistry Letters **6**, 2326 (2015).

[16] J. Cui and R. V. Krems, Journal of Physics B: Atomic, Molecular and Optical Physics **49**, 224001 (2016).

[17] A. Kamath, R. A. Vargas-Hernńdez, R. V. Krems, T. C. Jr., and S. Manzhos, Journal of Chemical Physics **148**, 241702 (2018).

[18] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, npj Computational Mathematics **3**, 37 (2017).

[19] A. A. Peterson, R. Christensen, and A. Khorshidi, Physical Chemistry Chemical Physics **19**, 10978 (2017).

[20] L. Bartok-Pártay, A. Bartók, and G. Csányi, **114**, 10502 (2010).

[21] T. Q. Yu, P. Y. Chen, M. Chen, A. Samanta, E. Vanden-Eijnden, and M. Tuckerman, Journal of Chemical Physics **140**, 214109 (2014).

[22] A. Samanta, M. Chen, T. Q. Yu, M. Tuckerman, and W. E, Journal of Chemical Physics **140**, 164109 (2014).

[23] A. Samanta, M. A. Morales, and E. Schwegler, Journal of Chemical Physics **144**, 164101 (2016).

[24] L. Koziol, L. E. Fried, and N. Goldman, Journal of Chemical Theory and Computation **13**, 135 (2017).

[25] R. K. Lindsey, L. E. Fried, and N. Goldman, Journal of Chemical Theory and Computation **13**, 6222 (2017).

[26] L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton, et al., Journal of Chemical Physics **144**, 034203 (2016).

[27] A. V. Shapeev, Multiscale Modeling and Simulation **14**, 1153 (2016).

[28] J. Behler, Physical Chemistry Chemical Physics **13**, 17930 (2011).

[29] J. Behler, The Journal of Chemical Physics **134**, 074106 (2011).

[30] J. Behler, Journal of Physics: Condensed Matter **26**, 183001 (2014).

[31] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, Journal of Chemical Physics **148**, 241709 (2018).

[32] A. Takahashi, A. Seko, and I. Tanaka, Journal of Chemical Physics **148**, 234106 (2018).

[33] L. Zhang, J. Han, H. Wang, R. Car, and W. E, Physical Review Letters **120**, 143001 (2018).

[34] R. Kondor and J. D. Lafferty, Proceedings of the International Conference on Machine Learning p. 315 (2002).

[35] M. Belkin and P. Niyogi, Neural Computation **15**, 1373 (2003).

[36] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proceedings of the National Academy of Sciences of the United States of America **102**, 7426 (2005).

[37] A. E. Brouwer and W. H. Haemers, *Spectra of graphs* (Springer, New York, NY, 2012), ISBN 978-1-4614-1938-9.

[38] J. Shi and J. Malik, IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 888 (2000).

[39] A. Ng, M. Jordan, and Y. Weiss, Advances in Neural Information Processing Systems **14**, 849 (2002).

[40] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis, Advances in Neural Information Processing Systems **18**, 955 (2006).

[41] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt, Journal of Machine Learning Research **11**, 1201 (2010).

[42] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, Journal of Machine Learning Research **12**, 2539 (2011).

[43] G. Ferré, T. Haut, and K. Barros, Journal of Chemical Physics **146**, 114107 (2017).

[44] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (2006).

[45] D. Duvenaud, H. Nickisch, and C. E. Rasmussen, Advances in Neural Information Processing Systems **24**, 226 (2011).

[46] A. P. Bartók, R. Kondor, and G. Csányi, Physical Review B **96**, 019902 (2017).

[47] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, SIAM Journal on Optimization **9**, 112 (1998).

[48] M. Randić, Journal of Chemical Information and Computer Sciences **15**, 105 (1975).

[49] N. Artrith and J. Behler, Physical Review B **85**, 045439 (2012).