The Journal of Chemical Physics

Maximum volume simplex method for automatic selection and classification of atomic environments and environment descriptor compression

Cite as: J. Chem. Phys. **153**, 214104 (2020); https://doi.org/10.1063/5.0030061
Submitted: 22 September 2020 • Accepted: 10 November 2020 • Published Online: 01 December 2020

Behnam Parsaeifard, Daniele Tomerini, Deb Sankar De, et al.

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Maximum volume simplex method for automatic selection and classification of atomic environments and environment descriptor compression

Cite as: J. Chem. Phys. 153, 214104 (2020); doi: 10.1063/5.0030061
Submitted: 22 September 2020 • Accepted: 10 November 2020 •
Published Online: 1 December 2020

Behnam Parsaeifard,[1,2] Daniele Tomerini,[1,2] Deb Sankar De,[1,2] and Stefan Goedecker[1,2,a)]

AFFILIATIONS

[1] Department of Physics, Universitaet Basel, Klingelbergstrasse 82, 4056 Basel, Switzerland
[2] National Center for Computational Design and Discovery of Novel Materials (MARVEL), Lausanne, Switzerland

Note: This paper is part of the JCP Special Topic on Computational Materials Discovery.
[a)] **Author to whom correspondence should be addressed:** stefan.goedecker@unibas.ch

ABSTRACT

Fingerprint distances, which measure the similarity of atomic environments, are commonly calculated from atomic environment fingerprint vectors. In this work, we present the simplex method that can perform the inverse operation, i.e., calculating fingerprint vectors from fingerprint distances. The fingerprint vectors found in this way point to the corners of a simplex. For a large dataset of fingerprints, we can find a particular largest simplex, whose dimension gives the effective dimension of the fingerprint vector space. We show that the corners of this simplex correspond to landmark environments that can be used in a fully automatic way to analyze structures. In this way, we can, for instance, detect atoms in grain boundaries or on edges of carbon flakes without any human input about the expected environment. By projecting fingerprints on the largest simplex, we can also obtain fingerprint vectors that are considerably shorter than the original ones but whose information content is not significantly reduced.

I. INTRODUCTION

Materials science has become, to a large extent, a data driven science. Several data banks exist that contain not only structural data but also calculated properties; many exceed the hundreds of thousands of structural properties in number, with their number growing dramatically.[1–4] Molecular dynamics (MD) simulations typically also generate very large datasets. Such large datasets cannot anymore be inspected by eye and tools for classifying the structures in an automatic way are needed. Atomic environments can be described in a quantitative fashion by descriptors called "atomic environment fingerprints,"[5–9] which can also provide a description for entire crystalline structures.[10] Atomic environment fingerprints are also used as inputs for supervised machine learning schemes[11–13] of potential energy surfaces. For such a use, it is desirable that the fingerprint is able to detect any difference in the environment[14] while keeping the fingerprint vector as short as possible.

One of our goals will be the detection of grain boundaries, which are the disordered regions between one or two ordered phases. Grain boundaries have an important influence on physical properties of the system including strength, conductivity, ductility, and crack resistance.[15–20]

Several methods have been proposed in the literature to distinguish between certain reference crystalline structures and disordered and mainly liquid structures in melting and nucleation simulations such as Steinhardt parameters[21] and common neighbor analysis (CNA).[22] These methods have also been used to study dislocations,

local ordering, and grain boundaries.[23–27] One of the disadvantages of these methods is that they are based on a sharp cutoff, and they end up lacking smoothness with respect to particle displacements occurring in MD or during relaxations. As its name suggests, in the adaptive common neighbor analysis,[28] the cutoff is adapted to the environment of each atom. Although more robust compared to CNA, it remains sensitive to thermal vibrations. Different predefined crystalline structures can be distinguished by polyhedral template matching.[29] SOAP[5] fingerprints coupled to machine learning methods were recently also used to predict properties of grain boundaries.[30] Based on a formula to calculate the entropy for a system interacting only via pairwise forces, an atomic entropy can be obtained, which allows us to distinguish between liquid, FCC, BCC, and HCP crystalline phases.[31] Several other methods exist in the computational physics and machine learning communities for the selection of fingerprint components and atomic environments. In the Pearson correlation method, the correlation between the selected features and the atomic environments is optimized.[32] In the farthest point method,[33] the Euclidian fingerprint distance between the data points is maximized. Sketch maps[32] try to map faithfully distances from a high dimensional into a low dimensional space. The unsupervised landmark analysis of Kahle *et al.* is based on a Voronoi tessellation of the space such that all points in a certain region are closer to the points in the same region than the points in other regions.[34] CUR decomposition finds a low rank of the fingerprint matrix such that the least information is lost.[35] In the principal component analysis (PCA), the covariance matrix is diagonalized and the most important directions are selected.[36]

In this work, we introduce a method that selects all the relevant structures fully automatically based on a large pool of structures. The method is also applicable without any adjustments to any molecular system whose atomic environments can be represented by fingerprints.

II. THE LARGEST SIMPLEX METHOD

A. Fingerprints and fingerprint distances

In this section, we provide a short review of the overlap matrix (OM) fingerprint method that we use to describe the local atomic environment. A complete description can be found in the original paper detailing the method.[10,37]

In order to calculate the overlap matrix (OM) fingerprint for an atom k in a structure, we take into account the relative position of all the neighbors of that atom within a cutoff sphere (centered on atom k) of radius R_c. Neighbors include all the relevant periodic images of an atom when dealing with an atom at the edge of a repeating unit for a periodic system. Each of the atoms is associated with a minimal set of normalized atom-centered Gaussians $G_\nu(\boldsymbol{r}-\boldsymbol{R}_i)$, centered on the atom itself. The width of each Gaussian is given by the covalent radius of the atom on which it is centered. For carbon with its strong directional bonding, we have used a set of s and p-type orbitals ($\nu = s, p_x, p_y, p_z$) and denote the resulting fingerprint by OM[sp], and for aluminum with its metallic bonding, we have used only $\nu = s$ and denote the fingerprint by OM[s]. We then calculate the overlap between Gaussian functions in the sphere,

$$S^k_{i,\nu,j,\mu} = \int G_\nu(\boldsymbol{r}-\boldsymbol{R}_i)G_\mu(\boldsymbol{r}-\boldsymbol{R}_j)d\boldsymbol{r}. \tag{1}$$

Next, the overlap matrix $S^k_{i,\nu,j,\mu}$ is multiplied by the amplitude functions $f_c(|\boldsymbol{R}_k-\boldsymbol{R}_i|)$ and $f_c(|\boldsymbol{R}_k-\boldsymbol{R}_j|)$ to obtain a modified overlap matrix $\tilde{S}$,

$$\tilde{S}^k_{i,\nu,j,\mu} = f_c(|\boldsymbol{R}_k-\boldsymbol{R}_i)S^k_{i,\nu,j,\mu}f_c(|\boldsymbol{R}_k-\boldsymbol{R}_j|). \tag{2}$$

$f_c(r) = (1-\frac{r^2}{4w^2})^2$ is a cutoff function that vanishes at and beyond $r = 2w = R_c$ with two continuous derivatives. w gives the length scale over which $f_c(r)$ drops to zero and we typically choose it so that about 50 atoms are contained within the cutoff radius R_c. The matrix whose columns are denoted by the composite index i, ν and whose rows are given by the composite index j, μ is then diagonalized to obtain the eigenvalues. Finally, the vector $\boldsymbol{V}^k$ containing all the sorted eigenvalues of the matrix $\tilde{S}^k_{i,\nu,j,\mu}$ is the fingerprint of atom k. It has a length $L = 4N_{sphere}$ for OM[sp] and $L = N_{sphere}$ for OM[s], where N_{sphere} is the number of atoms in the sphere around the central atom.

By construction, the fingerprint is robust against displacements of the atoms across the boundary of the sphere radius, and therefore, it is possible to calculate derivatives of the fingerprints with respect to infinitesimal structural change around the atom k. The fingerprint vectors $\boldsymbol{V}^k$ characterize the atomic environments around atom k, and the fingerprint distance $d_{i,j}$ is a measure of the dissimilarity between two environments i and j. The fingerprint distance is obtained from the Euclidean norm of the difference vector throughout this study,

$$d_{i,j} = |\boldsymbol{V}^i - \boldsymbol{V}^j|. \tag{3}$$

B. Obtaining fingerprint vectors from fingerprint distances

The above formula (3) gives a trivial recipe to obtain fingerprint distances $d_{i,j}$ from a set of points represented by the fingerprint vectors in a space of dimension L. In the following, we will derive the formulas for the inverse operation. Given a set of pairwise fingerprint distances $d_{i,j}$, we want to construct a set of points $\boldsymbol{x}^i$ that will satisfy these constraints. The solution of this problem is not unique. The solution can, however, be made unique by requiring that the first point be the origin, $\boldsymbol{x}^0 = 0$, and that for each consecutive point, the number of nonzero components increases by one. Hence, the points $\boldsymbol{x}^i$ have the following structure:

$$(\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N) = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,N} \\ 0 & x_{2,2} & \cdots & x_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & x_{N,N} \end{pmatrix}. \tag{4}$$

Hence, after placing the first point at the origin, the next point lies on the positive x-axis at the right distance, the following on the xy plane ($y > 0$), and so on. The components of the set of points $\boldsymbol{x}^i$'s can be obtained recursively from simple relations between the distances among the vectors $\boldsymbol{V}^i$'s.

The distance between $\boldsymbol{x}^N$ and the origin, $\boldsymbol{x}^0$, is simply given by the norm of the vector,

$$d^2_{0,N} = \sum_{i=1}^{N} x^2_{i,N}. \tag{5}$$

For $M < N$, the difference between columns N and M is related to the distance between points $\boldsymbol{x}^N$ and $\boldsymbol{x}^M$ as

$$d^2_{M,N} = \sum_{i=1}^{M} \left(x_{i,N} - x_{i,M}\right)^2 + \sum_{j=M+1}^{N} x^2_{j,N}. \tag{6}$$

By taking the difference between $d^2_{M,N}$ and $d^2_{0,N}$, we obtain a simplified set of equations,

$$d^2_{M,N} - d^2_{0,N} = \sum_{i=1}^{M} \left(x_{i,N} - x_{i,M}\right)^2 - x^2_{i,N} = \sum_{i=1}^{M} -x_{i,M}\left(2x_{i,N} - x_{i,M}\right). \tag{7}$$

In Eq. (7), the unknowns $x_{i,N}$ depend only on other column elements $x_{j,M}$ with $M < N$,

$$x_{1,1} = d_{0,1}, \tag{8}$$

$$x_{1,2} = \frac{d^2_{0,1} + d^2_{0,2} - d^2_{1,2}}{2x_{1,1}}, \tag{9}$$

$$x_{2,2} = \sqrt{d^2_{0,2} - x^2_{1,2}}. \tag{10}$$

We can write for $M < N$, in general,

$$x_{M,N} = \frac{d^2_{0,M} + d^2_{0,N} - d^2_{M,N} - 2\sum_{i=1}^{M-1} x_{i,M}x_{i,N}}{2x_{M,M}}, \tag{11}$$

and for $M = N$, we have

$$x_{N,N} = \sqrt{d^2_{0,N} - \sum_{i=1}^{N-1} x^2_{i,N}}. \tag{12}$$

The geometrical body having as corners the above calculated points is a N-dimensional simplex with volume $x_{1,1}x_{2,2}\cdots x_{N,N}/N!$. The above construction can be done for any set of $\frac{N_{env}(N_{env}-1)}{2}$ distances as long as the original $\boldsymbol{V}^i$'s giving rise to the distances via Eq. (3) are linearly independent. Since the number of environments N_{env} is typically much larger than the length L of the fingerprint vectors, at most, L points (including in the count the origin) can be obtained. If the number of linearly independent fingerprint vectors is less than L, $x_{i,i}$ will become zero for some $i < L$, and it is thus not possible to increase the dimension of the simplex. In the context of our fingerprints, it turns out that the $x_{i,i}$ typically are not exactly zero but become very small, which means that all the fingerprint vectors are essentially contained in a sub-volume whose dimension is smaller than L. The component that is orthogonal to this subspace is then very small and can be neglected. This is the basic property that will be exploited for the fingerprint compression later in the paper.

C. Construction of the largest simplex

Now, we will describe how we can use the construction outlined above to obtain the largest simplex, which we will simply denote by the largest simplex (LS). We do this since we are interested in finding the effective dimension l of the space spanned by the fingerprints, which gives the number of the highly distinctive landmark environments together with these environments. We start by identifying the two environments characterized by the largest distance. This defines the origin $\boldsymbol{x}^0$ and the first point along the x-axis, i.e., $\boldsymbol{x}^1$, and in this way, the first two corners of the simplex, which is, at this stage, just a line. To enlarge, in the next step, the dimension of the simplex by one, we search for the environment that will give the largest area triangle if the point $\boldsymbol{x}^2$, corresponding to this environment, is used as the third corner. We then increase the dimension of the simplex step by step and we choose the new corners in each step in such a way that the volume of the new simplex will be maximal. The procedure is stopped if in a certain step l, the volume collapses to a very small value because additional fingerprint vectors are quasi linearly dependent on the previous ones. In this way an effective dimension l of the entire fingerprint space can be determined. Once this largest simplex is constructed, we can express other fingerprint vectors in the basis of the vectors $\boldsymbol{x}^i$ spanning the LS. To get the expansion coefficients, we just perform the same steps of Eqs. (8)–(12) that would be needed to add a corner to the simplex. However, in this case, we know already that the $x_{l+1,l+1}$ from Eq. (12) will be negligible because we stopped the largest simplex construction exactly for the reason that we could not find any point that gave a large $x_{l+1,l+1}$.

III. APPLICATIONS

In this section, we show some applications of the LS. In Sec. III A, we apply the methodology to the study of a variety of C_{60} molecules to identify the most distinct environments and group the most similar ones. In Sec. III B, we use the method to find the grain boundaries in a Al nanocrystalline material. In Sec. III C, we exploit the LS to reduce the dimensions of the fingerprint and compare its performance with CUR decomposition method.[35]

A. C_{60} clusters

Our first system to be studied consists of 5000 C_{60} structures, i.e., 5000 × 60 atomic environments, that exhibit several structural motifs including sheets, chains, and cages. These structures were generated by minima hopping[38] runs coupled to Density Functional based Tight Binding (DFTB).[39] Our aim is to identify the most distinct atomic environments as well as to classify the environments. We use OM[sp] with a cutoff radius of $R_c = 2w = 6$ Å and follow the approach described in Sec. II to generate the LS with $N = 60$. In Fig. 1, we show the first 20 corners of the LS, which represent 20 highly distinct landmark environments in the dataset. In agreement with the basic chemical intuition, the first two corners representing the two most different chemical environments are a fourfold

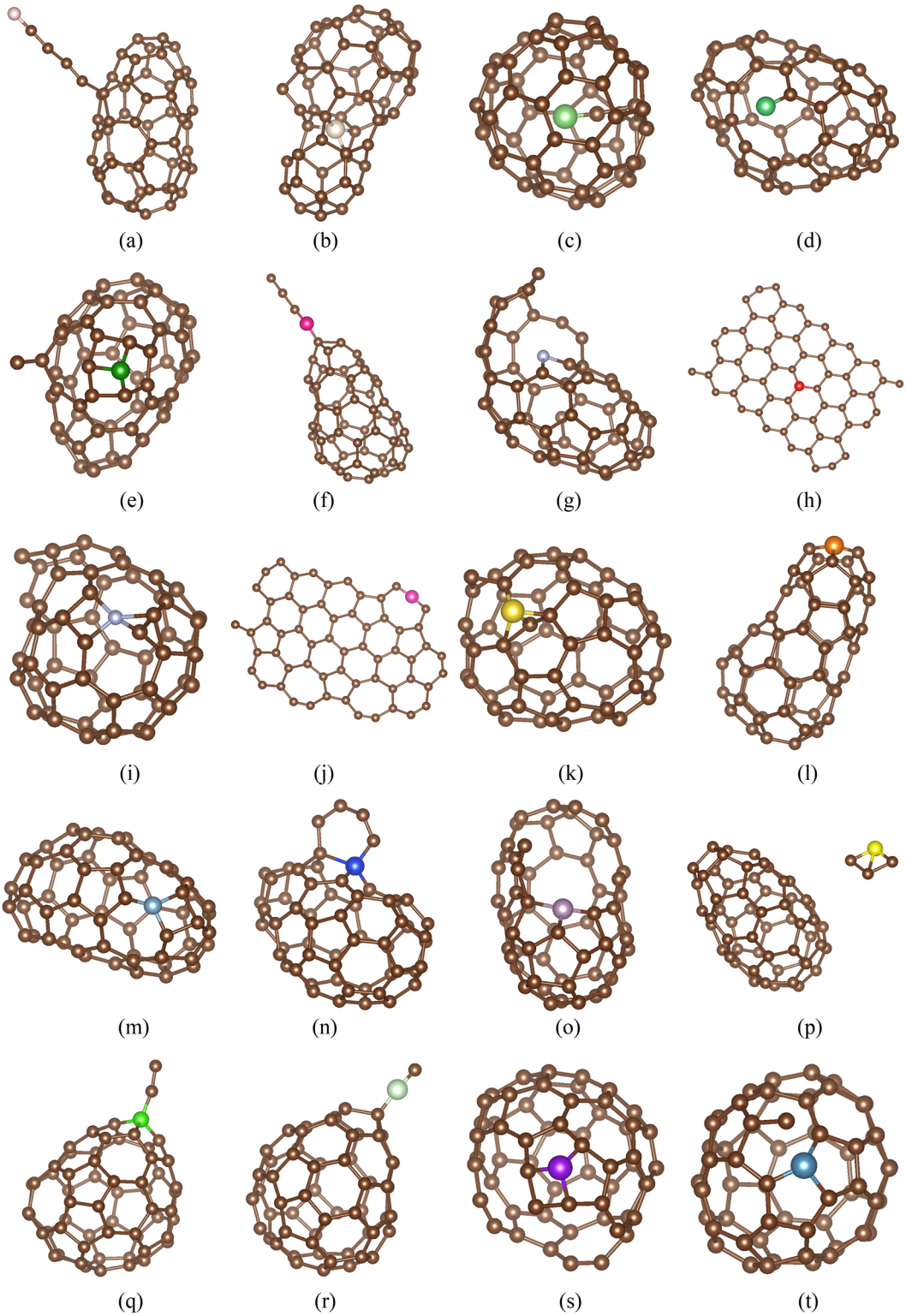

FIG. 1. (a)–(t) The first 20 corners of the LS, i.e., the 20 most distinct atomic environments. The central atoms are shown in a different color than the rest of the atoms. The relative size and the colors of the atoms are for visualization purposes and are not physically important.

coordinated atom and a carbon atom at the end of a linear chain with only one nearest neighbor, as shown in Figs. 1(b) and 1(a). Other twofold coordinated atoms in chains are also represented by higher order corners of the LS, as shown in Figs. 1(f), 1(q), 1(r), and 1(c). In Fig. 1(c), the reference atom is part of a chain, but the chain points inside the cage, which shows that our method can distinguish between chains that point inward or outward since it is not based solely on its nearest neighbors, but on its general environment.

The fourth corner of the LS is an atom with one nearest neighbor and near a hole in the C_{60} shown in Fig. 1(d). Other corners of the LS also clearly represent truly different environments. For instance, the 8th corner of the LS shown in Fig. 1(h) is an atom in a graphite flake and the 16th corner of the LS is an atom in a fragmented part shown in Fig. 1(p). Our dataset contains only a few fragmented structures in the dataset, which are of type Fig. 1(p) and the LS could correctly recognize them as highly distinct environments.

Next, we employ the corners of the LS to analyze structures. Based on the fact that each corner represents highly distinct landmark environments, we can assume that each fingerprint that has a small fingerprint distance to any of these corners represents an environment that is similar to the corresponding landmark environment. Hence, we assign each atomic environment to its closest corner if the fingerprint distance is less than a threshold value δ, which we take to be 0.5. With this criterion, we calculate the number of environments that belong to each class, as shown in Fig. 2. The environments that do not belong to any corner of the LS because their fingerprint distance to the their closest corner is larger than δ are shown in the blue bar in Fig. 2. Since the first corner is at the origin, Fig. 2 starts at zero.

The energetic minimum of the C_{60} molecule is the fullerene molecule. In this structural motif, the atomic environments for all of the carbon atoms are equivalent. This is not true anymore if the fullerene has a so-called Stone–Wales defect.[40] In the following, we look at such a structure as well as a 60 atom graphite flake and categorize the atoms according to their fingerprint distance to the landmark environments, i.e., the corners of the LS. None of the atomic environments of these two structures is actually a landmark

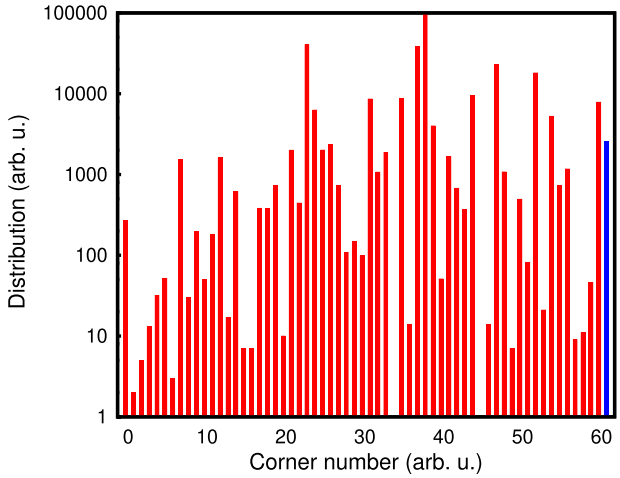

FIG. 2. The number of atomic environments in the dataset of C_{60} structures, which are similar to one of the corners of the LS. The blue bar represents environments that are not similar to any corner based on the threshold value $\delta = 0.5$.

environment of the LS. For the visualization, we assign a color to each corner of the LS. All the atomic environments in the data that have a short fingerprint distance to this corner are then shown in this color.

Our method automatically classifies the atoms of the structure shown in Fig. 3(a) into three types, and we can easily verify by visual inspection that these three classes are in agreement with chemical intuition: We see an atom surrounded by two pentagons and one hexagon [corner 47 shown in Fig. 3(b)], one pentagon and two hexagons [corner 38 shown in Fig. 3(c)], or three hexagons [corner 23 shown in Fig. 3(d)]. As can be seen from Fig. 2, a large number of atomic environments in our dataset are similar to these corners.

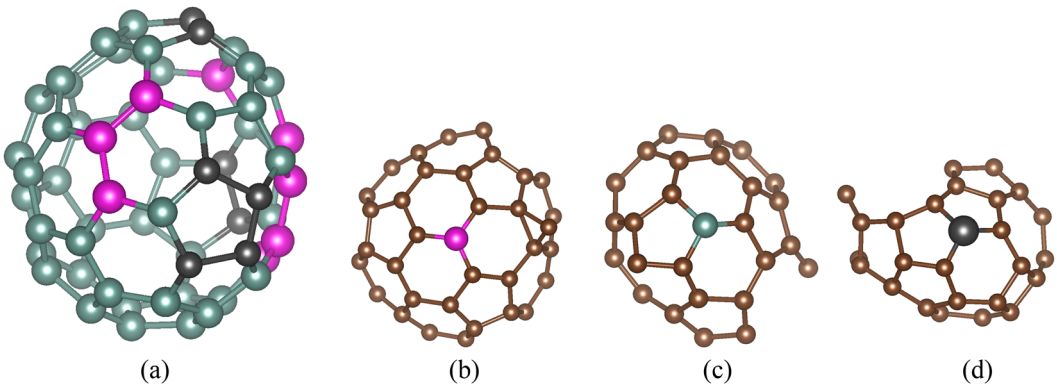

FIG. 3. (a) A C_{60} with a Stone–Wales defect: the atoms are colored according to their closest corners, which is shown by the same color in the other three images. (b) Corner 47, (c) corner 38, and (d) corner 23 of the LS.

Another example is shown in Fig. 4. The atoms of the structure in Fig. 4(a) are similar to one of the six different corners of the LS. These are shown in Figs. 4(b)–4(g). Hence, indeed groups of environments that have a short distances to a landmark environment share similar chemical environments.

B. Grain boundary networks in nanocrystalline Al

In our second application, we study a nanocrystalline Al aggregate with 255 064 atoms containing grain boundary networks. The details on the generation of the nanocrystalline Al used here can be found elsewhere.[31] We use the OM[s] fingerprint with a cutoff radius of R_c = 5 Å to build the LS. We take N = 46, which is the same as the length of the fingerprint. Having generated the LS, we assign a different color to each of the corners of the LS for the following visualizations. These corners are the most distinct environments in the nanocrystalline Al, i.e., each corner can represent a class of diverse environments in the data. We again categorize the atoms in the system according to their similarity to the corners of the LS and assign them the same color as the corners they resemble most. Visual inspection of Fig. 5 shows that the LS can find all the grain boundary networks, in agreement with the findings of Piaggi.[31] In addition, it can also recognize differences between different grain boundaries and find different kinds of ordered–disordered phases, as shown in Fig. 6.

In Fig. 6, we showed the first 20 corners of the LS. Figure 6(a) shows a perfect crystalline FCC phase. Figures 6(c) and 6(r) show the defective crystalline FCC phases where one nearest neighbor of the central atom is missing. The corners shown in Figs. 6(e), 6(n), 6(p), and 6(s) correspond to atoms on a twisted grain boundary. The configurations from Figs. 6(b), 6(d), 6(h), 6(l), and 6(t) represent environments located on the boundary between ordered and disordered phases. Finally, some corners of the LS represent atoms in disordered phases such as those shown in Figs. 6(i) and 6(j).

C. The compression of the fingerprints

In Sec. II, we showed that once the LS is found, the original fingerprints can be projected onto the LS. In this section, we will show that these projections can be regarded as a new fingerprint whose length is much shorter than the original fingerprint while containing most of the information of the original fingerprint. This is an example of data compression, a problem for which many algorithms are available, such as CUR[35] decomposition. Assuming that F is the fingerprint matrix with dimension $L \times N'$, where L is the length of the fingerprint and N' is the number of atomic environments $N' = N_{env}$, i.e., ith column of F contains the fingerprint vector of atomic environment i, one can write $F \sim CUR$ in which C and R contain k selected columns and rows of F and $U = C^{+}FR^{+}$, where A^{+}

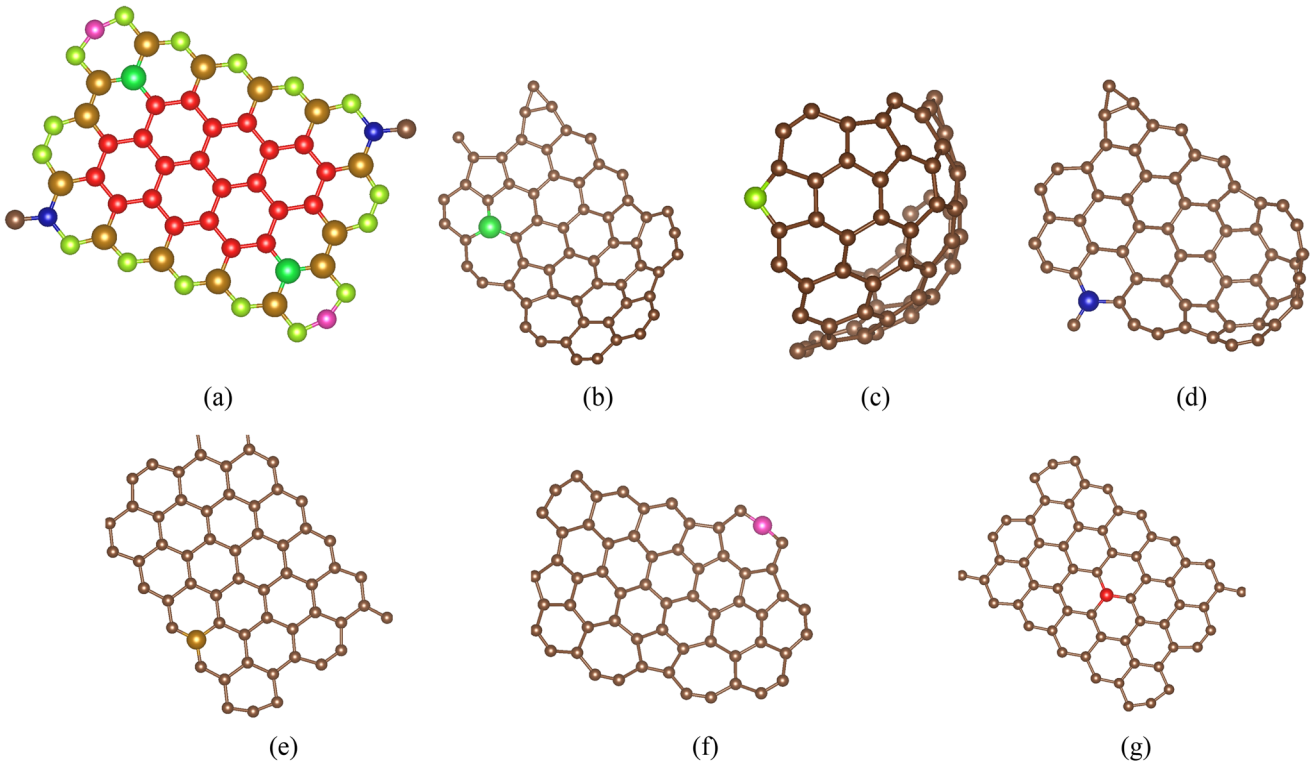

FIG. 4. (a) A graphite flake whose atoms are colored according to their closest corners. (b) Corner 55, (c) corner 33, (d) corner 26, (e) corner 25, (f) corner 9, and (g) corner 7 of the LS.

indicates the pseudo-inverse of A and $k < r = rank(F)$. In order to find the reduced selected number of rows of matrix F, one writes its Singular Value Decomposition (SVD) as $F = \tilde{U}DV^T$, where $\tilde{U}$ (left singular matrix) and V (right singular matrix) are $L \times L$ and $N' \times N'$ unitary matrices and D is a $L \times N'$ rectangular diagonal matrix with non-negative real numbers on the diagonal. The diagonal entries of D are known as the singular values of F. Then, the leverage score for each row i is calculated as $\pi_i = \frac{1}{k}\sum_{\xi=1}^{k}(u_i^{\xi})^2$, where u_i^{ξ} is the ith component of ξth left singular vector and k is the number of rows that should be selected. Frequently, rows are selected with probability proportional to the leverage score. We employed a deterministic method[32,42] and select the row with the highest leverage score at each time. Then, the selected row is removed from the matrix, and the rest of the rows become orthogonalized with respect to it. To select other rows, this procedure is repeated. The selected rows are the most important features. One can also select columns of the matrix F, i.e., the most important atomic environments by following the same procedure but for F^T. The selected rows and column are stored in R and C, respectively.

In the following, we employ the LS and CUR method to reduce the length of the fingerprint by selecting the components of the fingerprint that contain the most important information.

In order to investigate whether the compressed fingerprint conserves the information encoded in the original fingerprint, we correlate all the pairwise fingerprint distances obtained by the original and compressed fingerprints.[14]

Obviously, fingerprint distances that are large with the original fingerprint should remain large with the compressed fingerprint. In

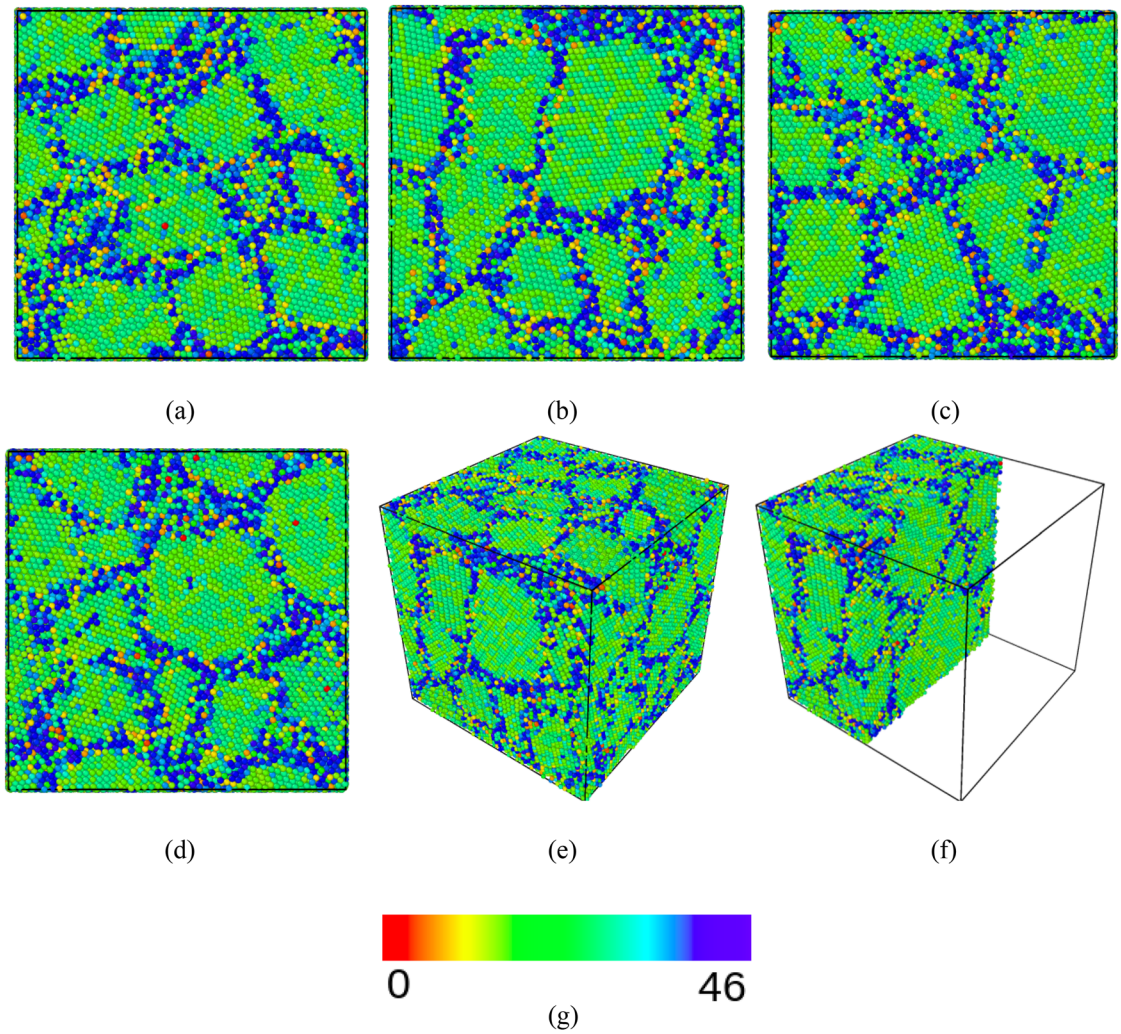

FIG. 5. Nanocrystalline Al containing grain boundaries. The LS is employed to find the grain boundary networks. Atoms are assigned the same color as their closest corner of the LS, i.e., the corner with which they have the smallest fingerprint distance. (a) View from top. (b) View from front. (c) View from left. (d) View from right. (e) Perspective view. (f) Slice view. Software Ovito[41] is used for the visualization.

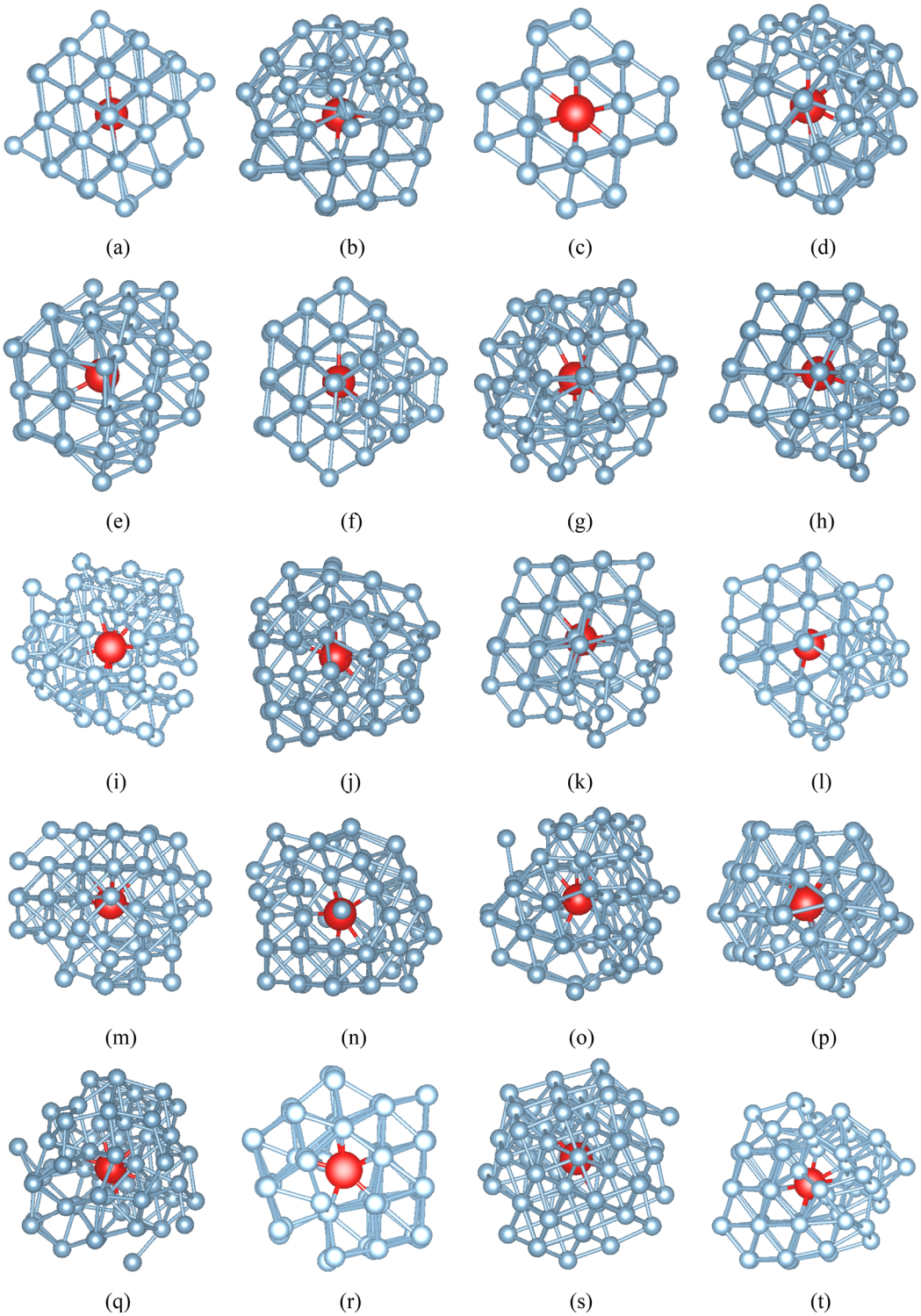

FIG. 6. The first 20 most distinctive atomic environments in the nanocrystalline Al found by the LS are shown in (a)–(t). Red atoms are the central atoms whose local environment is one of the corners of the LS and the atoms in their vicinity are depicted in blue.

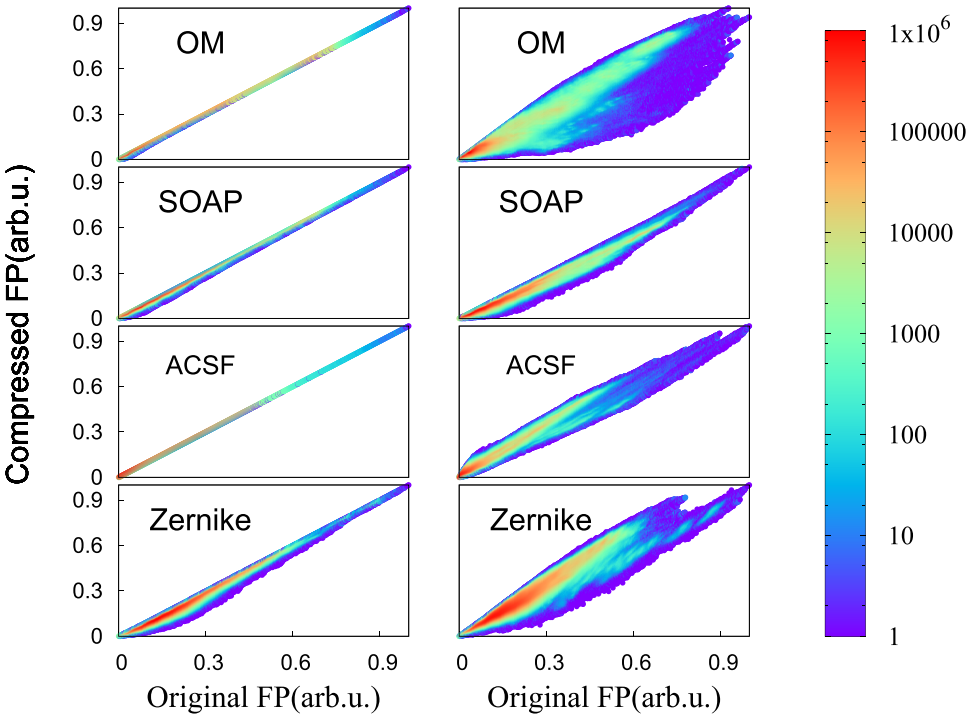

FIG. 7. The correlation between the original fingerprints and LS-reduced fingerprints (left column) and CUR-reduced fingerprints (right column) for OM, SOAP, atom-centered symmetry functions (ACSF), and Zernike. The length of the reduced fingerprints is l = 16, while the length of the original fingerprint L is 240 for OM, 325 for SOAP, 58 for ACSF, and 121 for Zernike. The fingerprint distances are scaled such that the maximum fingerprint is 1 in each case.

the same way, short distances should remain short. If this is the case, all the points in a correlation plot between the fingerprint distances arising from the original and the compressed fingerprint will lie on or close to the diagonal. If there are points far away from the diagonal and, in particular, if some fingerprint distances of the compressed fingerprint are small, whereas the original distances are large, there is a loss of information.

In Fig. 7, we show the correlation plot between the original fingerprints and the LS- and CUR-reduced fingerprints using OM, SOAP,[5] atom-centered Behler–Parrinello symmetry functions (ACSF),[6,43] and Zernike fingerprints[44] for our above-mentioned test of 1000 C_{60} clusters with 1000 × 60 atomic environments. We used the same fingerprint parameters for OM as in Sec. III A. For SOAP, we used the following parameters: $l_{max} = n_{max} = 8$, $r_\delta = 4.0$ Å, and $\sigma = 0.5$ Å. We used the standard parameters for ACSF.[6] For Zernike, we used $n_{max} = 20$. The cutoff radius is 6 Å for all the fingerprints. The software QUIP[45] is used to generate the ACSF and SOAP fingerprints. For the Zernike fingerprint, we used the software atomistic machine-learning package (AMP).[44] We reduced the length of the fingerprints to $l = 16$ in all cases. As can be seen in Fig. 7, the correlation is almost diagonal in the case of LS, which indicates that vast majority of the information of the original fingerprint is retained in the LS-reduced fingerprint. There are, however, some deviations from the diagonal in the correlation plot between the original fingerprint and CUR-reduced fingerprint, which indicates that some information is lost in the CUR-decomposition.

IV. CONCLUSION

We have introduced an algorithm to construct a largest simplex in the space spanned by a large set of atomic environment fingerprint vectors. The number of corners of this LS gives the effective dimension of the fingerprint vector space. The corners themselves represent landmark environments that can be used to analyze structures with a large number of atoms in a fully automatic way. Hence, in contrast to other methods, it is not necessary to include into our analysis tool criteria that are based on human expectations of what kind of environments are expected to be encountered in this system. We show that this analysis method can be used to detect grain boundaries and other typical environments in multigrain metallic systems and to classify atomic environments in a carbon cluster in a way that is consistent with basic chemical intuition. Since only those components of the fingerprint vector that are inside the space spanned by the LS are relevant, projecting the fingerprint into the space spanned by the LS reduces the length of the fingerprint without any significant loss of information. Therefore, the method can also be used as a data compression method for fingerprints.

SUPPLEMENTARY MATERIAL

See the supplementary material for the structures in Fig. 6.

ACKNOWLEDGMENTS

The authors thank Dr. Pablo Piaggi for providing us the nanocrystalline Al data. The authors acknowledge that this research was supported by NCCR MARVEL and funded by the Swiss National Science Foundation. Structures were visualized using VESTA[46] and Ovito[41] packages. The calculations were performed on the computational resources of the Swiss National Supercomputer (CSCS) under project s963 and on the Scicore computing center of the University of Basel.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] A. Jain, G. Hautier, C. J. Moore, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, Comput. Mater. Sci. **50**, 2295 (2011).
[2] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).
[3] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, Comput. Mater. Sci. **58**, 227 (2012).
[4] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi, and N. Marzari, Sci. Data **7**, 299 (2020).
[5] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[6] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[7] F. A. Faber, A. S. Christensen, B. Huang, and O. A. Von Lilienfeld, J. Chem. Phys. **148**, 241717 (2018).
[8] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. A. von Lilienfeld, J. Chem. Phys. **152**(4), 044107 (2020).
[9] M. Hirn, S. Mallat, and N. Poilvert, Multiscale Model. Simul. **15**, 827 (2017).
[10] L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton *et al.*, J. Chem. Phys. **144**, 034203 (2016).
[11] J. Behler, Int. J. Quantum Chem. **115**, 1032 (2015).
[12] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[13] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[14] B. Parsaeifard, D. S. De, A. S. Christensen, F. A. Faber, E. Kocer, S. De, J. Behler, A. von Lilienfeld, and S. Goedecker, Mach. Learn.: Sci. Technol. (published online 2020).
[15] N. Hansen, Scr. Mater. **51**, 801 (2004).
[16] A. Chiba, S. Hanada, S. Watanabe, T. Abe, and T. Obana, Acta Metall. Mater. **42**, 1733 (1994).
[17] T. H. Fang, W. L. Li, N. R. Tao, and K. Lu, Science **331**, 1587 (2011).
[18] M. Shimada, H. Kokawa, Z. J. Wang, Y. S. Sato, and I. Karibe, Acta Mater. **50**, 2331 (2002).
[19] L. Lu, Y. Shen, X. Chen, L. Qian, and K. Lu, Science **304**, 422 (2004).
[20] M. A. Meyers, A. Mishra, and D. J. Benson, Prog. Mater. Sci. **51**, 427 (2006).
[21] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).
[22] D. Faken and H. Jónsson, Comput. Mater. Sci. **2**, 279 (1994).
[23] J. Schiotz and K. W. Jacobsen, Science **301**, 1357 (2003).
[24] V. Yamakov, D. Wolf, S. R. Phillpot, A. K. Mukherjee, and H. Gleiter, Philos. Mag. Lett. **83**, 385 (2003).
[25] C. Brandl, P. M. Derlet, and H. Van Swygenhoven, Modell. Simul. Mater. Sci. Eng. **19**, 074005 (2011).
[26] H. Jónsson and H. C. Andersen, Phys. Rev. Lett. **60**, 2295 (1988).
[27] N. P. Bailey, J. Schiøtz, and K. W. Jacobsen, Phys. Rev. B **69**, 144205 (2004).
[28] A. Stukowski, Modell. Simul. Mater. Sci. Eng. **20**, 045021 (2012).
[29] P. M. Larsen, S. Schmidt, and J. Schiøtz, Modell. Simul. Mater. Sci. Eng. **24**, 055007 (2016).
[30] C. W. Rosenbrock, E. R. Homer, G. Csányi, and G. L. Hart, npj Comput. Mater. **3**, 29 (2017).
[31] P. M. Piaggi and M. Parrinello, J. Chem. Phys. **147**, 114112 (2017).
[32] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).
[33] M. Ceriotti, G. A. Tribello, and M. Parrinello, J. Chem. Theory Comput. **9**, 1521 (2013).
[34] L. Kahle, A. Musaelian, N. Marzari, and B. Kozinsky, Phys. Rev. Mater. **3**, 055404 (2019).
[35] M. W. Mahoney and P. Drineas, Proc. Natl. Acad. Sci. U. S. A. **106**, 697 (2009).
[36] I. T. Jolliffe and J. Cadima, Philos. Trans. R. Soc., A **374**, 20150202 (2016).
[37] A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, and S. Goedecker, J. Chem. Phys. **139**, 184118 (2013).
[38] S. Goedecker, J. Chem. Phys. **120**, 9911 (2004).
[39] B. Aradi, B. Hourahine, and T. Frauenheim, J. Phys. Chem. A **111**, 5678 (2007).
[40] A. J. Stone and D. J. Wales, Chem. Phys. Lett. **128**, 501 (1986).
[41] A. Stukowski, Modell. Simul. Mater. Sci. Eng. **18**, 015012 (2009).
[42] M. Ceriotti, M. J. Willatt, and G. Csányi, "Machine-learning of atomic-scale properties based on physical principles," in *Handbook of Materials Modeling: Methods: Theory and Modeling*, edited by W. Andreoni and S. Yip (Springer, Cham, 2020), p. 1911.
[43] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[44] A. Khorshidi and A. A. Peterson, Comput. Phys. Commun. **207**, 310 (2016).
[45] N. Bernstein, G. Csanyi, and J. Kermode, Quip and Quippy Documentation.
[46] K. Momma and F. Izumi, J. Appl. Crystallogr. **44**, 1272 (2011).