Metrics for measuring distances in configuration spaces

Ali Sadeghi,[1,a)] S. Alireza Ghasemi,[1,b)] Bastian Schaefer,[1] Stephan Mohr,[1] Markus A. Lill,[2] and Stefan Goedecker[1]
[1]*Department of Physics, Universität Basel, Klingelbergstr. 82, 4056 Basel, Switzerland*
[2]*Department of Medicinal Chemistry and Molecular Pharmacology, College of Pharmacy, Purdue University, 575 Stadium Mall Drive, West Lafayette, Indiana 47907, USA*

(Received 10 September 2013; accepted 15 October 2013; published online 14 November 2013)

In order to characterize molecular structures we introduce configurational fingerprint vectors which are counterparts of quantities used experimentally to identify structures. The Euclidean distance between the configurational fingerprint vectors satisfies the properties of a metric and can therefore safely be used to measure dissimilarities between configurations in the high dimensional configuration space. In particular we show that these metrics are a perfect and computationally cheap replacement for the root-mean-square distance (RMSD) when one has to decide whether two noise contaminated configurations are identical or not. We introduce a Monte Carlo approach to obtain the global minimum of the RMSD between configurations, which is obtained from a global minimization over all translations, rotations, and permutations of atomic indices. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4828704]

I. INTRODUCTION

Quantifying dissimilarities between molecular structures is an essential problem encountered in physics and chemistry. Comparisons based on structural data obtained either from experiments or computer simulations can help identifying or synthesising new molecules and crystals. A broad diversity of structures can only be obtained if identical configurations are eliminated. It is therefore highly desirable to have numerically affordable fingerprints that allow in a reliable way to detect identical configurations in the presence of noise which can either arise from experimental measurements or from structural relaxations in numerical simulations. Maintaining a broad diversity of structures is also a prerequisite for efficiency in any structure prediction method in material science and solid state physics[1–5] and conformer search in structural biology and drug discovery.[6–12] In the latter case, most of the proposed approaches[13–16] use approximate methods that reduce the structure description information, e.g., by excluding the side chains in a protein or a two-dimensional representations of the molecule,[17] to speed up the searching procedure.[18] In the case of solid state physics fairly accurate dissimilarity measures are required. Within the structure prediction methods based on the evolutionary algorithms,[1] the required diversity of populations can only be maintained if strongly similar configuration are eliminated. Within the Minima Hopping structure prediction method,[2] an identification of identical configurations is required as well to prevent trapping in funnels that do not contain the global minimum. Some machine learning approaches[19] are also based on similarity measures.

It is natural to characterize the dissimilarity between two structures p and q by a real number $d(p, q) \geq 0$. In order to give meaningful results $d(p, q)$ should satisfy the properties of a metric, namely,

- coincidence axiom: $d(p, q) = 0$ if and only if $p \equiv q$,
- symmetry: $d(p, q) = d(q, p)$,
- triangle inequality: $d(p, q) + d(q, r) \geq d(p, r)$.

The coincidence axiom ensures that two configurations p and q are identical if their distance is zero, and vice versa. The triangle inequality is essential for clustering algorithms. If it is not satisfied, then it could happen that a configuration that belongs to one cluster in configuration space is also part of another cluster even though the distance between the two clusters is very large in the configuration space.

Since measuring distances between configurations is required in many applications, a considerable effort has been made to find cheap, yet reliable, distance measures that are not affected by the alignment of the two structures whose distance is being measured and by the indexing of the atoms in the structures. In the field of chemoinformatics a large number of different descriptors have been proposed to establish relations between structure and functionality.[20] For example, a structure can be represented by a binary string whose elements are set depending on whether some specific patterns exist in the structure. Then the similarity between structures is described by the Tanimoto coefficient.[16,21] Another class of approaches is based on a generalizations of standard physical descriptors such as coordination numbers. Cheng and Fournier[22] used, for instance, the statistical properties (average, variance, and bounds) of the coordination numbers while Lee *et al.*[23] used their weighted histograms in order to characterize the structures. Histogram-based methods were also used for the identification of crystalline structures.[24] All these methods have several tuning parameters such as the width of histogram bins or cutoff radii for the determination of coordination numbers[23] and their performance can critically depend on the choice of these parameters.

a)ali.sadeghi@unibas.ch
b)Present address: Institute for Advanced Studies in Basic Sciences, P.O. Box 45195-1159, Zanjan, Iran.

In this article we will introduce a family of parameter free metrics for measuring distances in configuration spaces. We show that these metrics fulfil all the mathematical requirements and demonstrate their excellent performance for a representative set of benchmark systems including covalent, metallic (simple or transition), ionic, and organic structures. In the case of periodic systems, additional complexity comes into play because of non-uniqueness of the elementary cell. In the present work, our focus is on isolated molecules. The configurations in our test set are metastable low energy configurations obtained during a structure search using the Minima Hopping method[2] on the density functional theory (DFT) level as implemented in the BigDFT code.[25]

II. RMSD

A configuration of n alike atoms is uniquely represented by $\boldsymbol{R} \equiv (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n) \in \mathbb{R}^{3\times n}$, where the column vector $\boldsymbol{r}_i$ represents the Cartesian coordinates of atom i. A distance based on the naive Frobenius norm

$$\|\boldsymbol{R}^p - \boldsymbol{R}^q\| = \left(\sum_{i=1}^{n} \|\boldsymbol{r}_i^p - \boldsymbol{r}_i^q\|^2\right)^{1/2} \tag{1}$$

cannot be used to compare two configurations p and q, because it is not invariant with respect to translations or rotations of one configuration relative to the other. For this reason the commonly used root-mean-square distance (RMSD) is defined as the minimum Frobenius distance over all translations and rotations. By minimizing $\sum_i^n \|\boldsymbol{r}_i^p + \boldsymbol{d} - \boldsymbol{r}_i^q\|^2$ with respect to the translation $\boldsymbol{d}$ one obtains $\sum_i^n (\boldsymbol{r}_i^p + \boldsymbol{d} - \boldsymbol{r}_i^q) = 0$, i.e., the required translation is the difference between the centroids $\boldsymbol{d} = \frac{1}{n}\sum_i^n \boldsymbol{r}_i^q - \frac{1}{n}\sum_i^n \boldsymbol{r}_i^p$. Therefore, we will assume in the following that all $\boldsymbol{r}_i$ are measured with respect to the centroids of the corresponding configuration which allows us to drop the minimization with respect to the translation $\boldsymbol{d}$. Then, finding the rotation U around the common centroid which minimizes

$$RMSD_l(p,q) = \frac{1}{\sqrt{n}} \min_U \|\boldsymbol{R}^p - U\boldsymbol{R}^q\| \tag{2}$$

is a local minimization problem and hence we denote this version of the RMSD by RMSD$_l$. The Kabsch algorithm[26] provides the solution to this problem based on the Euler angles. Like many others, we perform the local minimization by an alternative method based on quaternions[27] (see Appendix A) which is more stable and numerically very cheap.[28,29]

The RMSD$_l$ is, however, not invariant under index permutations of chemically identical atoms. If the configuration p and q are identical, Eq. (2) will be different from zero if we permute, for instance, in $\boldsymbol{R}^q$ the positions $\boldsymbol{r}_i^q$ and $\boldsymbol{r}_j^q$ of atoms i and j. The minimum Frobenius distance obtained by considering all possible index permutations for an arbitrary rotation U is

$$RMSD_P(p,q) = \frac{1}{\sqrt{n}} \min_P \|\boldsymbol{R}^p - U\boldsymbol{R}^q P\|, \tag{3}$$

where P is an $n \times n$ permutation matrix. This assignment problem is solved in polynomial time using the Hungarian algorithm.[30]

What is really needed is a solution of the combined problem of the global minimization over all rotations and permutations, namely,

$$RMSD(p,q) = \frac{1}{\sqrt{n}} \min_{P,U} \|\boldsymbol{R}^p - U\boldsymbol{R}^q P\|. \tag{4}$$

The global minimum RMSD fulfills all the properties of a metric. The coincidence and symmetry properties are easy to see. Using the standard triangle inequality, the proof of the triangle property is as follows:

$$\begin{aligned}
&RMSD(p,q) + RMSD(q,r)\\
&= \frac{1}{\sqrt{n}} \min_{P,U} \|U\boldsymbol{R}^p P - \boldsymbol{R}^q\| + \frac{1}{\sqrt{n}} \min_{P,U} \|\boldsymbol{R}^q - U\boldsymbol{R}^r P\|\\
&= \frac{1}{\sqrt{n}} \|U_{pq}\boldsymbol{R}^p P_{pq} - \boldsymbol{R}^q\| + \frac{1}{\sqrt{n}} \|\boldsymbol{R}^q - U_{rq}\boldsymbol{R}^r P_{rq}\|\\
&\geq \frac{1}{\sqrt{n}} \|U_{pq}\boldsymbol{R}^p P_{pq} - \boldsymbol{R}^q + \boldsymbol{R}^q - U_{rq}\boldsymbol{R}^r P_{rq}\|\\
&\geq \frac{1}{\sqrt{n}} \|\boldsymbol{R}^p - U_{rp}\boldsymbol{R}^r P_{rp}\|\\
&= RMSD(p,r),
\end{aligned}$$

where $\min_{P,U} \|U\boldsymbol{R}^p P - \boldsymbol{R}^q\|$ is shown by $\|U_{pq}\boldsymbol{R}^p P_{pq} - \boldsymbol{R}^q\|$ for convenience.

Since U and P are not independent, no algorithm exists which can find the global RMSD within polynomial time. Just doing a search by alternating rotation and permutation steps using local minimizations and the Hungarian algorithm, respectively, is not guaranteed to converge to the global minimum with a finite number of steps. Trying out all possible permutations would lead to a factorial increase of the computing time with respect to n and this approach is therefore not feasible except for very small systems. In some applications, one might apply restrictions into the permutations in order to reduce the size of the permutation space. For instance, in an application to organic molecules only equivalent atoms has to be permuted, e.g., see Ref. 31. Equivalent atoms in an organic molecule are considered for example those that have identical connectives determined by the Morgan algorithm.[32,33] For all kind of molecular structures, however, such a grouping of identical atoms ones is not possible.

We use a two-stage method for finding the global RMSD with moderate computational effort. The flowchart of the algorithm is depicted in Fig. 1 with the two different stages shown on the left and right sides. In the first stage we try to find the optimal global alignment of the two structures being compared. We first align two of the three principal axes of inertia of one configurations with the corresponding axes of the other one. A trial alignment is always followed by the application of the Hungarian algorithm to find the index permutation that gives the smallest RMSD.[34] The index matching in the Hungarian algorithm is done in the Cartesian space by associating to each atom $i \in$ p the closest atom $j \in$ q such that $\sum_i^n \|\boldsymbol{r}_i^p - \boldsymbol{r}_j^q\|$ is minimal. In other words, the columns of the $n \times n$ matrix made by $\|\boldsymbol{r}_i^p - \boldsymbol{r}_j^q\|$ are reordered such that its trace is minimal. The implementation of the Hungarian algorithm based on Ref. 35 finds the optimal index permutation

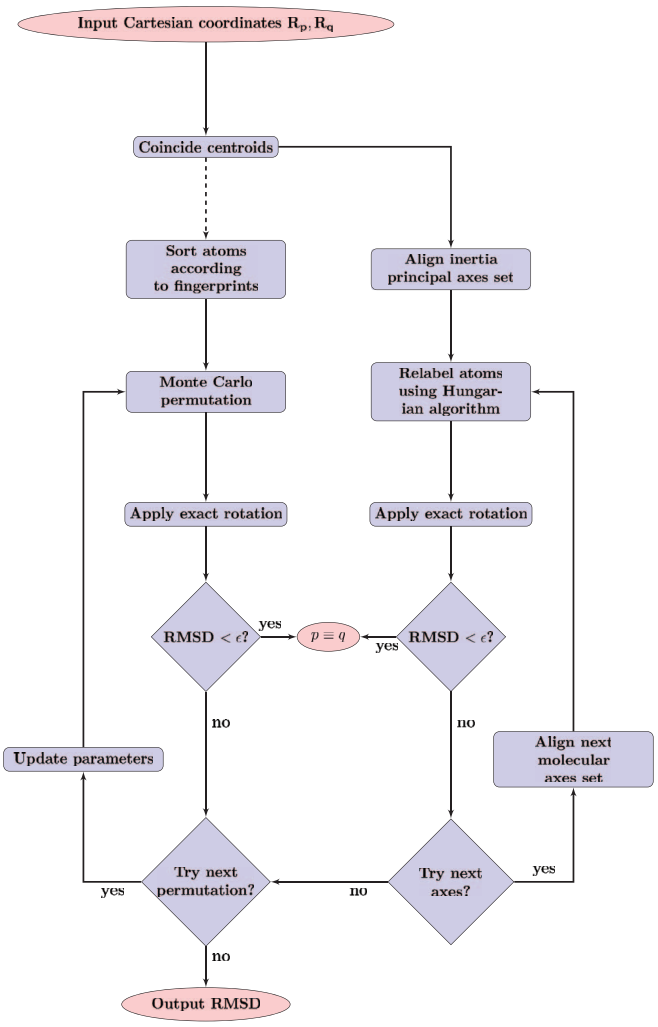

FIG. 1. Flowchart of the algorithm of global minimization of RMSD in two major steps. The loop on the right runs over several sets of axes and matches atoms of a pair of configurations via aligning their molecular axes. The left loop shows the Monte Carlo (MC) permutation of identical particles while the parameters are dynamically tuned to obtain an acceptance rate close to 50%. The dashed line means that the right loop can be excluded.

within polynomial time and with a small prefactor. After this initial index matching, a rotation using quaternions is applied to refine the molecular alignment. If the required rotation is significant, the atomic index assignment should be repeated. This whole procedure is iterated until the atomic indices remain fixed after applying the rotation. This procedure has allowed us to detect all identical configuration in this first stage, as seen in Table I.

Since all the global alignment methods are empirical and can fail we apply several of them successively. After the first global alignment based on the principal axes of inertia we apply some more alignments steps based on axes which are derived from local atomic fingerprints. We set up an overlap matrix with s- and p-type Gaussian orbitals (see Appendix B) and find its principal eigenvector (i.e., the eigenvector with the largest eigenvalue; see Fig. 7). Defining $\boldsymbol{w}_i = s_i \boldsymbol{p}_i$, where s_i and $\boldsymbol{p}_i$ are, respectively, s- and p-type components of the principal eigenvector belonging to atom i we can form two axes $\boldsymbol{W}$ and $\boldsymbol{W}'$

$$\boldsymbol{W} = \sum_i^n \boldsymbol{w}_i, \tag{5}$$

$$\boldsymbol{W}' = \sum_i^n \boldsymbol{w}_i \times \boldsymbol{r}_i, \tag{6}$$

where the sum runs over the atoms, $\boldsymbol{r}_i$ represents the positions of atoms with respect to the center of mass and $\times$ denotes the cross product. First, we align $\boldsymbol{W}^q$ with $\boldsymbol{W}^p$ and then rotate q around it such that the plane made by $(\boldsymbol{W}^q, \boldsymbol{W}'^q)$ coincides with the plane made by $(\boldsymbol{W}^p, \boldsymbol{W}'^p)$. Depending on the width of the Gaussian used to construct the overlap matrix, several sets of axes may be constructed and tried one-by-one in this stage. If the alignment according to a new set of axes results in a smaller RMSD, we accept it. In Table I we show the results of the alignment of the principal axes of inertia as well as three sets of $(\boldsymbol{W}, \boldsymbol{W}')$ axes obtained by three different Gaussian widths α.

If a small enough RMSD is not found, we enter into an iterative stage (see left side of Fig. 1) where randomly chosen atoms are permuted within a thresholding Monte Carlo (MC) approach followed by applying the optimal rotation. In the thresholding MC step, two chemically identical atoms are selected according to a uniform random distribution. If by swapping them the RMSD is reduced, the permutation is accepted. To exclude the possibility of getting stuck in a local minimum, the permutation is also accepted if it causes the RMSD to increase by less than an adjustable parameter ξ. This parameter is dynamically updated at each step: if the acceptance rate so far is less/greater than 50%, then ξ is increased/decreased by a factor of 1.1. In this way, the average acceptance rate approaches 50% during the minimization. The iteration stops when the global minimum RMSD does not decrease any more for a large number of iterations.

As seen in Table I, the number of required MC iterations depends on the system size. For instance, for the biomolecule 10^4 MC iterations (which take on average 0.13 s on a single 2.4 GHz Intel core) are sufficient to find the global minimum RMSD between two configurations of this molecule. For a more systematic investigation of the scaling, we take the global minima of the Lenard-Jones (LJ) clusters with different sizes and apply random displacements of the unit magnitude to every atom (i.e., the RMSD between the randomized structures is almost one in the LJ length units). The averaged number of required MC iterations to get the asymptotic value of the RMSD (as obtained by 10^7 iterations), as a function of the cluster size n is shown in Fig. 2. Even though the number of iterations increases exponentially it is several orders of magnitude smaller than the number of possible permutations, i.e., $n!$.

III. FINGERPRINT DISTANCES AS METRICS

While the RMSD can be considered as the most basic quantity to measure the dissimilarities, finding the global minimum RMSD is numerically costly. Only in case that two structures are nearly identical the global minimum of RMSD is calculated with a polynomial computational time because

TABLE I. Number of remaining distinct configurations, average RMSD and average CPU-time (on single 2.4 GHz Intel core) at different steps of the two-stage RMSD global minimization. All the test sets consist of the energetically lowest metastable configurations on the DFT level obtained in a Minima Hopping run. In the first stage (Axes Alignment, AA), the principal axes of inertia as well as three molecular sets of axes obtained from vectorial atomic fingerprints are used (the Gaussian widths are $\alpha_1 = 0.5d_c^{-2}$, $\alpha_2 = \alpha_1/1.21$ and $\alpha_3 = \alpha_1/1.44$, where d_c is the sum of covalent radii of the two atoms; see Appendix B). Every molecular alignment is always followed by the application of the Hungarian algorithm to find the optimal index permutation. In the second stage (Monte Carlo, MC), random permutations are tried out which are followed by local minimization to get the optimal rotation. Because of the stochastic nature of the MC part, the reported values might change in different runs.

		Si_{32}			Mg_{26}			$C_{22}H_{24}N_2O_3$		
		remaining distinct	$\overline{\text{RMSD}}$ (Å)	$\bar{t}_{\text{CPU}}$ (s)	remaining distinct	$\overline{\text{RMSD}}$ (Å)	$\bar{t}_{\text{CPU}}$ (s)	remaining distinct	$\overline{\text{RMSD}}$ (Å)	$\bar{t}_{\text{CPU}}$ (s)
	Unanalyzed	317	1.40		111	3.44		60	2.75	
	Inertia axes	184	1.16		60	1.08		42	1.93	
	$(\boldsymbol{W}, \boldsymbol{W}')_{\alpha_1}$	184	1.06		59	1.06		42	1.89	
AA	$(\boldsymbol{W}, \boldsymbol{W}')_{\alpha_2}$	184	1.04		59	1.03		42	1.81	
	$(\boldsymbol{W}, \boldsymbol{W}')_{\alpha_3}$	184	1.02	<0.001	59	1.01	<0.001	42	1.78	<0.001
	iter.=10^3	184	0.978	0.03	59	0.985	0.02	42	1.52	0.05
	iter.=10^4	184	0.910	0.13	59	0.864	0.11	42	1.51	0.13
MC	iter.=10^5	184	0.852	1.1	59	0.852	1.0	42	1.51	1.6
	iter.=10^6	184	0.792	12.1	59	0.824	10	42	1.51	15
	iter.=10^7	184	0.791	132	59	0.824	119	42	1.51	163

no MC permutation is then required. Otherwise, even if the above described algorithm is used, the computational time increases exponentially with the number of permutable particles. In the following we will therefore introduce a family of metrics which are cheaper to calculate than the global RMSD yet in good agreement with it. We consider symmetric $N \times N$ matrices whose elements depend only on the interatomic distances $r_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|$ of an n-atom configuration. Vectors $\boldsymbol{V}$ containing eigenvalues of such a matrix form a configurational fingerprint which allows to identify a structure. The normalized Euclidean distance

$$\Delta_V(p, q) = \frac{1}{\sqrt{N}} \|\boldsymbol{V}^p - \boldsymbol{V}^q\| \tag{7}$$

measures the dissimilarly between p and q with no need to superimpose them.[36] Since the matrix depends only on interatomic distances, the same holds true for the eigenvalues, and $\boldsymbol{V}$ is thus invariant under translations, rotations and reflections of the configuration. In order to make Δ_V also independent of the atomic indices, the elements of each $\boldsymbol{V}$ are sorted in an ascending order. This sorting can introduce discontinuities in the first derivative of the fingerprint distance with respect to changes in the atomic coordinates (e.g., when there is a crossing of eigenvalues) but does not destroy the important continuity of the fingerprint distance itself.

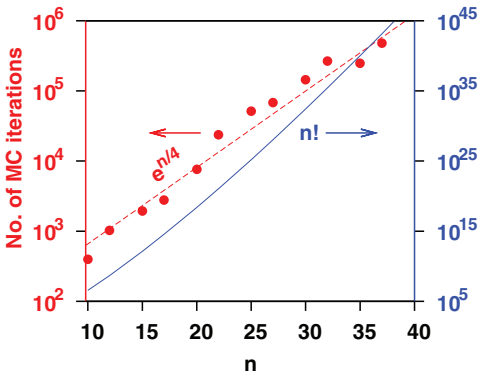

FIG. 2. Average number of the MC iterations required to obtain the global RMSD between randomized LJ clusters as a function of the number of particles n. The dashed line ($41 \exp(n/4)$) is obtained by least square fit. For comparison, $n!$ is also plotted with solid line.

The coincidence axiom for a configurational fingerprint is satisfied if the dimension N of the matrix is sufficiently large and if therefore the resulting fingerprint vector is sufficiently long. We show in Appendix C that how a hypersurface of constant fingerprint can be constructed if the length of the fingerprint is short. What we would like to show, however, is the opposite, namely that no distinct configurations with identical fingerprints exist if the fingerprint is long enough. Since the fingerprint distance is a nonlinear function, it can in principle not be excluded that two distinct configurations with identical fingerprints exist even if the fingerprint vector is longer than the threshold value. Since we recommend for a unique identification fingerprints which are considerably longer than the threshold value, namely, fingerprints of length $3n$ or even $4n$ it is, however, extremely unlikely that such configurations exist and the coincidence axiom can be taken to be fulfilled. To confirm this assumption numerically as well, we did extensive numerical searches where we tried to find a second configuration which has a fingerprint which is identical to the fingerprint of a reference configuration. The initial guess for the second configuration was random and then this second configuration was moved in such a way as to minimize the difference between the fingerprints. All these numerical minimizations lead to non-zero local minima, i.e., we were not able to find numerically any violation of the coincidence axiom for vectors of length $3n - 3$ based on the Hessian matrix and vectors of length $4n$ based on an overlap matrix with s and p orbitals.

Even though the eigenvalue vector is much shorter than the vector containing all matrix elements, the fingerprint distances based on the eigenvalues are better than those obtained

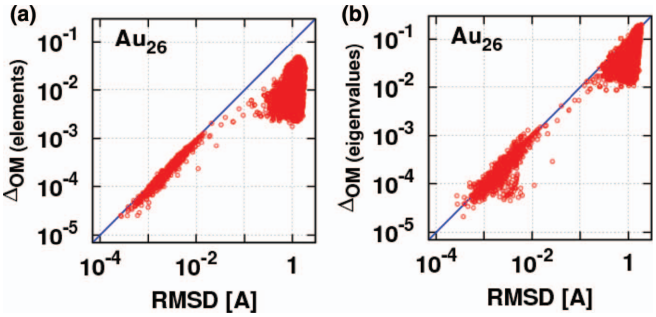

FIG. 3. Correlation of the pairwise Euclidean distances based on vectors consisting either of all of the sorted elements of the overlap matrix (a) or eigenvalues of this matrix (b) and the RMSD for 1000 metastable configurations of a 26 atom gold cluster. The gap in the fingerprint distances between identical and distinct configuration is larger if eigenvalues are used (panel a).

by sorting all the matrix elements depending on interatomic distances into a vector. One can in some cases construct distinct so-called homometric configurations[37] for which the fingerprint vectors of the sorted matrix elements are identical whereas the eigenvalue vectors are not identical and allow thus to distinguish between them. In addition, our empirical results of Fig. 3 show that the gap between identical and distinct pairs is larger for the eigenvalues than for the sorted matrix elements. Because the geometry relaxations were stopped when the force on each atom is within 0.01 eV/Å, identical configurations are in practice identical only up to some finite precision, i.e., the atomic positions of the configurations are contaminated by noise. Two configurations are considered to be identical if their distance is below a certain threshold. An unambiguous threshold for distinguishing between distinct and non-distinct configurations can only be found if a well detectable gap exists in the distance space. Hence, the existence of a large gap is an important benefit of a fingerprint method.

In an application to Ni clusters Grigoryan and Springborg[38] used the sorted interatomic distances to find the similarities between an $(n-1)$-atom cluster and $(n-1)$-atom parts of an n-atom cluster. This similarity measure also leads to a gap which is smaller than the one obtained from eigenvalue based fingerprints of either the corresponding r_{ij} matrix or the matrices proposed in this article (cf. Figs. 4 and 5). So it seems to be a general feature that fingerprints based on

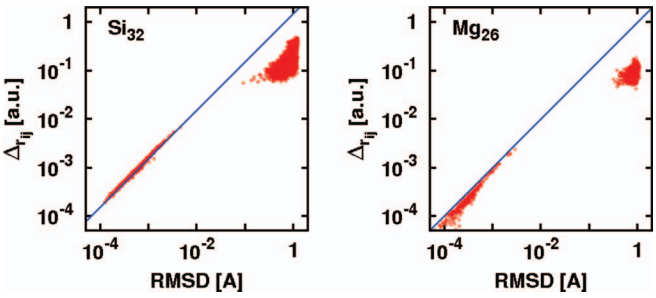

FIG. 4. Correlation of Euclidean distance of the sorted interatomic distances with RMSD for the metastable configurations of the Si_{32} and Mg_{26} clusters. The gap that allows to discriminate distinct from non-distinct configurations is smaller in both cases compared to the fingerprints based on eigenvalues.

the eigenvalues are better than those based on sorted matrix elements.

In the following we will describe several matrix constructions which can be used for fingerprinting. These matrices are closely related to measurable quantities that are traditionally used by experimentalists to identify structures.

A. Hamiltonian matrix

Emission and absorption spectra arise from transitions between discrete electronic energy levels. Each element has its characteristic energetic levels and therefore atomic spectra can be used as elemental fingerprints. When atoms are assembled into structures the electronic states of the constituent atoms are modified depending on the arrangement of the atoms. A computational analogue to electronic energy levels probed by various spectroscopic experiments are the Kohn-Sham energy eigenvalues, even though they do not represent the physical excitation energies. Since the Kohn-Sham Hamiltonian matrix depends only on the interatomic distances, the sorted Kohn-Sham eigenvalues are invariant to translations, rotations, reflections, and permutations of atoms.

We examine fingerprints that are based on the occupied Kohn-Sham eigenvalues only as well as fingerprints that are based both on the occupied and unoccupied eigenvalues. The former were obtained from the self-consistent eigenvalues calculated in a large wavelet basis,[25] whereas, for simplicity, the latter were obtained from the non-self-consistent input guess eigenvalues calculated in a minimal Gaussian type atomic orbitals (GTO's) basis set for a charge density which is a superposition of atomic charge densities. Even though the length requirement of the coincidence axiom is violated in all cases, the configurational distances $\Delta_{KS}(p, q)$ obtained from the occupied Kohn-Sham eigenvalues correlates with the RMSD for the five test sets, see Fig. 5. Fingerprint distances based on the vector $\boldsymbol{V}_{GTO}$ do not much better correlate with the RMSD than fingerprint distances based on $\boldsymbol{V}_{KS}$, even though the vector $\boldsymbol{V}_{GTO}$ is in all cases longer than the vector $\boldsymbol{V}_{KS}$ (e.g., $4n$ in case of the Si cluster, i.e., two times longer) and hence the coincidence axiom is satisfied in all cases. Notice that different distances measure different kinds of dissimilarities, and it is therefore not expected that they correlate well for large distances. All metrics should, however, be in agreement concerning the detection of distinct and non-distinct pairs. This requirement is satisfied in all scatter plots shown in Fig. 5. It has also to be stressed that no specific metric, including the RMSD, is *a priori* the best dissimilarity measure for all applications. Configurations belonging to the same structural motif are, for instance, not necessarily close in RMSD distance.

B. Overlap matrix

A matrix which has similar properties as the Hamiltonian matrix is the overlap matrix expressed in terms of GTO's. Contrary to the Hamiltonian, all elements of the overlap matrix can easily be calculated analytically (Appendix B). In the simplest case where only uncontracted s-type GTO's are

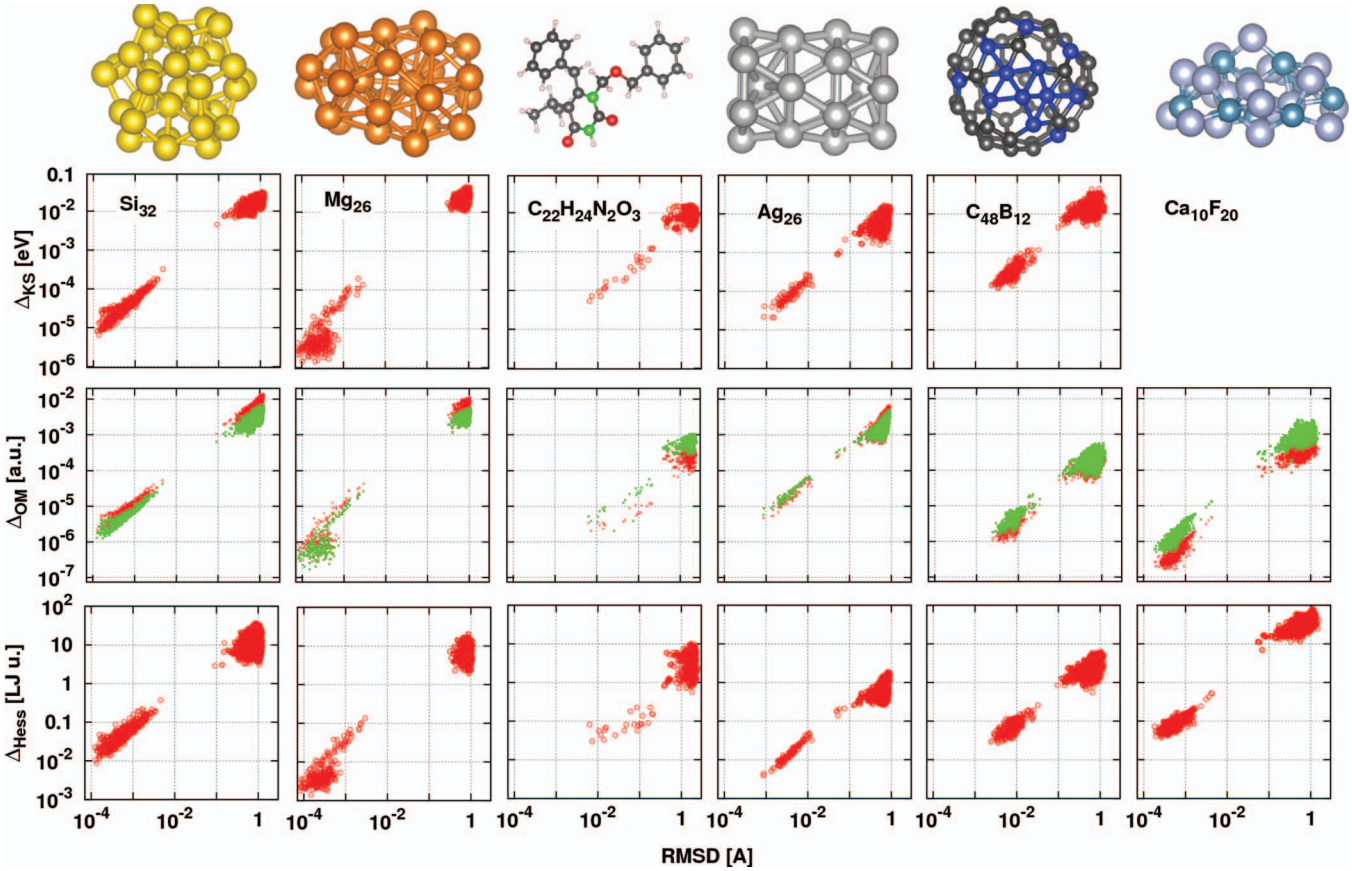

FIG. 5. Comparison of fingerprint distances based on the eigenvalues of the Kohn-Sham Hamiltonian matrix (first row), the overlap matrix (second row), and the Lennard-Jones Hessian matrix with the RMSD for sets of semiconductor (silicon), simple metal (magnesium), organic (6-benzyl-1-benzyloxymethyl-5-isopropyl uracil), transition metal (silver), covalent fullerene-type ($C_{48}B_{12}$), and ionic (calcium fluoride) clusters. Shown on top are representative configurations. Each set consists of a few hundred configurations, all being low-energy local minima within DFT, except those of $Ca_{10}F_{20}$ which are local minima of the Tosi-Fumi potential (parameters from Ref. 39). For the latter system the Kohn-Sham eigenvalues are obviously not calculated. For the five sets where Kohn-Sham eigenvalues can be calculated their number is determined by the occupied valence states and is given, respectively from left to right, by $64 = 2n$, $26 = n$, $n < 70 < 2n$, $26 = n$, and $n < 114 < 2n$, n being the number of atoms. For the overlap matrix, results for both s-only (red) and s-and-p (green) overlap matrices are shown, leading to fingerprint vectors of lengths n and $4n$, respectively. For the Hessian matrix $3n - 3$ eigenvalues are non-zero. Even in the cases where the length of the fingerprint vector is shorter than $3n - 6$ the agreement with the RMSD is good and allows always to identify distinct and non-distinct configurations.

used, the resulting fingerprint consists of n scalars. Information about the radial distribution can be incorporated in the overlap matrix by adding p and d type GTO's. In this way the configurational fingerprint vector becomes also longer than $3n - 6$ and the coincidence axiom will be satisfied.

If the fingerprint is used to calculate distances between our test set of local minima configurations, it turns out that adding p-type orbitals gives only a marginal improvement, in the sense that the distance gap separating identical and distinct configurations gets larger. Adding additional d-type orbitals has virtually no effect. This is related to the fact that it is very unlikely that two local minima lie on the hypersurface that leaves the fingerprint invariant (see Appendix C). The width of the GTO's was in all our tests given by the covalent radius of the atom on which the GTO was centered.

C. Hessian matrix

The vibrational properties, which are frequently used experimentally to identify structures, are closely related to the Hessian matrix which consists of the second order derivatives of the energy with respect to the atomic positions. The vibrational frequencies are up to a scaling factor related to the mass of the atoms equal to the square root of the eigenvalues of the Hessian matrix. This matrix also belongs to the class of matrices with the desired properties. Unfortunately, the calculation of the Hessian is rather expensive in the context of a DFT calculation and can also be cumbersome with sophisticated force fields. We will therefore not further pursue approaches based on an Hessian which is calculated within the same high level method as the energy and forces. It, however, turns out that eigenvalues or eigenvectors of the Hessian matrices which are derived from another cheaper potential such as the LJ potential give also good fingerprints. This is shown in Fig. 5 for our six test systems after the lengths were scaled to the equilibrium bond length of the LJ potential.

D. Discussion

Various $n \times n$ matrices, have been used previously to characterize molecular configurations. The definition of a molecular descriptor can be based on either eigenvalues, spectral moments (defined as the kth power of the eigenvalues, where the natural number $k \leq n$ is then the order of the

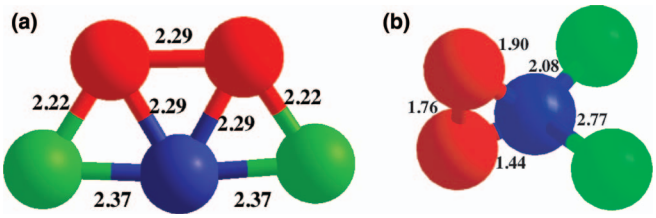

FIG. 6. Two distinct configurations of the Si_5 cluster with an identical set of SPRINT coordinates, i.e., 3.59 (green), 4.37 (red), 4.85 (blue), using the parameters given in the supplementary material of Ref. 40. The planar structure shown in (a) is a local minimum in DFT, but that in (b) is not. The numbers show the bond lengths in Å.

moment) or even the elements of the eigenvector associated with the largest eigenvalue (i.e., the principal eigenvector) of many matrices, e.g., adjacency, Laplacian, distance, reciprocal distance, distance-path, etc.; for review see Ref. 20. The contact matrix from the graph theory exhibits discontinuities when the atomic distances cross the cutoff radius. By introducing a smooth cutoff these discontinuities disappear and the resulting matrix has been used as a fingerprinting tool in the SPRINT method.[40] Presumably not only the contact matrix but also other matrices from spectral graph theory such as the Laplace matrix could be used in a similar way. We did, for instance, not find significant differences in performance between the contact and Laplacian matrices. We found, however, that fingerprints based on either of them are rather sensitive to the form of the smooth cutoff function. Tuning of the parameters of this cutoff function is therefore required to obtain good results. In both cases, the resulting atomic fingerprints are real scalars which mostly contain information about the number of nearest neighbours of each atom and might be insufficient to characterize the chemical environment of an atom. Better chemical environment descriptors can however be obtained by adding information about the radial distribution of the neighbours.[41,42] The Coulomb matrix is another matrix whose eigenvalues have been used to characterize configurations.[19] The off-diagonal elements of this n-by-n matrix are the pairwise Coulomb repulsions $q_i q_j / r_{ij}$, while the diagonal is filled with $q_i^{2.4}/2$, q_i being the core charge of atom i.

As discussed before, such a fingerprint of length n is not long enough to satisfy the coincidence axiom and can thus fail to detect structural differences. This has already been shown for the Coulomb matrix.[43] We show in Fig. 6 two distinct configurations of a Si_5 cluster which have identical sets of SPRINT coordinates. Note that the Si atoms with identical SPRINT coordinates in the configuration shown in Fig. 6(b), have very different environments. This shows that SPRINT, like any other $n \times n$ matrix-based fingerprint, fails to describe uniquely the entire structure and/or the chemical environment of an atom.

IV. CONCLUSIONS

In summary, we have shown that the RMSD, the most natural measure of dissimilarity between two configurations, satisfies the properties of a metric when it is obtained by a global minimization over all rotations and index permutations. We have presented a Monte Carlo method to calculate the global minimal RMSD which does not require to try out all possible index permutations and which is thus computationally feasible. At the same time we have introduced fingerprints which are much cheaper to calculate because they do not require a structural superposition. Nevertheless, the fingerprint based distances correlate in all our test cases with the RMSD, in the sense that small RMSD distances correspond to small fingerprint distance and vice versa. In contrast to numerous previously proposed fingerprints they satisfy the coincidence axiom and allow therefore to distinguish distinct from non-distinct configurations in a unique way. Within a DFT calculation the metric based on the Kohn-Sham eigenvalues is a good choice since the eigenvalues are a byproduct of any DFT calculation and thus no extra effort is required to obtain them. For the coincidence axiom to be satisfied, the number of bound eigenstates whose Kohn-Sham eigenvalues can be included in the fingerprint vector has, however, to be larger than $3n - 6$. If Kohn-Sham eigenvalues are not available, the method based on the eigenvalues of the overlap matrix constructed from s and p orbitals is recommended, since it leads to matrices whose elements can be calculated analytically and because the fingerprint vector is long enough ($4n$) to make the probability of a violation of the coincidence axiom vanishingly small. Even if the coincidence axiom is violated, it turns out in practice that it is very rare that different physically reasonable metastable configurations give rise to identical fingerprints. For our test sets of low energy local minima configurations metrics which violated the coincidence axiom therefore allowed nevertheless in all cases to distinguish between distinct and non-distinct configurations. In other applications where small movements away from metastable configurations lead to a change of physical properties, such as in force fields based on machine learning, a violation of the coincidence theorem cannot, however, be tolerated. All the proposed variants of our approach are parameter free and no parameter tuning is therefore required.

ACKNOWLEDGMENTS

We gratefully thank D. G. Kanhere, S. De, R. Schneider, and R. Ebrahimian for interesting and helpful discussions. This work has been supported by the Swiss National Science Foundation (SNF) and the Swiss National Center of Competence in Research (NCCR) on Nanoscale Science. Structures were visualized using V_Sim[44] and VESTA[45] packages. Computing time was provided by the CSCS.

APPENDIX A: CLOSED-FORM OF SUPERIMPOSING ROTATION

A quaternion $\mathcal{Q} = (\mathcal{Q}_0, \mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3)$ is an extension of the idea of complex numbers to one real ($\mathcal{Q}_0$) and three imaginary parts. According to the Euler's rotation theorem, a rotation in space which keeps one point on the rigid body (centroid in our case) fixed, can be represented by four real numbers: one for the rotation angle and three for the rotation axis (we assume that the center of rotation is at the origin). A unit quaternion, i.e., $\|\mathcal{Q}\|^2 = \mathcal{Q}_0^2 + \mathcal{Q}_1^2 + \mathcal{Q}_3^2 + \mathcal{Q}_4^2 = 1$, can

conveniently represent this axis-angle couple as

$$\mathcal{Q} = \left(\cos\left(\frac{\theta}{2}\right), \hat{\mathbf{u}}\sin\left(\frac{\theta}{2}\right)\right),$$

where θ is the rotation angle around the unit axis $\hat{\mathbf{u}} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$. The corresponding orthogonal rotation matrix is

$$U = \begin{bmatrix} \mathcal{Q}_0^2 + \mathcal{Q}_1^2 - \mathcal{Q}_2^2 - \mathcal{Q}_3^2 & 2\mathcal{Q}_1\mathcal{Q}_2 - 2\mathcal{Q}_0\mathcal{Q}_3 & 2\mathcal{Q}_1\mathcal{Q}_3 + 2\mathcal{Q}_0\mathcal{Q}_2 \\ 2\mathcal{Q}_1\mathcal{Q}_2 + 2\mathcal{Q}_0\mathcal{Q}_3 & \mathcal{Q}_0^2 - \mathcal{Q}_1^2 + \mathcal{Q}_2^2 - \mathcal{Q}_3^2 & 2\mathcal{Q}_2\mathcal{Q}_3 - 2\mathcal{Q}_0\mathcal{Q}_1 \\ 2\mathcal{Q}_1\mathcal{Q}_3 - 2\mathcal{Q}_0\mathcal{Q}_2 & 2\mathcal{Q}_2\mathcal{Q}_3 + 2\mathcal{Q}_0\mathcal{Q}_1 & \mathcal{Q}_0^2 - \mathcal{Q}_1^2 - \mathcal{Q}_2^2 + \mathcal{Q}_3^2 \end{bmatrix}. \quad \text{(A1)}$$

The optimum rotation U which minimizes RMSD, indeed maximizes the correlation between $\boldsymbol{R}^p$ and $\boldsymbol{R}^q$, i.e., the atomic Cartesian coordinates with respect to the common center of mass. Based on quaternions,[27] the optimum U is given by $\mathcal{Q}$ which is identical to the principal eigenvector of the 4×4 symmetric, traceless matrix

$$\mathcal{F} = \begin{bmatrix} \mathcal{R}_{xx} + \mathcal{R}_{yy} + \mathcal{R}_{zz} & \mathcal{R}_{yz} - \mathcal{R}_{zy} & \mathcal{R}_{zx} - \mathcal{R}_{xz} & \mathcal{R}_{xy} - \mathcal{R}_{yx} \\ \mathcal{R}_{yz} - \mathcal{R}_{zy} & \mathcal{R}_{xx} - \mathcal{R}_{yy} - \mathcal{R}_{zz} & \mathcal{R}_{xy} + \mathcal{R}_{yx} & \mathcal{R}_{xz} + \mathcal{R}_{zx} \\ \mathcal{R}_{zx} - \mathcal{R}_{xz} & \mathcal{R}_{xy} + \mathcal{R}_{yx} & -\mathcal{R}_{xx} + \mathcal{R}_{yy} - \mathcal{R}_{zz} & \mathcal{R}_{yz} + \mathcal{R}_{zy} \\ \mathcal{R}_{xy} - \mathcal{R}_{yx} & \mathcal{R}_{xz} + \mathcal{R}_{zx} & \mathcal{R}_{yz} + \mathcal{R}_{zy} & -\mathcal{R}_{xx} - \mathcal{R}_{yy} + \mathcal{R}_{zz} \end{bmatrix} \quad \text{(A2)}$$

where $\mathcal{R}$ is the correlation matrix whose elements are $\mathcal{R}_{xy} = \sum_i^n x_i^p y_i^q$ and so on. Eq. (2) is then given by

$$RMSD(p,q) = \sqrt{\frac{1}{n}(\|\boldsymbol{R}^p\|^2 + \|\boldsymbol{R}^q\|^2 - 2\lambda^*)}, \quad \text{(A3)}$$

where λ^* is the largest eigenvalue of $\mathcal{F}$.

APPENDIX B: OVERLAPS BETWEEN GTO's

The normalized GTO's centered at the atomic positions $\boldsymbol{r}_i$ in Cartesian coordinates are given by

$$\phi_i^{\boldsymbol{l}}(\boldsymbol{r}) = N_{\boldsymbol{l}}(x - x_i)^{l_x}(y - y_j)^{l_y}(z - z_i)^{l_z} e^{-\alpha_i \|\boldsymbol{r} - \boldsymbol{r}_i\|^2},$$

where $\boldsymbol{l} = (l_x, l_y, l_z)$ and $N_{\boldsymbol{l}}$ is the normalization factor. Depending on the angular moment $L = l_x + l_y + l_z$ the functions are labeled as s-type (L= 0), p-type ($L = 1$), d-type ($L = 2$), and so on. We take the Gaussian width α_i inversely proportional to the square of the covalent radius of atom i throughout this work.

The Gaussian product theorem says that the product of two Gaussian functions is again a Gaussian function. Therefore, the overlap integrals between a pair of GTO's, namely,

$$\langle \phi_i^{\boldsymbol{l}} | \phi_j^{\boldsymbol{l}'} \rangle = \int d\boldsymbol{r} \phi_i^{\boldsymbol{l}}(\boldsymbol{r}) \phi_j^{\boldsymbol{l}'}(\boldsymbol{r}) \quad \text{(B1)}$$

can be evaluated analytically. This gives the normalization factors as

$$N_{\boldsymbol{l}}(\alpha_i) = \frac{1}{\sqrt{\langle \phi_i^{\boldsymbol{l}} | \phi_i^{\boldsymbol{l}} \rangle}} = (2\alpha_i/\pi)^{3/4} \sqrt{n_{l_x} n_{l_y} n_{l_z}},$$

$$n_k = \frac{(4\alpha_i)^k}{(2k-1)!!}.$$

All GTO's are recursively obtained by differentiating

$$\phi_i^s(\boldsymbol{r}) = \left(\frac{2\alpha_i}{\pi}\right)^{3/4} e^{-\alpha_i \|\boldsymbol{r} - \boldsymbol{r}_i\|^2}$$

with respect to the Cartesian components of $\boldsymbol{r}_i$. For instance,

$$\phi_i^{p_x}(\boldsymbol{r}) = 2\sqrt{\alpha_i}(x - x_i)\phi_i^s(\boldsymbol{r})$$

can also be expressed as

$$\phi_i^{p_x}(\boldsymbol{r}) = \frac{1}{\sqrt{\alpha_i}} \frac{\partial \phi_i^s(\boldsymbol{r})}{\partial x_i}. \quad \text{(B2)}$$

The general formula for the overlap integrals, i.e., the elements of the overlap matrix, is given, e.g., by Eq. (3.5) in Ref. 46 and can also be calculated from recursion relations.[47] For convenience, we restate the simplified relations for the special cases involving s and p-type GTO's all in terms of the basic quantity

$$S_{ij} = S_{ji} = \left(\frac{2\sqrt{\alpha_i \alpha_j}}{\alpha_i + \alpha_j}\right)^{3/2} \exp\left[\frac{-\alpha_i \alpha_j}{\alpha_i + \alpha_j} r_{ij}^2\right] \quad \text{(B3)}$$

where $r_{ij} = \|\boldsymbol{r}_i - \boldsymbol{r}_j\|$, which is indeed the s-s overlap integral

$$\langle \phi_i^s | \phi_j^s \rangle = S_{ij}$$

Using Eq. (B2) we obtain

$$\langle \phi_i^{p_x} | \phi_j^s \rangle = \frac{1}{\sqrt{\alpha_i}} \frac{\partial S_{ij}}{\partial x_i} = -\left(\frac{2\sqrt{\alpha_i}\alpha_j}{\alpha_i + \alpha_j}\right)(x_i - x_j) S_{ij} \quad \text{(B4)}$$

and

$$\langle \phi_i^{p_x} | \phi_j^{p_{x'}} \rangle = \left(\frac{2\sqrt{\alpha_i \alpha_j}}{\alpha_i + \alpha_j}\right) S_{ij} \times \left[\delta_{x,x'} - \frac{2\alpha_i \alpha_j}{\alpha_i + \alpha_j}(x_i - x_j)\left(x_i' - x_j'\right)\right], \quad \text{(B5)}$$

where $x, x' \in \{x, y, z\}$ and δ denotes the Kronecker delta. The derivative of the basic quantity S_{ij} with respect to the atomic

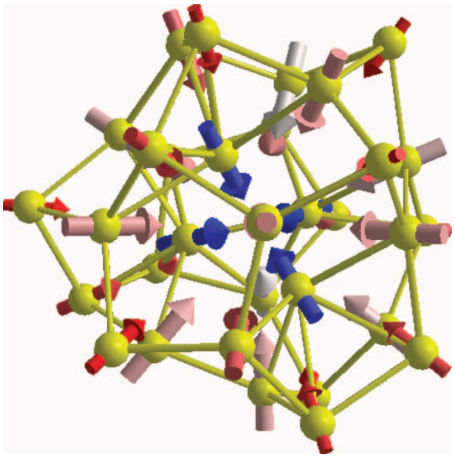

FIG. 7. Description of atomic environments for a Si_{32} cluster using the combined scalar and vectorial atomic fingerprints. Each atomic fingerprint consists of a scalar and a vector which are the corresponding s and (p_x, p_y, p_z) components of the principal eigenvector of the $4n \times 4n$ overlap matrix. The color of the vectors indicates the value (red corresponds to small values and blue to large values) of the scalar (s-type) fingerprint.

positions

$$\frac{\partial S_{ij}}{\partial x_k} = (\delta_{ik} - \delta_{jk})\left(\frac{-2\alpha_i\alpha_j}{\alpha_i + \alpha_j}\right)(x_i - x_j)S_{ij} \tag{B6}$$

is required to calculate the derivative of the overlap matrix elements, which in turn determine the derivative of its eigenvalues (see Eq. (C1))

$$D_{\nu,x_k} \equiv \frac{\partial V_\nu}{\partial x_k} = \left\langle \nu \left| \frac{\partial O}{\partial x_k} \right| \nu \right\rangle, \tag{B7}$$

where the eigenvector $|\nu\rangle$ corresponds to the eigenvalue V_ν of the overlap matrix O.

Eigenvectors associated to small eigenvalues seem not to contain any useful information. We therefore use the principal eigenvector of the overlap matrix as an atomic fingerprint, see Fig. 7. This vector gives the coefficients required to construct the pseudo-orbital with the largest pseudo charge density. This charge density has similarities to a true charge density since it is large in regions between neighboring atoms where covalent bonding can occur (Fig. 8).

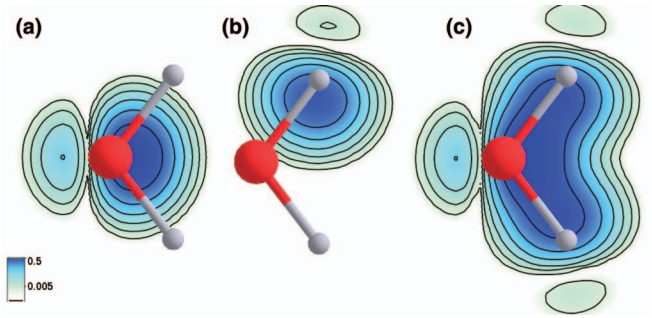

FIG. 8. Contributions of an oxygen (a) or hydrogen atom (b) to the total (c) pseudo-charge density $|\psi(\boldsymbol{r})|^2$ on the molecular plane for a water molecule. The coefficients of the orbitals ϕ_i^I from which the pseudo-wavefunction ψ is made, are the elements of the principal eigenvector of the overlap matrix constructed from s- and p-type GTO's.

APPENDIX C: CONSTANT-FINGERPRINT HYPERSURFACES

Using a constructive iterative procedure, we show in the following that the coincidence axiom for a configurational fingerprint is not satisfied if the dimension of the matrix is not sufficiently large and if therefore the resulting fingerprint vector is not sufficiently long. Consider two configurations p and q which are close. The difference of the fingerprint vectors is then given by a first order Taylor expansion

$$\boldsymbol{V}^p - \boldsymbol{V}^q \simeq D(q)(\boldsymbol{R}^p - \boldsymbol{R}^q). \tag{C1}$$

Note that, instead of the $3 \times n$ matrix notation used in Sec. II, hereafter we use a column vector $\boldsymbol{R} \in \mathbb{R}^{3n}$ for representing the atomic coordinates. Since $\boldsymbol{V}$ is a column vector of length N, the first derivative $D(q) \equiv \frac{\partial V}{\partial R}\big|_{R=R^q}$ is a $N \times 3n$ matrix. We assume that D has always the largest possible rank for the three types of matrices discussed in more detail in this section. For the Hamiltonian matrix this maximal rank $r_{\max}$ equals $\min(N, 3n - 6)$ if all N eigenstates included in the fingerprint vector are bound. For the overlap matrix $r_{\max}$ equals $\min(N - 1, 3n - 6)$ because the diagonal elements are independent of the configuration. For the Hessian matrix $r_{\max} = 3n - 6$ for configurations that are local minima with respect to the interaction potential and $r_{\max} = 3n - 3$ for all other cases.[48]

If $r_{\max}$ is less than $3n - 6$ one can find on a hypersurface of dimension $3n - 6 - r_{\max}$ (i.e., the nullity of D) configurations with identical fingerprint vectors, which are given as a solution of the equation

$$D\delta\boldsymbol{R} = \boldsymbol{0}. \tag{C2}$$

Formulated in words, configurational displacement vectors $\delta\boldsymbol{R}$ which are in the null space of D leave the fingerprint invariant to first order. For configurations which are further apart the first order approximation breaks down but Eq. (C2) can still be used as a starting point for mapping out such a hypersurface iteratively. We perform a move with a small amplitude along a vector $\delta\boldsymbol{R}$ in the null space of D. To correct for the small second and higher order deviations of the eigenvalues away from the hypersurface of constant eigenvalues defined as $\boldsymbol{V} = \boldsymbol{V}_{\text{ref}}$ we then solve

$$D\delta\boldsymbol{R}' = \boldsymbol{V}_{\text{ref}} - \boldsymbol{V} \tag{C3}$$

for the required displacement $\delta\boldsymbol{R}'$. Like Eq. (C1), the latter equation does not have a unique solution and we can therefore choose an arbitrary set of $r_{\max}$ coordinates which we want to modify in order to go back onto the hypersurface of constant eigenvalues. If the corresponding $r_{\max} \times r_{\max}$ matrix made out of D was ill-conditioned, we select another set of $r_{\max}$ atomic modification coordinates to ensure that Eq. (C3) is solved accurately. Since this moving back to the hypersurface requires only tiny displacements a single solution of the linear system is sufficient. If this was not the case it could be repeated which would correspond to a Newton iteration. By iterating this procedure of moves along the null space followed by moves that bring us exactly back on the hypersurface we can obtain clearly distinct configurations whose fingerprints are identical up to machine precision. Such examples are shown in Fig. 9 where the procedure is also illustrated schematically.

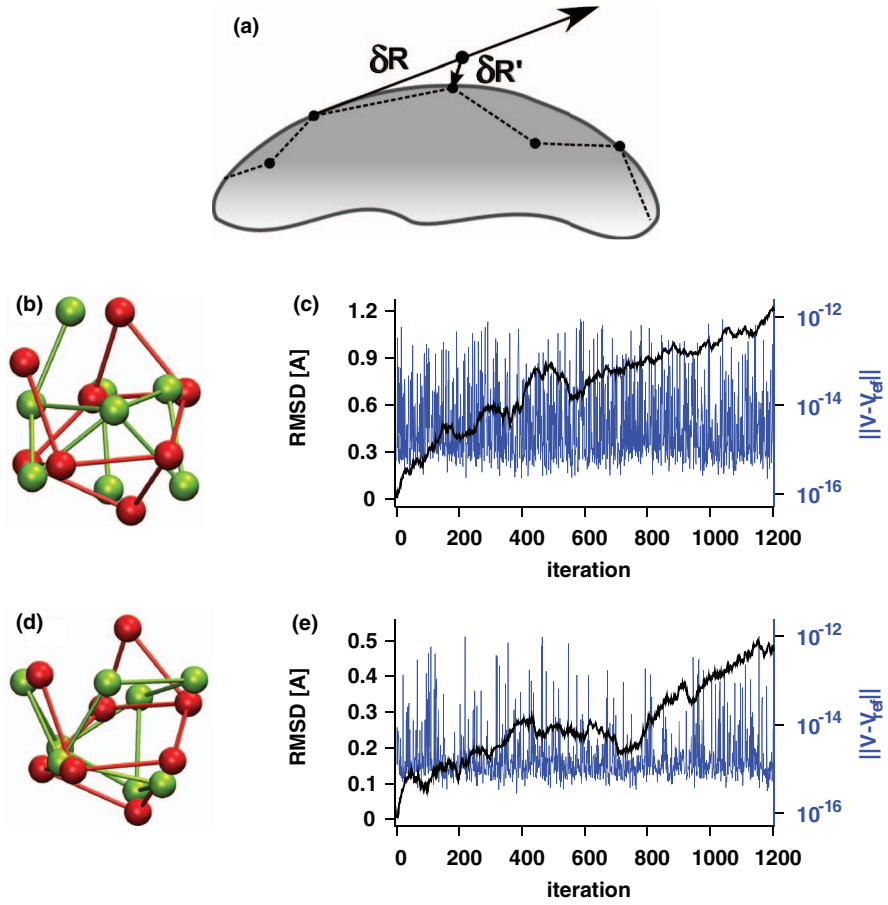

FIG. 9. (a) Schematic illustration of the exploration of the hypersurface defined by $\boldsymbol{V} = \boldsymbol{V}_{\text{ref}}$ consisting of iterative movements along $\delta\boldsymbol{R}$ (in the null space of D) followed by Newton step(s) $\delta\boldsymbol{R}'$ to come back to the hypersurface. Panel (b) shows two configurations (in red and green) of a Si_8 cluster whose fingerprint vectors of length n, obtained from an overlap matrix with one set of s-type GTO's, are identical. Panel (c) shows the evolution of the RMSD during the exploration of the hypersurface leading from the red structure to the green structure. Panels (d) and (e) contain the some information as panels (b) and (c) but for a fingerprint of length $2n$ obtained from an overlap matrix with two sets of s-type GTO's. In both cases $\|\boldsymbol{V} - \boldsymbol{V}_{\text{ref}}\|$ is vanishingly small.

Note that at each iteration we orthogonalize $\delta\boldsymbol{R}$ of the previous iteration to the row space of current D. This reduces the probability of moving backwards to the starting point.

[1]A. R. Oganov, *Modern Methods of Crystal Structure Prediction* (Wiley-VCH Verlag GmbH & Co. KGaA, 2010).
[2]S. Goedecker, J. Chem. Phys. **120**, 9911 (2004).
[3]M. Amsler and S. Goedecker, J. Chem. Phys. **133**, 224104 (2010).
[4]M. Neumann, F. Leusen, and J. Kendrick, Angew. Chem., Int. Ed. **47**, 2427 (2008).
[5]A. R. Oganov and M. Valle, J. Chem. Phys. **130**, 104504 (2009).
[6]G. M. Downs and P. Willett, Rev. Comput. Chem. **7**, 1 (1996).
[7]E. Velasquez, E. R. Yera, and R. Singh, in *IEEE Symposium on Bion-Informatics and BioEngineering*, BIBE (IEEE Computer Society, 2006), pp. 261–268.
[8]E. Karakoc, A. Cherkasov, and S. C. Sahinalp, Bioinformatics **22**, e243 (2006).
[9]I. D. Kuntz, E. C. Meng, and B. K. Shoichet, Acc. Chem. Res. **27**, 117 (1994).
[10]G. M. Downs, P. Willett, and W. Fisanick, J. Chem. Inf. Comput. Sci. **34**, 1094 (1994).
[11]Y. Zhang, Curr. Opin. Strut. Biol. **18**, 342 (2008).
[12]V. J. Gillet, D. J. Wild, P. Willett, and J. Bradshaw, Comput. J. **41**, 547 (1998).
[13]R. P. Sheridan and S. K. Kearsley, Drug Discovery Today **7**, 903 (2002).
[14]R. Ponec, L. Amat, and R. Carb-dorca, J. Comput.-Aided Mol. Des. **13**, 259 (1999).
[15]C. Lemmen and T. Lengauer, J. Comput.-Aided Mol. Des. **14**, 215 (2000).
[16]D. R. Flower, J. Chem. Inf. Comput. Sci. **38**, 379 (1998).
[17]B. C. P. Allen, G. H. Grant, and W. G. Richards, J. Chem. Inf. Comput. Sci. **41**, 330 (2001).
[18]F. Schwarzer and I. Lotan, in *Proceedings of RECOMB'03* (ACM, New York, NY, USA, 2003), pp. 267–276.
[19]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[20]R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics* (Wiley-VCH, Weinheim, 2009).
[21]R. L. Martin, B. Smit, and M. Haranczyk, J. Chem. Inf. Model. **52**, 308 (2012).
[22]J. Cheng and R. Fournier, Theor. Chem. Acc. **112**, 7 (2004).
[23]J. Lee, I.-H. Lee, and J. Lee, Phys. Rev. Lett. **91**, 080201 (2003).
[24]M. Valle and A. R. Oganov, Acta Crystallogr., Sect. A: Found. Crystallogr. **66**, 507 (2010).
[25]L. Genovese, A. Neelov, S. Goedecker, T. Deutsch, S. A. Ghasemi, A. Willand, D. Caliste, O. Zilberberg, M. Rayson, A. Bergman, and R. Schneider, J. Chem. Phys. **129**, 014109 (2008).
[26]W. Kabsch, Acta Crystallogr. **A34**, 827 (1978).
[27]B. K. P. Horn, H. Hilden, and S. Negahdaripour, J. Opt. Soc. Am. **5**, 1127 (1988).
[28]E. A. Coutsias, C. Seok, and K. A. Dill, J. Comput. Chem. **25**, 1849 (2004).
[29]D. L. Theobald, Acta Cryst. A **61**, 478 (2005).
[30]H. W. Kuhn, Naval Res. Logistics Quart. **2**, 83 (1955).
[31]D. J. Wales and J. M. Carr, J. Chem. Theory Comput. **8**, 5020 (2012).
[32]H. L. Morgan, J. Chem. Doc. **5**, 107 (1965).
[33]Z. Ouyang, S. Yuan, J. Brandt, and C. Zheng, J. Chem. Inf. Comput. Sci. **39**, 299 (1999).
[34]B. Helmich and M. Sierka, J. Comput. Chem. **33**, 134 (2012).

[35]G. Carpaneto, S. Martello, and P. Toth, Ann. Operat. Res. **13**, 191 (1988).
[36]Although we use the eigenvalues to form the vector $\boldsymbol{V}$ for describing entire structures throughout this work, one can also fill the vector $\boldsymbol{V}$ by the elements of selected eigenvectors. Then each element of $\boldsymbol{V}$ corresponds to an atom and the ensemble belonging to one atom forms an atomic fingerprint or descriptor of the local environment of the atom. For instance, if the principal eigenvector of the overlap matrix of one *s* and one *p*-type GTO per atom is used (as in Eqs. (5) and (6)), each individual atom is accordingly described by four numbers, as depicted in Fig. 7.
[37]A. Patterson, Nature (London) **143**, 939 (1939).
[38]V. G. Grigoryan and M. Springborg, Chem. Phys. Lett. **375**, 219 (2003).
[39]G. Benson and E. Dempsey, Proc. R. Soc. London, Ser. A **266**, 344 (1962).
[40]F. Pietrucci and W. Andreoni, Phys. Rev. Lett. **107**, 085504 (2011).
[41]P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).
[42]A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[43]J. E. Moussa, Phys. Rev. Lett. **109**, 059801 (2012).
[44]D. Caliste *et al.*, V_Sim visualization software, CEA-INAC, Grenoble, France; http://inac.cea.fr/L_Sim/V_Sim.
[45]K. Momma and F. Izumi, J. Appl. Crystallogr. **44**, 1272 (2011).
[46]E. Clementi and D. Davis, J. Comput. Phys. **1**, 223 (1966).
[47]S. Obara and A. Saika, J. Chem. Phys. **89**, 1540 (1988).
[48]M. J. Field, *A Practical Introduction to the Simulation of Molecular Systems* (Cambridge University Press, 1999).