How Different Are Two Chemical Structures?

J. M. C. Marques,*,† J. L. Llanio-Trujillo,† P. E. Abreu,† and F. B. Pereira‡,¶

Departamento de Química, Universidade de Coimbra, 3004-535 Coimbra, Portugal, Instituto Superior de Engenharia de Coimbra, Quinta da Nora, 3030-199 Coimbra, Portugal, and Centro de Informática e Sistemas da Universidade de Coimbra (CISUC), 3030-290 Coimbra, Portugal

Received June 4, 2010

We extend the scope of a recent method for superimposing two molecules (*J. Chem. Phys.* **2009**, *131*, 124126-1−124126-10) to include the identification of chiral structures. This methodology is tested by applying it to several organic molecules and water clusters that were subjected to geometry optimization. The accuracy of four simpler, non-superimposing approaches is then analyzed by comparing pairs of structures for argon and water clusters. The structures considered in this work were obtained by a Markovian walk in the coordinate space. First, a random geometry is generated, and then, the iterative application of a mutation operator ensures the creation of increasingly dissimilar structures. The discriminating power of the non-superimposing approaches is tested by comparing the corresponding dissimilarity measures with the root-mean-square distance obtained from the superimposing method. Finally, we showcase the application of those methods to characterize the diversity of solutions in global geometry optimization by evolutionary algorithms.

INTRODUCTION

The recognition of structural similarities between molecules is a major issue for chemical science, especially in the subjects at the frontier with biology and pharmaceutics. In particular, the identification of molecules with a geometry that resembles that of a given query structure (e.g., a pharmacophore) is crucial for computer-aided drug discovery. This involves the exploration of large databases of molecular structures to search for the best candidates that can be used in the design of new drugs. In the case of flexible molecules with different accessible low-energy structures, databases often include various conformers of the same compound, which constitutes an additional difficulty for the specific shape identification. Obviously, the daunting task of comparison among a huge number of structures in a database requires the development of efficient methods for shape recognition, and indeed, many approaches have been proposed to achieve such a goal.

Methods for comparing molecular structures are essentially of two types: (i) superimposing methodologies rely on the best superposition of the molecules to be compared, while (ii) non-superimposing approaches use internal quantities of each molecule that are independent of a particular orientation. The methods described in (i) are “exact” in the sense that they can lead, in principle, to the best overlap between two structures (including the correct assignment of the atoms); after achieving the best superposition, the root-mean-square distance (rmsd) between the two structures can be easily calculated. Because this procedure allows for an unambiguous comparison between structures, superimposing methods are preferable over those in (ii). Indeed, non-superimposing approaches rely on the comparison of only a small number of descriptors that seek to encode the essential features of the structure, i.e., without any attempt to ensure completeness. However, non-superimposing approaches often allow for the identification of similarities between structures at a low computation cost; this is an obvious advantage over superimposing methods that are quite time-consuming. Several superimposing methods are described in the literature,[1−6] and some of them are reviewed in ref 7. Nevertheless, most of the methods require a correct assignment between atoms of the two structures to be compared in order to determine the best overlap. Recently, Vásquez-Pérez et al.[8] designed a superimposing algorithm that solves the atom-labeling problem and finds out the best superposition in a self-consistent way. Concerning non-superimposing approaches, there are also various computational schemes proposed in the literature,[9−16] though the ultrafast shape recognition (USR) method[14] is perhaps the most popular for identifying similar structures.

In this work, we are interested in the relevance of structural comparison methods to atomic and molecular geometry optimization. In particular, evolutionary algorithms[17−21] (EAs) are among the most successful optimization methods for discovering the global minimum structure of complex systems. Recently, we have developed state-of-the-art EAs to be applied in the geometry optimization of Morse[21−23] and argon[24] clusters, as well as binary Lennard-Jones and mixed Ar−Kr clusters.[25] Thus, in the context of geometry optimization, we identify two related issues that benefit from using accurate algorithms for structural comparison. One is associated with the post-optimization verification whether similar structures had been obtained, e.g., by other optimization methods, which is particularly difficult to be done by eye-inspection for clusters composed by many atoms or molecules. The other issue is due to the identification of similarities between structures arising during a global

* To whom correspondence should be addressed. E-mail: qtmarque@ci.uc.pt.
† Universidade de Coimbra.
‡ Instituto Superior de Engenharia de Coimbra.
¶ Centro de Informática e Sistemas da Universidade de Coimbra (CISUC).

10.1021/ci100219f
Published on Web 12/07/2010

geometry optimization procedure, e.g., by using an EA where it is important to guarantee a certain degree of diversity among the set of structures forming the population. Clearly, the first issue may be addressed with superimposing methods. In contrast, the second involves a large number of comparisons between the structures forming the population, and hence, the less time-consuming non-superimposing methods are perhaps the only ones that can be applied. To be reliable, such methods should be able to answer the title question, i.e., they should catch adequately the differences between two structures and correlate well with the corresponding superimposing outcome.

The goal of the present paper is 2-fold. First, we develop a ready-to-use computational program that extends the scope of the superimposing method recently proposed by Vásquez-Pérez et al.[8] to identify chiral structures. This is relevant for many chemical applications, because this computational tool is able to qualify a particular difference (i.e., chirality) between two structures, which is sometimes nontrivial to assign. Second, we analyze the application of non-superimposing approaches for fast shape recognition during the optimization of clusters with EAs; the reliability of these simpler approaches is tested against the superimposing method. Among the available schemes proposed in literature for fast recognition of structures, we study the USR[14] (and its CSR[16] extension for identification of chiral molecules) and a method based on the comparison of center of mass distances.[15] In addition, we propose and analyze a very simple descriptor approach based on the global shape of the molecule as given by the inertia principal momenta. This relies on the global shape of the chemical species and, hence, does not involve local structure details.

The structure of the paper is as follows. In "Superimposing Method", we describe our implementation of the superimposing method proposed by Vásquez-Pérez et al.,[8] as well as its extension to identify enantiomers. This new method is, then, tested for identifying the enantiomers of several chemical structures. Also shown is the importance of the method for cluster geometry optimization. In turn, "Non-Superimposing Approaches" describes briefly four algorithms for fast shape recognition, while "Application to Cluster Optimization" analyses the reliability of such non-superimposing methods for the application in global optimization of atomic and molecular clusters by EAs. Finally, the main conclusions are gathered in "Conclusions".

SUPERIMPOSING METHOD

The Algorithm. The superposition of two structures with arbitrary atom labeling is achieved by applying the algorithm recently proposed by Vásquez-Pérez et al.[8] Such a method has been validated by del Campo and Köster[26] in a reactant−product alignment for the transition state search of 28 reactions. Our own implementation of the superimposing method[8] is illustrated through a flux diagram in Figure 1 and, for clarity, it is briefly described as follows.

Given two structures with N atoms and Cartesian coordinates vectors $\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1$ and $\mathbf{x}_2, \mathbf{y}_2, \mathbf{z}_2$, their best superposition is achieved by minimizing the function

$$d^2 = \sum_{i=1}^{N} [(\mathbf{x}_2^i - \mathbf{x}_1^i)^2 + (\mathbf{y}_2^i - \mathbf{y}_1^i)^2 + (\mathbf{z}_2^i - \mathbf{z}_1^i)^2] \quad (1)$$

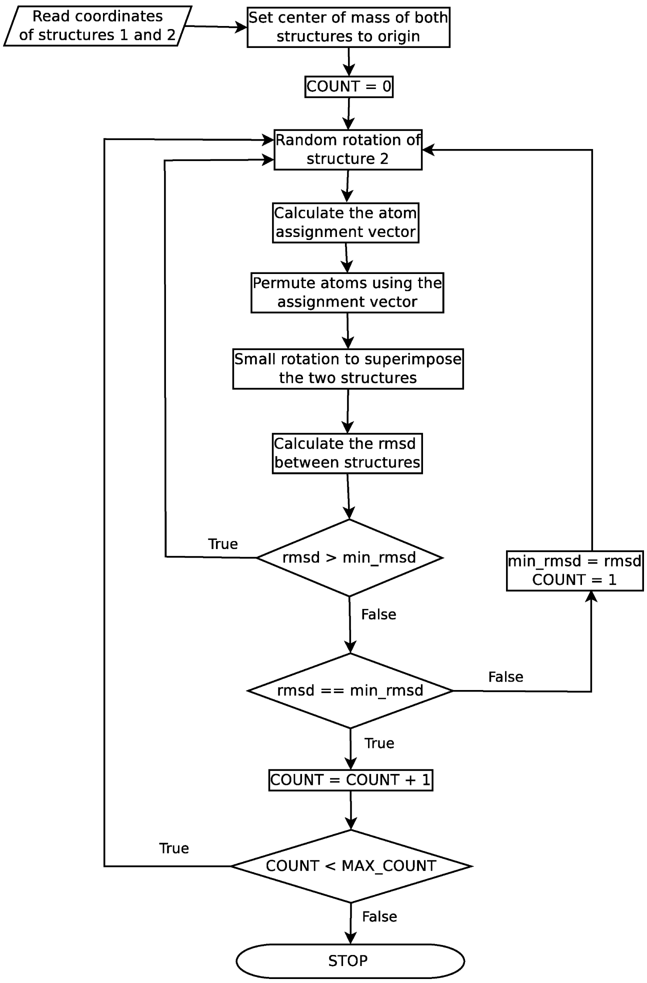
Read coordinates of structures 1 and 2
Set center of mass of both structures to origin
COUNT = 0
Random rotation of structure 2
Calculate the atom assignment vector
Permute atoms using the assignment vector
Small rotation to superimpose the two structures
Calculate the rmsd between structures
True
rmsd > min_rmsd
False
min_rmsd = rmsd
COUNT = 1
rmsd == min_rmsd
False
True
COUNT = COUNT + 1
True
COUNT < MAX_COUNT
False
STOP

Figure 1. Flux diagram illustrating the present implementation of the superimposing algorithm proposed in ref 8.

which represents the sum over all the N square distances defined between each atom in structure **2** and the corresponding one in structure **1** (taken as reference). As suggested by the flux diagram of Figure 1, the first step in the minimization of d^2 corresponds to making a translation of coordinates, i.e., to move the two centers of mass to the origin of the axis. We note that such translation needs to be applied only once, since the center of mass is unchanged by performing rigid-body rotations to find the best orientation that superimposes the two structures.

However, the main difficulty in minimizing d^2 results from the fact that atom i of structure **2** does not generally correspond to atom i of structure **1**. This happens for instance in cluster geometry optimization with EAs, where the application of genetic operators (e.g., crossover) destroys the original order of the atoms, and hence, the correct assignment is unknown a priori. Thus, the superposition of two structures requires the solution of the assignment problem in advance, so that both sets of atoms become in the same order. The so-called Hungarian algorithm developed by Kuhn[27] (also known as the Kuhn−Munkres algorithm[27,28]) was adopted within the present superimposing method to reorder the atoms and, hence, obtain the optimal mapping between them; here, we have applied the Hungarian algorithm as implemented in a routine converted from Fortran 77 to Fortran 90 by Miller.[29] Basically, this routine calculates the atom-assignment vector of the two structures once a cost matrix is provided as input. Since we do not define mass-weighted coordinates as in ref 8, the cost matrix is slightly modified here to account for chemical systems presenting isotope substitution of structural equivalent atoms (e.g., a hydrogen atom by a deuterium in a methyl group). As applied in this work, the elements of the cost matrix are written as

$$F_{ij} = [(\mathbf{x}_2^j - \mathbf{x}_1^i)^2 + (\mathbf{y}_2^j - \mathbf{y}_1^i)^2 + (\mathbf{z}_2^j - \mathbf{z}_1^i)^2]^{1/2} + b \tag{2}$$

where $b = 0$ if atoms i and j have the same mass and $b = 10$ otherwise. Since the Hungarian algorithm finds the mapping that minimizes the total cost, the parameter b in eq 2 works as a penalty to avoid the wrong assignment between two structural equivalent atoms having different masses; the value of b for unlike atoms is arbitrary and one can use another positive quantity, since the algorithm is not sensitive to it.

In general, structures **1** and **2** are not oriented in such a way to obtain the best assignment given by the Hungarian algorithm (see ref 8 for a detailed discussion). This can be solved by using a probability driven approach,[8] which involves the application of the Hungarian algorithm to several randomly generated orientations of structure **2** (as displayed in Figure 1), followed by reorientation to tune for the best superposition of the two structures. The reorientation is carried out by employing a rotation matrix defined through the components of the quaternion corresponding to the eigenvector associated with the lowest eigenvalue of the well-known[6,7] Kearsley's matrix.[30] Whenever this process is repeated, the rmsd between structure **1** and structure **2** is calculated. The lowest rmsd is booked and the cycle stops when the lowest rmsd has appeared a number of times defined by the MAX_COUNT parameter (cf. Figure 1); as in ref 8, we have considered MAX_COUNT = 50 for all applications in this work. In turn, the algorithm also takes into account a maximum number of iterations to avoid exausting the computational resources in cases of late convergence. If this value is reached before MAX_COUNT is attained, then the lowest rmsd found so far is returned. In the applications described in this paper, the maximum number of iterations was set to 1×10^5.

Discovering Enantiomers. When optimization algorithms are applied to chemical systems, they search for arrangements of atoms with low (and sometimes the lowest) energy, but usually do not consider whether an arising structure is an enantiomer of another previously localized. Moreover, the interconversion between two enantiomers is probably difficult for searching procedures based on molecular dynamics, while a recent work on cluster optimization has shown[31] the importance of applying operators that make changes in some specific bond arrangements. It is reasonable to consider that state-of-the-art algorithms for geometry optimization should incorporate chirality detection, because many chemical systems present enantiomers. Indeed, minima structures belonging to common symmetry-point groups, such as C_n or D_n (but not only these), are always chiral and, hence, have enantiomers. For instance, we have localized 13 (over a total of 77) putative global minima of argon clusters[24] and 10 (over a total of 39) minima of mixed Ar−Kr clusters[25] with C_1 or C_2 symmetry, just to mention our recent experience on rare-gas atomic clusters. In the particular case of Ar_{21}, whose global minimum is of C_1 symmetry,[24] we have verified with the algorithm described below that, from 30 runs of the EA,[24] one of the enantiomers arises 17 times while the other is obtained in the remaining 13.

Although enantiomers exhibit some relevant differences in their properties (see, e.g., ref 32 for a dicussion on chirality), the structural similarity is obvious. This and the fact that, in many examples, enantiomers present the same potential energy lead such structures to be undetected in molecular geometry optimization. Even a detailed eye-inspection procedure is not conclusive in many situations (especially for complex chemical systems) and one cannot easily catch the structural relationship between both molecules.

The application of a superimposing method to discover enantiomers is, by its own definition, impossible! From the historical definition of chirality,[33] the concept of enantiomer relies on a non-superimposing mirror-image structure. Hence, if two structures are enantiomers, they cannot be superimposed and the single application of the algorithm of Vásquez-Pérez et al.[8] described above cannot identify those structures. Here, we devise a simple procedure that allows for the identification of enantiomers while using the above-mentioned superimposing algorithm. Thus, given two structures (hereafter designated as structure **1** and structure **2** with vectors of Cartesian coordinates defined as $\mathbf{x_1}$, $\mathbf{y_1}$, $\mathbf{z_1}$ and $\mathbf{x_2}$, $\mathbf{y_2}$, $\mathbf{z_2}$, respectively), the enantiomer assigment procedure is described as follows.

(1) The superimposing algorithm described in the previous subsection is applied to structures **1** and **2**, and the rmsd between them is calculated; the value is kept as $\text{rmsd}_\mathbf{1}$.

(2) If $\text{rmsd}_\mathbf{1}$ is small (say, $\text{rmsd}_\mathbf{1} < 5.0 \times 10^{-2}$ Å), then they are assigned as the same isomers. Otherwise, the superimposing algorithm is applied again, but for structure **1** and the specular image of structure **2**. This is obtained by reflecting the structure in a plane (e.g., the xz-plane) so that the image structure arises with coordinate vectors $\mathbf{x_2}$, $-\mathbf{y_2}$, $\mathbf{z_2}$. The second application of the superimposing algorithm leads to the calculation of $\text{rmsd}_\mathbf{2}$.

(3) If $\text{rmsd}_\mathbf{2}$ is small (and, in general, smaller than $\text{rmsd}_\mathbf{1}$), structures **1** and **2** are assigned as enantiomers.

The Fortran code that implements the superimposing algorithm with enantiomer assignment is available through download of the Supporting Information of this paper.

Some Applications. To test the superimposing algorithm against the discovery of chiral molecules, we have applied it to the enantiomer structures of bromochlorofluoromethane, 2,3-dibromobutane, 2-cloro-3-bromobutane, hexahelicene, and [6.6]chiralane. Bromochlorofluoromethane is a typical molecule with enantiomers due to the presence of a chiral carbon, while stereoisomers of helical molecules like hexahelicene (or DNA) result from an axial dissymmetry of the structure. In contrast, 2,3-dibromobutane and 2-cloro-3-bromobutane show two chiral centers, which leads to a classification of the four possible conformers as *RR*, *SS*, *SR*, and *RS* (in agreement with the well-known Cahn−Ingold−Prelog rules[34,35]). For 2,3-dibromobutane, conformers *RR* (*RS*) and *SS* (*SR*) are equal, and hence, there are only two enantiomers. The 2-cloro-3-bromobutane has distinct halogen atoms bonded to each chiral center, leading to two pairs of non-superimposing structures (i.e., *RR* with *SS* and *RS* with *SR*), while *RR* (*SS*) and *RS* (*SR*) are superimposing diastereomers. Another interesting molecule is [6.6]chiralane, which

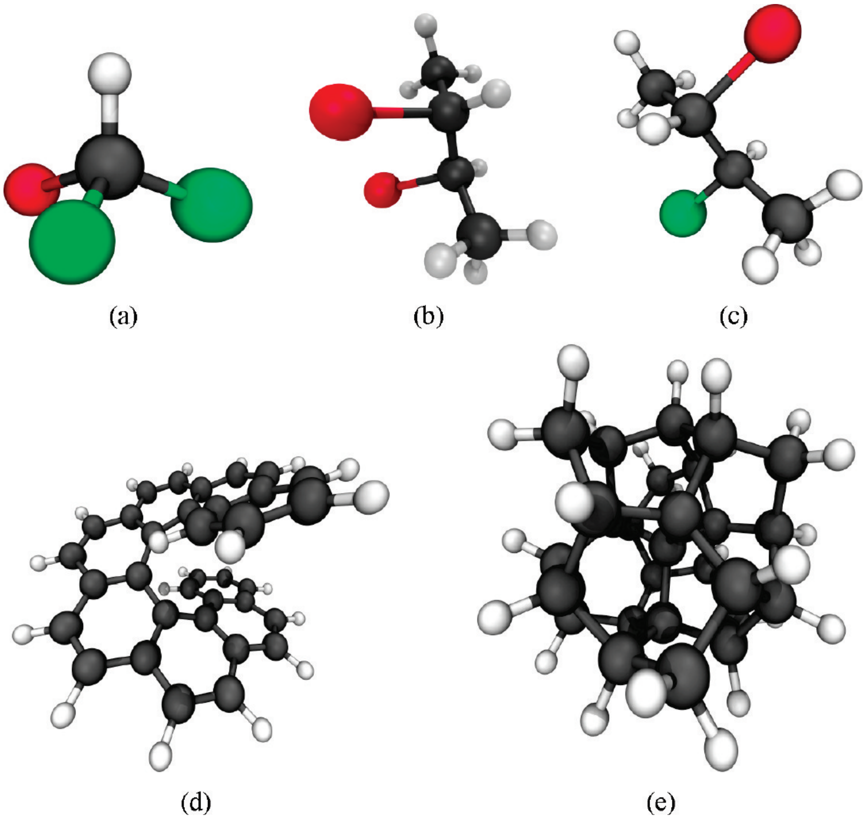

Figure 2. Five molecules presenting enantiomers that have been used to test the superimposing algorithm with chirality assignment: (a) bromochlorofluoromethane, (b) 2,3-dibromobutane, (c) 2-cloro-3-bromobutane, (d) hexahelicene, and (e) [6.6]chiralane. Only one enantiomer is represented in each case.

belongs to the *T* point group, with all carbon atoms forming various homochiral twist-boat cyclohexane rings.[36] All these five structures are illustrated in Figure 2.

The atomic coordinates of each enantiomer have been obtained as follows. First, we used Marvin[37] to draw the two-dimensional structures and then used molconvert (Molecule File Converter, bundled with Marvin) to convert the structures to three-dimensional coordinates. These structures were optimized with MOPAC2009[38] using the PM3[39,40] Hamiltonian. Since the pairs of enantiomers resulting from the optimization can hardly be exact specular images, such structures constitute more real-life instances to test our methodology. Finally, a random translation and rotation are applied to the coordinates of one enantiomer of each pair; also a random shuffling of the atoms in the structure is carried out. This procedure is expected to unbias the outcome from the application of the algorithm. When applying the superimposing algorithm without chirality assignment, one expects to obtain the best overlap between each pair of enantiomers. This outcome is shown in Figure 3 for 2,3-dibromobutane (panel a) and hexahelicene (panel b). It is clear from this figure that the two structures for each instance cannot be superimposed. In turn, the superposition is perfect after using the chirality assignment method, i.e, when one structure and the specular image of the second are brought to overlap. This is illustrated in Figure 3 for 2,3-dibromobutane (panel c) and hexahelicene (panel d). Although not shown, similar outcomes were obtained for the other molecules.

Moreover, the robustness of the superimposing method with chirality assignment was tested, for each instance, by performing comparisons between one structure and 1000 randomly modified replicas of the second one (as described above); all the relevant pairs of the stereoisomers have been tested with the present methodology. The main results are given in Table 1, presenting the average values of rmsd_1 ($\langle\text{rmsd}_1\rangle$) and rmsd_2 ($\langle\text{rmsd}_2\rangle$) and the corresponding average number of random rotations ($\langle n_{\text{rot}}\rangle$) needed to obtain the lowest rmsd between two structures 50 times (i.e., the value of MAX_COUNT in Figure 1). Although we represent average root-mean-square distances, it is important to emphasize that the largest deviation from those values was always below the precision displayed in each case. In addition, the last column of Table 1 gives the classification of each pair of structures obtained for all the 1000 trials according to the values of rmsd_1 and rmsd_2.

From a detailed inspection of Table 1 one can make the following remarks concerning the present algorithm. For the simplest case, bromochlorofluormethane, the algorithm performs as expected: the average number of rotations needed to find the best superposition is equal to the 50 (the MAX_COUNT parameter value, in this case), showing that all assignments always result in the smallest rmsd. In contrast, the other instances require a large number of random rotations, which is closely related with the difficulty to obtain the best overlap between the two structures, and hence, $\langle n_{\text{rot}}\rangle$ becomes smaller for superimposing identical systems. In turn, for the 2,3-dibromobutane and 2-cloro-3-bromobutane cases, the algorithm is capable of distinguishing between diastereomers and enantiomers. It also shows the similarities between structures: for 2-cloro-3-bromobutane, notice the symmetry in the comparison of the *RR* isomer with both *SR* and *RS*, showing that *SR* and *RS* are identical. For the more

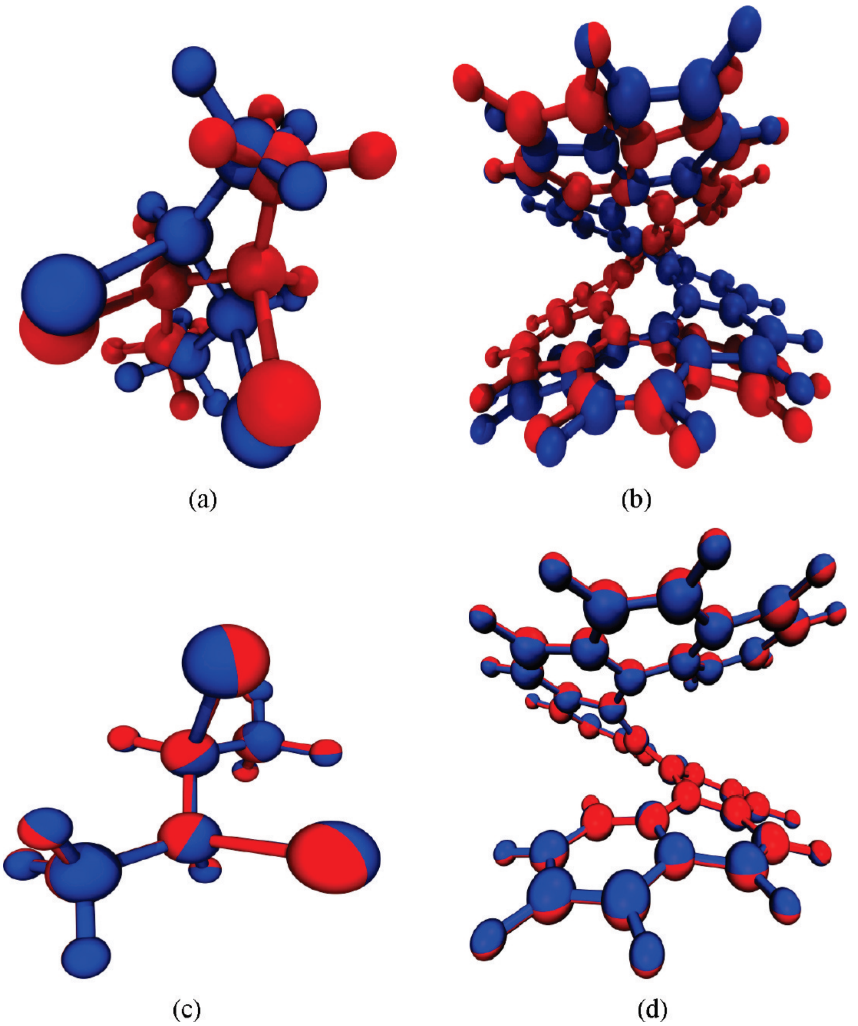

Figure 3. Best superposition after applying the superimposing algorithm without (a and b) and with (c and d) chirality assignment: *RR* and *SS* enantiomers of 2,3-dibromobutane (a and c); *M* and *P* enantiomers of hexahelicene (b and d).

Table 1. Summary of the Computational Tests for Discovering Enantiomers of Several Structures

		first superposition[a]		second superposition[a]		
system	structures	$\langle \text{rmsd}_{\mathbf{1}} \rangle$[b]	$\langle n_{\text{rot}} \rangle$	$\langle \text{rmsd}_2 \rangle$[b]	$\langle n_{\text{rot}} \rangle$	classification
bromochlorofluoromethane	*R*, *S*	1.274	50	1.093(−4)	50	enantiomers
2,3-dibromobutane	*RR*, *RS*	1.562	2624	1.588	2744	diastereomers
	RR, *SS*	0.705	4251	1.296(−2)	1697	enantiomers
	SS, *RS*	1.585	2755	1.560	2666	diastereomers
	SR, *RS*	4.534(−2)	1410			identical
2-cloro-3-bromobutane	*RR*, *SS*	0.752	13585	5.411(−3)	1688	enantiomers
	RR, *SR*	1.095	1928	1.332	4389	diastereomers
	RR, *RS*	1.333	4843	1.092	1914	diastereomers
	RS, *SR*	0.579	3144	8.652(−3)	1440	enantiomers
hexahelicene	*P*, *M*	1.248	30533	2.291(−3)	10475	enantiomers
[6.6]chiralane	*R*, *S*	0.774	2462	1.803(−4)	1554	enantiomers

[a] According to the present algorithm, the "first superposition" refers to the superposition of structures **1** and **2**, while the specular image of structure **2** is superimposed on structure **1** in the "second superposition". [b] Rmsd values are in Å.

complicated hexahelicene case, where the usual *R* and *S* labels are replaced by the *M* and *P* labels (these labels constitute the equivalent notation for helical molecules), it is worth noting the extra effort needed for the assignment, which is apparent from the large number of random rotations ($\langle n_{\text{rot}} \rangle$). Even for [6.6]chiralane, which is a "rotationally symmetric, extreme chiral molecule",[36] the algorithm performs equally well, being able to fully identify the enantiomers.

Since one of the goals of our work is the study of atomic and molecular clusters, it would be interesting to apply the present superimposing algorithms to those systems. Very recently, we have carried out a global geometry optimization of TIP4P[41] water clusters up to $(H_2O)_{20}$ to test a new EA developed in our group and reproduced all the putative global minima obtained independently by Wales and Hodges[42] and Takeuchi.[31] For the $(H_2O)_{16}$ cluster, the global minimum

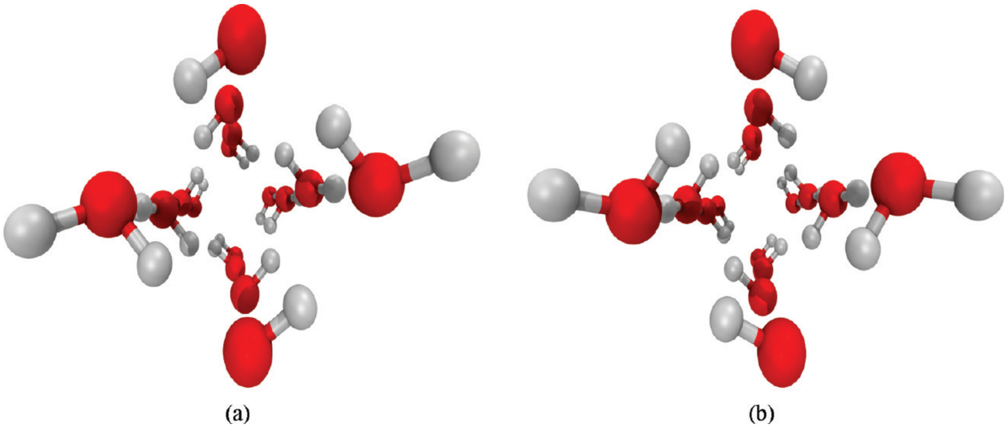

Figure 4. Global minimum structures of the $(H_2O)_{16}$ cluster obtained by (a) Wales and Hodges[42] and by (b) Takeuchi.[31] As shown, the two structures differ by the orientation of the hydrogen atoms.

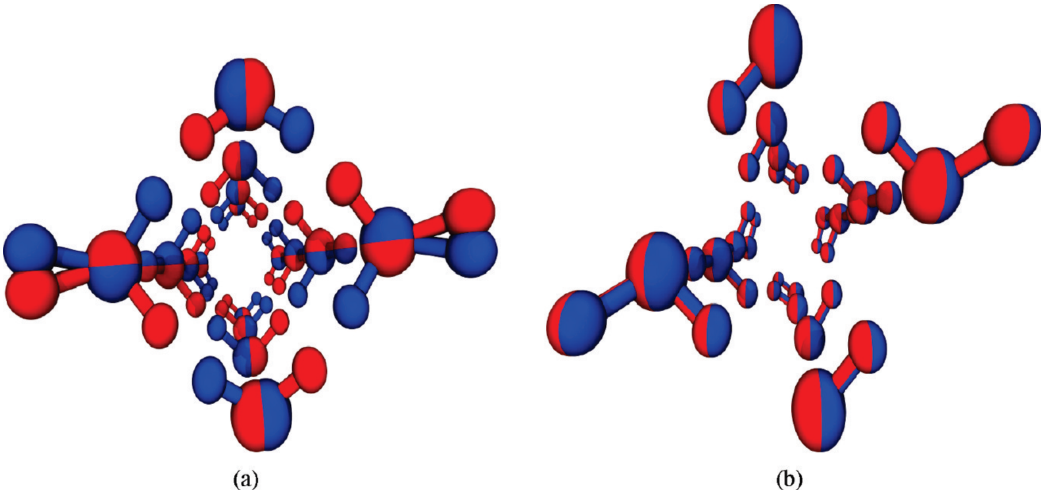

Figure 5. Best superposition of the two global minimum structures of the $(H_2O)_{16}$ shown in Figure 4: after aplication of the superimposing algorithm without (a) and with (b) chirality assignment.

structure of Takeuchi[31] differs from the corresponding one of Wales and Hodges[42] in the directions of hydrogen bonds in the four-membered ring; both minimum structures are shown in Figure 4. Besides the structural differences, the two structures have approximately the same energy. By applying the superimposing algorithm without chirality assignment, one observes in Figure 5a that the two structures cannot be superimposed, while the rmsd is 0.495 Å. In contrast, the application of the superimposing algorithm with chirality assignment leads to a very good superimposition (the rmsd is now 6×10^{-5} Å), and hence, the two structures are classified as enantiomers; the result is shown in Figure 5b. It is worth noting that the two structures were never assigned before as enantiomers, although Takeuchi has recognized that one configuration is easily transformed into the other "if the directions of the four hydrogen bonds in the ring are reversed".[31]

Finally, another application of the superimposing algorithm may be the comparison of structures corresponding to different local minima of a given system. As an example, we have recently identified[43] five low-energy minima of Ar_{38} belonging to the O_h (global minimum), C_{5v}, C_s, C_{5v} (hereafter denoted as C_{5v}' for clarity), and C_1 symmetry point groups, repectively; these five structures are illustrated in Figure 6. It is well-known that Ar_{38} presents a double-funnel energy landscape,[44] and hence, we can assign the global minimum to the deepest funnel and the remaining four structures to the other one (cf. ref 43). Then, if this corresponds to a structural signature, it should be apparent in the rmsd values resulting from the best superposition of those structures. The results are represented in Table 2. Indeed, one observes from Table 2 that the superposition of the O_h global minimum structure with the other clusters always leads to the largest rmsd values (i.e., around 1.5 Å or higher). Conversely, the superposition of the four structures (C_{5v}, C_s, C_{5v}', and C_1) placed in the second funnel lead to rmsd values always smaller than 1 Å. Clearly, this indicates that the structures of those four clusters are much more similar among each other than with the global minimum (which is located in a distinct funnel).

NON-SUPERIMPOSING APPROACHES

Although superimposing methods may be applied along with geometry optimization algorithms, they are computationally demanding for most of the interesting cluster systems. In turn, the use of non-superimposing approaches involves only the comparison of quantities based on a limited set of internal coordinates, and hence, it eliminates any need for atomic assignment and translation and orientation of the

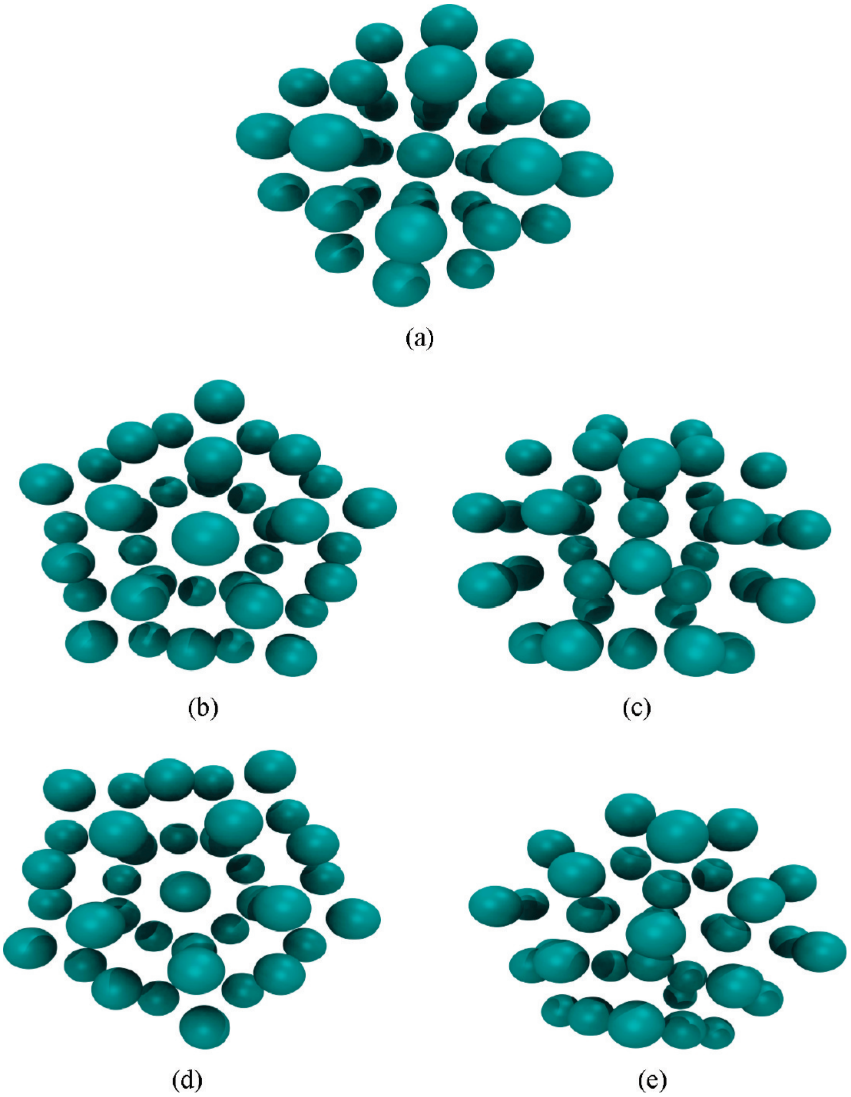

Figure 6. Five low-energy minimum structures of Ar_{38} located in our previous work.[43] They belong to the symmetry point groups (ordered from the lowest to the highest energy) (a) O_h, (b) C_{5v}, (c) C_s, (d) C_{5v} (denoted in the text as C_{5v}'), and (e) C_1.

Table 2. Root-Mean Square Distance (rmsd) of the Best Overlap between Different Pairs of Structures for Ar_{38}[a]

	C_{5v}	C_s	C_{5v}'	C_1
O_h	1.715	1.455	1.664	1.511
C_{5v}		0.978	0.829	0.824
C_s			0.802	0.582
C_{5v}'				0.952

[a] Rmsd values are in Å.

structure. Because of this, non-superimposing methods are in general suitable for the comparison of structures in large-scale computations such as in the application of global geometry optimization algorithms.

Here, we test three non-superimposing algorithms proposed in literature: ultrafast shape recognition[14] (USR), chiral structure recognition[16] (CSR), and center of mass (CM) measure.[15] In addition, we propose and test a global shape descriptor (GSD) method. All these methods are based on the calculation of descriptors that, in a certain way, encode the shape of the system. Then, the degree of similarity between two structures may be quantified from an inverse-scaled generalized mean distance. Thus, a normalized similarity measure between two structures, A and B, may be given by

$$S_{\mathrm{A,B}} = \frac{1}{1 + \left(\frac{1}{n}\sum_{i=1}^{n} |M_i^{\mathrm{A}} - M_i^{\mathrm{B}}|^p\right)^{1/p}} \tag{3}$$

where p indicates the L^p metrics used to calculate the inverse-scaled mean distance; for instance, $p = 1$ defines the Manhattan distance, while the Euclidean one is obtained for $p = 2$. In eq 3, M_i^{A} (M_i^{B}) is the i-descriptor of structure A (B) over a total of n descriptors.

We further note that $S_{\mathrm{A,B}}$ varies from 0 to 1. Accordingly, the dissimilarity between structures A and B is, then, defined as $D_{\mathrm{A,B}} = 1 - S_{\mathrm{A,B}}$; $D_{\mathrm{A,B}}$ can also vary from 0 (for identical structures) to a maximum dissimilarity of 1. In summary, the dissimilarity between two structures can be easily calculated, once the corresponding descriptor vectors $\mathbf{M}^{\mathrm{A}}$ and $\mathbf{M}^{\mathrm{B}}$ are obtained. In the following, we briefly describe how the descriptor shape vectors are computed for the four non-superimposing methods applied in this work.

USR. The USR method[14] considers the set of all atomic distances from four strategic reference locations in the structure, which provides it with sufficient discriminating power. The strategic anchor points are the structure centroid (*a*), the closest atom to the centroid (*b*), the farthest atom to the centroid (*c*) and the farthest atom to *c* (here designated as *d*). It is, then, possible to define four sets of distances from the *N* atoms to each anchor point. These sets may be regarded as distributions that are characterized by their moments. The USR method uses 12 descriptors enconding the shape of the structure that are based on the first moments of each distribution. In spite of applying directly the first, second, and third moments of each distribution (like in the original version of the USR method[14]), we followed here the suggestion of ref 45 to use moments with the same dimension, since it leads to a more balanced set of descriptors while showing to be extremely successful for prospective virtual screening.[46] Thus, the descriptors corresponding to the distribution of atomic distances (R_i^a) from the anchor point *a* are defined as[45]

$$M_1 = \frac{1}{N}\sum_{i=1}^{N} R_i^a \tag{4}$$

$$M_2 = \left[\frac{1}{N}\sum_{i=1}^{N}(R_i^a - M_1)^2\right]^{1/2} \tag{5}$$

$$M_3 = \left[\frac{1}{N}\sum_{i=1}^{N}(R_i^a - M_1)^3\right]^{1/3} \tag{6}$$

In addition, the corresponding sets of three descriptors associated with distributions of atomic distances from anchor points *b*, *c*, and *d* are calculated in a similar way: M_4, M_5, and M_6 are related to anchor point *b*, M_7, M_8, and M_9 use atomic distances from anchor point *c*, and M_{10}, M_{11}, and M_{12} are for anchor point *d*.

It is worth noting that M_1 corresponds to the first moment (i.e., the mean) of the distribution, which provides an estimate of the size of the structure. The descriptor M_2 is the standard deviation of the same distribution, which is related with the compactness of the atoms in the structure. Finally, M_3 is the cube root of the skewness of the distribution and, hence, is a measure of its asymmetry.

The USR method applies the Manhattan distance to measure the degree of similarity between two structures. Hence, we have used $p = 1$ in eq 3 to calculate $S_{A,B}$ and, from this, to obtain the corresponding dissimilarity, $D_{A,B}$.

CSR. The USR approach cannot distinguish between enantiomers, because all distances from the anchor points are preserved by reflections. To circumvent this disadvantage, Armstrong et al.[16] have proposed the CSR method, which is a variant of USR that is able to identify enantiomer structures.

In CSR method, the first three anchor points are similar to those defined in the USR approach. Indeed, anchor points *a*, *b*, and *c* are, respectively, the system centroid, the atom farthest from the centroid, and the atom farthest from that one. Conversely, the definition of anchor point *d* is done in such a way that the fourth distance distribution becomes different for two enantiomer structures. To achieve this purpose, one uses the previous anchor points to define two vectors: both $\vec{v}_1$ and $\vec{v}_2$ have their origin in anchor point *a*, and the direction of $\vec{v}_1$ ($\vec{v}_2$) is established by the line linking anchor points *a* and *b* (*c*). Then, we apply the cross product of $\vec{v}_1$ and $\vec{v}_2$ to define another vector:

$$\vec{v}_3 = \left(\frac{|\vec{v}_1| + |\vec{v}_2|}{2}\right)\frac{\vec{v}_1 \times \vec{v}_2}{|\vec{v}_1 \times \vec{v}_2|} \tag{7}$$

The term in parenthesis is used to normalize $\vec{v}_3$ so that it keeps the same units and similar magnitudes for the three vectors. Note that such normalization term slightly differs from the original paper,[16] where $|\vec{v}_1|/2$ was used. The present normalization has the advantage of accounting better for the possible difference of magnitude between $\vec{v}_1$ and $\vec{v}_2$. It is worth noting that $\vec{v}_1$ and $\vec{v}_2$, as well as $\vec{v}_3$, are invariant under translations of the structure and equivariant under rotations, but $\vec{v}_3$ changes its sign upon reflection.[16] Thus, using $\vec{v}_3$ to define the fourth anchor point, *d*, one guarantees that the corresponding distance distribution (and, hence, its first moments) is changed for the enantiomer of a given structure.

As in the USR method, 12 descriptors are calculated by using the mean, standard deviation, and cube root of skewness of the four distance distributions to the anchor points assigned in the CSR. Then, the Manhattan distance is also applied to calculate both $S_{A,B}$ and $D_{A,B}$.

CM Measure. The similarity (or dissimilarity) measure between two structures with the same number of particles (*N*) may be performed by considering the center of mass distances of all atoms of the system.[15] This approach has been applied in several works[15,23,25,47] from the area of cluster geometry optimization. Such a measure was used to keep a certain degree of diversity along the optimization procedure in population-based methods.

In the CM measure approach,[15] the shape descriptors coincide with the *N* atomic distances from the corresponding center of mass. However, these descriptors can only be sensitive to the differences in the shape of the two system if they are ordered by their magnitude before performing the comparison. This requires the use of a sorting algorithm, which is perhaps one of the main computational drawbacks in comparison with the other non-superimposing approaches applied in this work.

Once the *N* distances (i.e, the descriptors) of each structure have been ranked by their magnitude, the structural similarity may be calculated with eq 3, where the difference applies for descriptors of the same ranking position. According to previous works[15,23,25,47] on cluster geometry optimization (to which we want to relate the present results), the value of $S_{A,B}$ is calculated by using $p = 3$ in eq 3.

GSD. In a previous work[24] on global geometry optimization of argon clusters, we have shown that the hyperradius and the deformation indices[48] were able to characterize important structural differences on minima obtained with several potential functions. Motivated by this achievement, we propose here a global shape descriptor (GSD) approach that uses both hyperradius and deformation indices. These are global descriptors that, perhaps, encode significant structural dissimilarities between two systems. Whether GSD allows for an important discriminating power in the comparison of two structures is an issue that deserves to be investigated.

The hyperradius is related to the moments of inertia (I_1, I_2, and I_3) through the expression[48,49]

$$\rho^2 = \frac{I_1 + I_2 + I_3}{2M_N} \quad (8)$$

where M_N is the total mass of the system. In turn, the deformation indices are given as[48]

$$\xi_+ = \frac{I_1 - I_2}{M_N \rho^2} \quad (9)$$

$$\xi_- = \frac{I_3 - I_2}{M_N \rho^2} \quad (10)$$

where it was assumed that $I_1 \geq I_2 \geq I_3$ and, hence, $\xi_- \leq 0$ and $\xi_+ \geq 0$. We note that ρ represents a measure of the compactness of the structure, while ξ_+ and ξ_- characterizes its shape. In fact, on the basis of the values of the deformation indices, one may classify a given structure as spherical tops ($\xi_- = \xi_+ = 0$), prolate tops ($\xi_- < 0$ and $\xi_+ = 0$), oblate tops ($\xi_+ > 0$ and $\xi_- = 0$), and asymmetric tops ($\xi_- < 0$ and $\xi_+ > 0$).

Once ρ, ξ_+, and ξ_- are obtained, the GSD approach uses them as descriptors in eq 3. Just like for USR and CSR non-superimposing methods, the GSD algorithm applies the Manhattan distance to calculate both $S_{A,B}$ and $D_{A,B}$ (i.e., $p = 1$ in eq 3). Although it has to diagonalize the inertia tensor to obtain the principal moments of inertia and then order them, the GSD method has only three descriptors, which probably leads to a higher computational efficiency in comparison with the other non-superimposing approaches.

APPLICATION TO CLUSTER OPTIMIZATION

Hybrid EAs are state-of-the-art approaches for cluster geometry optimization.[17,24,50] In this framework, the EA performs a global exploration of the space, whereas a gradient-driven local search promotes a quick convergence to the nearest local optimum. It is well-known that the optimization efficacy of hybrid EAs is greatly enhanced if an appropriate level of diversity is maintained.[15,23,51] This ensures that the population is composed by different solutions, therefore postponing convergence and increasing the likelihood of discovering promising areas of the search space. Maintaining population diversity requires the application of a distance measure that estimates structural dissimilarity between solutions.

In this section, we describe a set of empirical tests that aim to determine if the measures previously described are accurate estimators for structural dissimilarity between two clusters. More specifically, we study the discriminating power of the above-described non-superimposing approaches by comparing their dissimilarity values with the rmsd obtained from the superimposing method. These tests require the generation of a set of increasingly dissimilar structures, which will be obtained through the application of a mutation operator. In the beginning, a random solution X is generated. Then, mutation helps to generate a Markovian random walk in the search space. We apply a sequence of N mutations and measure the distance between X and the clusters X^1, X^2, ..., X^N that are obtained after each step. In conformity with the described hybrid optimization framework, local search is applied after each mutation step and distance is measured using the solutions that result from this operation (i.e., dissimilarity is always estimated between two local optima). Local optimization is performed with the L-BFGS quasi-Newton method.[52] One expects that solutions obtained after the application of just a few mutations are structurally similar to the original cluster X, but as mutations start to accumulate, more dissimilar structures will tend to appear. This allows for a wide-ranging comparison between superimposing and non-superimposing methods.

Sigma mutation is the operator considered in this study. When modifying an atomic cluster, mutation is applied to atoms, i.e., it modifies the values of the three coordinates that specify the position of a particle in the 3D space. The new location of the particle is obtained by perturbing each coordinate with a random value sampled from a Gaussian distribution with mean 0 and standard deviation σ (which is a parameter of the optimization algorithm). As for molecular clusters, sigma mutation can either modify the center of mass of a given molecule or its orientation.

Atomic and molecular clusters are considered in this analysis: in the first case, we present results obtained with Ar_{38}, whereas for molecular cluster we selected $(H_2O)_{16}$. We performed additional tests for argon and water clusters with a different number of particles and obtained results similar to those presented here. The settings for all tests reported are the following: sample size (number of distinct solution pairs), 1000; mutation rate, 0.1; σ, 0.1.

The panels from Figure 7 display charts with the correlation between rmsd and the non-superimposing methods for the Ar_{38} cluster. Each entry in the charts displays the correlation between the rmsd of two solutions and the corresponding dissimilarity value estimated by a given measure. For completeness, we also present the correlation between rmsd and the variation in potential energy (panel e). It is interesting to notice that the application of mutation either generates solutions almost identical to the original cluster or substantially different structures (there are only a few rmsd values between 0.0 and 0.5 Å). This is probably a consequence of the application of local search to mutated solutions and suggests that different local optima correspond to distinct structures.

The best correlation with rmsd is obtained by the CM measure. Results show that CM distance is a fairly good estimator of the dissimilarity of two clusters: it correctly identifies similar solutions (when the rmsd is close to 0), and as the rmsd increases, there is a tendency for the CM distance to also increase. Additionally, it always classifies pairs with a high rmsd as dissimilar structures. Nonetheless, we show in Figure 8 two cases where the CM measure attributes dissimilarities of ~26% (panel a) and ~18% (panel b), while the rmsd values are 0.394 and 1.142 Å, respectively. In the first case, although the two structures almost coincide, the small differences arising in the corresponding distances to the center of mass are sufficient to get a large dissimilarity (in the normalized scale of the CM measure), when comparing with the corresponding rmsd value. Conversely, the second case has atoms that, clearly, cannot overlap each other. However, it is apparent from the figure that one has here a compensation effect, since the distances to the center of mass become similar for both structures.

The relationship between the other distance measures and rmsd is not so evident. In what concerns USR, one can see that it correctly identifies similar structures. However, results

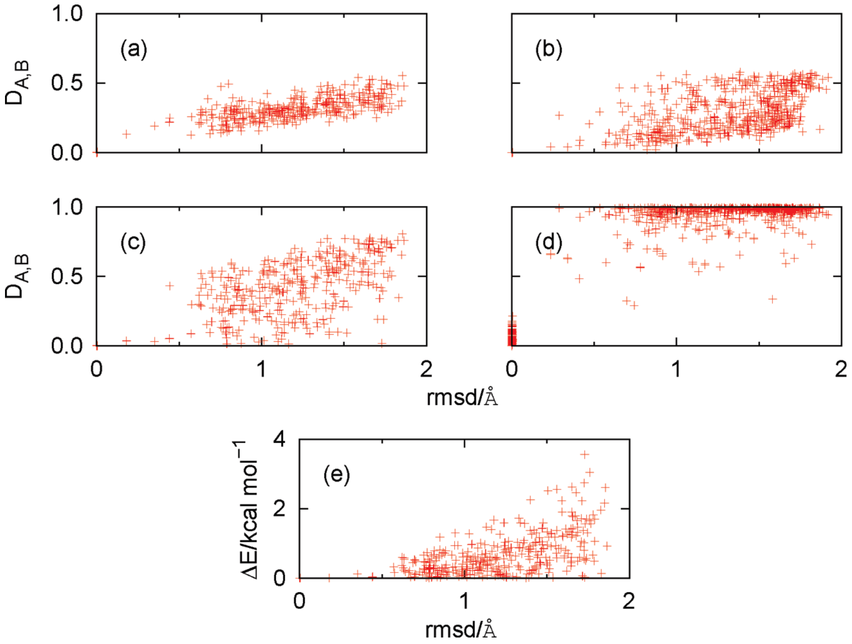

Figure 7. Correlation between various non-superimposing structural dissimilarity measures and the rmsd resulting from the best superposition of clusters of Ar_{38} (for 1000 comparisons): (a) CM measure, (b) USR distance, (c) GSD, (d) CSR, and (e) energy difference. The calculated Spearman rank correlation coefficients (with correction for ties) are 0.88, 0.72, 0.83, 0.57, and 0.66 for CM measure, USR, GSD, CSR, and energy difference, respectively. See the text.

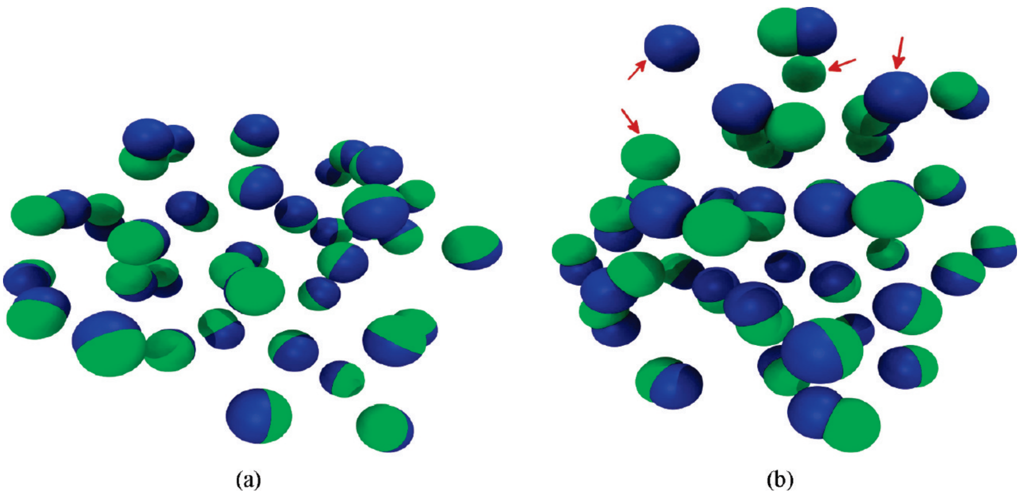

Figure 8. Best superposition of pairs of Ar_{38} clusters in two pathological cases for the CM measure: (a) $D_{A,B}$ = 0.26 and rmsd = 0.394 Å, (b) $D_{A,B}$ = 0.18 and rmsd = 1.142 Å. In panel b, arrows indicate the atoms that are not overlapping any other.

from structures with high rmsd are spread over a wide area, showing that USR is unable to provide an accurate estimate for the similarity between two argon clusters. The results obtained by GSD are similar to those achieved by USR. Once more, a large dispersion is visible in the area of high rmsd values, showing that GSD is unfit to identify distinct structures. There are many examples of structures with high rmsd that are classified as similar by GSD. This result can be explained by having in mind that GSD is based on global descriptors used to assign structures as spheres, oblate, prolate, or asymmetric tops. Obviously, this is an ambiguous measure, because the same global shape may correspond to different configurations of atoms. As for CSR, it has a poor correlation with rmsd. With the exception of identical clusters (those with rmsd close to 0.0 Å), this distance measure tends to consider all cluster pairs from the sample to be highly dissimilar, regardless of the true rmsd value. Finally, panel e from Figure 7 reveals a poor correlation between potential energy and structural changes. Results show that potential energy does not overestimate changes in structure, i.e., clusters with low rmsd have identical potential, but it fails to correctly classify dissimilar structures. This is in agreement with previous optimization studies that exposed the incapacity of potential energy to classify structural similarity.[15,23]

The panels from Figure 9 display charts with the correlation between rmsd and the non-superimposing methods for the $(H_2O)_{16}$ molecular cluster. A quick inspection of the charts immediately reveals a more uniform distribution of the rmsd values. This is probably due to the existence of structurally similar local optima, that just differ in the orientation of the rigid water molecules that compose the cluster. The correlation results follow the same trend of those obtained for the Ar_{38} cluster. CM exhibits the best correlation and it accurately classifies the similarity between structures. USR shows a reasonable correlation with rmsd and the dispersion identified in the test with argon clusters is not so

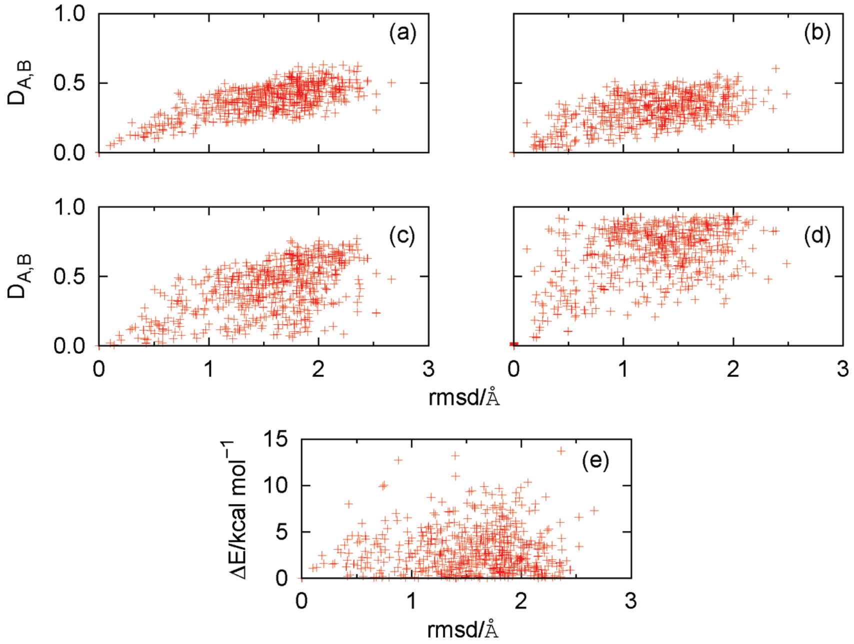

Figure 9. As in Figure 7, but for $(H_2O)_{16}$. The calculated Spearman rank correlation coefficients (with correction for ties) are 0.72, 0.69, 0.63, 0.54, and 0.28 for CM measure, USR, GSD, CSR, and energy difference, respectively. See the text.

evident here. As for GSD, it again fails to correctly classify dissimilar structures. In what concerns CSR, it still tends to overestimate the dissimilarity between clusters (although this effect is less visible here than it was with argon clusters). Anyway, CSR was able to identify and correctly classify similar structures. Finally, panel e confirms that it is not possible to establish a correlation between potential energy and structure.

In short, CM exhibits the best correlation with rmsd. This result is in agreement with optimization reports found in the literature[15,23] that point out this dissimilarity measure as a fast and accurate estimator for the structural dissimilarity between atomic clusters. One limitation of this approach is that it can only be applied to estimate the similarity of clusters with the same number of particles. USR has been successfully embedded in optimization algorithms used to detect low-energy water clusters,[53] but results presented here suggest that it might require a careful tuning to maximize its accuracy. As for GSD and CSR, the empirical analysis described in this section shows that they are not suitable to estimate structural dissimilarity.

CONCLUSIONS

We have extended the scope of a recently proposed[8] superimposing algorithm to include the enantiomer assignment. The method has been tested for several chemical structures subjected to a prior geometry optimization, followed by random rotation, translation, and shuffling of the atoms in the structure. The results show that the algorithm is robust and effective for the comparison of chemical structures and in discovering pairs of enantiomers resulting from geometry optimization. By using the present method, we were able to recognize, for the first time, that the two minimum structures of $(H_2O)_{16}$ with similar energies obtained, independently, by Wales and Hodges[42] and Takeuchi[31] are enantiomers. In addition, we have applied the superimposing algorithm to show that, among low-energy minimum structures[43] of Ar_{38}, those belonging to the same funnel present similar rmsd values (calculated after the best superposition is achieved), while being distinct from the global minimum located in the other funnel.

The superimposing algorithm was, then, used to gauge the effectiveness of simpler non-superimposing approaches to distinguish between structures that are different. Tests were carried out for both argon and water clusters. Since we are interested in discovering good diversity measures to apply in our EA, all tests were performed within the framework of global geometry optimization. It appears clear from the present results that CM measure is the best to estimate the population diversity during the geometry optimization with an EA. Moreover, this work confirms that both the CM measure and the USR approach are among the most promising methods to estimate diversity in global geometry optimization by population-based algorithms[15,23,25,47,53] like EAs; in contrast, approaches encoding the global shape of the structure (e.g., the GSD) are not adequate for measuring diversity. To our knowledge, this is the first study where different diversity measures have been tested against the unambiguous rmsd calculated after best superposition of pairs of structures. Thus, we advocate the application of the present methodology to test novel diversity measures before implementing them in global geometry optimization algorithms.

ACKNOWLEDGMENT

This work was supported by Fundação para a Ciência e Tecnologia (FCT), Portugal, under project PTDC/QUI/69422/2006, which is financed by Programa Operacional Factores de Competitividade (COMPETE) of QREN and FEDER programs. We are grateful to the John von Neumann Institut für Computing, Jülich, for the provision of supercomputer time on the Juropa-JSC (Project EPG01).

Supporting Information Available: Supplement_material.zip contains an easy-to-use Fortran code that implements the superimposing algorithm with assignment of enantiomers, as well as the files with the coordinates for several structures used in this work. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Nissink, J. W. M.; Verdonk, M. L.; Kroon, J.; Mietzner, T.; Klebe, G. Superposition of molecules: Electron density fitting by application of Fourier transforms. *J. Comput. Chem.* **1997**, *18*, 638–645.

(2) Gironés, X.; Robert, D.; Carbó-Dorca, R. TGSA. A molecular superposition program based on topo-geometrical considerations. *J. Comput. Chem.* **2001**, *22*, 255–263.

(3) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches. *J. Comput. Chem.* **1997**, *18*, 934–954.

(4) Barakat, M. T.; Dean, P. M. Molecular-structure matching by simulated annealing. 1. A comparison between different cooling schedules. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 295–316.

(5) Bayada, D. M.; Simpson, R. W.; Johnson, A. P.; Laurenço, C. An algorithm for the multiple common subgraph problem. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 680–685.

(6) Karney, C. F. F. Quaternions in molecular modeling. *J. Mol. Graphics Modell.* **2007**, *25*, 595–604.

(7) Flower, D. R. Rotational superposition: A review of methods. *J. Mol. Graphics Modell.* **1999**, *17*, 238–244.

(8) Vásquez-Pérez, J. M.; Martínez, G. U. G.; Köster, A. M.; Calaminici, P. The discovery of unexpected isomers in sodium heptamers by Born−Oppenheimer molecular dynamics. *J. Chem. Phys.* **2009**, *131*, 124126-1–124126-10.

(9) Bemis, G. W.; Kuntz, I. D. A fast and efficient method for 2D and 3D molecular shape-description. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 607–628.

(10) Nilakantan, R.; Baunman, N.; Venkataraghavan, R. New method for rapid characterization of molecular shapes−Applications in drug design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79–85.

(11) Good, A. C.; Ewing, T. J. A.; Gschwend, D. A.; Kuntz, I. D. New molecular shape descriptors−Application in database screening. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 1–12.

(12) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(13) Zauhar, R. J.; Moyna, G.; Tian, L. F.; Li, Z. J.; Welsh, W. J. Shape signatures: A new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* **2003**, *46*, 5674–5690.

(14) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711–1723.

(15) Grosso, A.; Locatelli, M.; Schoen, F. A population-based approach for hard global optimization problems based on dissimilarity measures. *Math. Program. Ser. A* **2007**, *110*, 373–404.

(16) Armstrong, M. S.; Morris, G. M.; Finn, P. W.; Sharma, R.; Richards, W. G. Molecular similarity including chirality. *J. Mol. Graphics Modell.* **2009**, *28*, 368–370.

(17) Deaven, D. M.; Ho, K. M. Molecular geometry optimization with a genetic algorithm. *Phys. Rev. Lett.* **1995**, *75*, 288–291.

(18) Gregurick, S. K.; Alexander, M. H.; Hartke, B. Global geometry optimization of $(Ar)_n$ and $B(Ar)_n$ clusters using a modified genetic algorithm. *J. Chem. Phys.* **1996**, *104*, 2684–2691.

(19) Niesse, J. A.; Mayne, H. R. Global optimization of atomic and molecular clusters using the space-fixed modified genetic algorithm method. *J. Comput. Chem.* **1997**, *18*, 1233–1244.

(20) Roberts, C.; Johnston, R. L.; Wilson, N. T. A genetic algorithm for the structural optimization of Morse clusters. *Theor. Chem. Acc.* **2000**, *104*, 123–130.

(21) Pereira, F. B.; Marques, J. M. C.; Leitão, T.; Tavares, J. Analysis of locality in hybrid evolutionary cluster optimization. *Proceedings of the 2006 IEEE Congress on Evolutionary Computation*; IEEE: New York, 2006; Vols. 1−6, pp 2270−2277.

(22) Pereira, F. B.; Marques, J. M. C.; Leitão, T.; Tavares, J. Designing efficient evolutionary algorithms for cluster optimization: A study on locality. In *Advances in Metaheuristics for Hard Optimization*; Springer Natural Computing Series; Siarry, P., Michalewicz, Z., Eds.; Springer: Berlin, 2008; pp 223−250.

(23) Pereira, F. B.; Marques, J. M. C. A study on diversity for cluster geometry optimization. *Evol. Intel.* **2009**, *2*, 121–140.

(24) Marques, J. M. C.; Pereira, F. B.; Leitão, T. On the use of different potential energy functions in rare-gas cluster optimization by genetic algorithms: Application to argon. *J. Phys. Chem. A* **2008**, *112*, 6079–6089.

(25) Marques, J. M. C.; Pereira, F. B. An evolutionary algorithm for global minimum search of binary atomic clusters. *Chem. Phys. Lett.* **2010**, *485*, 211–216.

(26) del Campo, J. M.; Köster, A. M. A hierarchical transition state search algorithm. *J. Chem. Phys.* **2008**, *129*, 024107−1024107−12.

(27) Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Res. Logistics Quart.* **1955**, *2*, 83–97.

(28) Munkres, J. Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.* **1957**, *5*, 32–38.

(29) Code obtained from http://jblevins.org/mirror/amiller/, which is a mirror of the Fortran source code repository of Alan Miller previously located at http://users.bigpond.net.au/amiller/ (accessed June 4, 2010).

(30) Kearsley, S. K. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr., Sect. A* **1989**, *45*, 208–210.

(31) Takeuchi, H. Development of an efficient geometry optimization method for water clusters. *J. Chem. Inf. Model.* **2008**, *48*, 2226–2233.

(32) Barron, L. D. On the definition of chirality. *Chem. Eur. J.* **1996**, *2*, 743–744.

(33) As mentioned in ref 32, Lord Kelvin stated that any geometrical figure (or group of points) has chirality if its image in a plane mirror, ideally realized, cannot be brought to coincide with itself (*cf.*, Lord Kelvin, Baltimore Lectures, C. J. Clay, London, 1904).

(34) Cahn, R. S.; Ingold, C.; Prelog, V. Specification of molecular chirality. *Angew. Chem., Int. Ed.* **1966**, *5*, 385–415.

(35) Prelog, V.; Helmchen, G. Basic principles of the CIP-system and proposals for a revision. *Angew. Chem., Int. Ed.* **1982**, *21*, 567–583.

(36) Schwartz, A. M.; Petitjean, M. [6.6]Chiralane: A remarkably symmetric chiral molecule. *Symmetry Cult. Sci.* **2008**, *19*, 307–316.

(37) *Marvin, 5.3.2*, ChemAxon, 2010, http://www.chemaxon.com (accessed Jun 1, 2010).

(38) Stewart, J. J. P. *MOPAC2009, 10.153 L*; Stewart Computational Chemistry: Colorado Springs, CO, 2008; http://OpenMOPAC.net (accessed Jun 1, 2010).

(39) Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.

(40) Stewart, J. J. P. Optimization of parameters for semiempirical methods II. Applications. *J. Comput. Chem.* **1989**, *10*, 221–264.

(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(42) Wales, D. J.; Hodges, M. P. Global minima of water clusters $(H_2O)_n$, $n \leq 21$, described by an empirical potential. *Chem. Phys. Lett.* **1998**, *286*, 65–72.

(43) Marques, J. M. C.; Pais, A. A. C. C.; Abreu, P. E. Generation and characterization of low-energy structures in atomic clusters. *J. Comput. Chem.* **2010**, *31*, 1495–1503.

(44) Wales, D. J. *Energy Landscapes: With Applications to Clusters, Biomolecules and Glasses*; Cambridge University Press; Cambridge, UK, 2003.

(45) Ballester, P. J.; Finn, P. W.; Richards, W. G. Ultrafast shape recognition: Evaluating a new ligand-based virtual screening. *J. Mol. Graphics Modell.* **2009**, *27*, 836–845.

(46) Ballester, P. J.; Westwood, I.; Laurieri, N.; Sim, E.; Richards, W. G. Prospective virtual screening with ultrafast shape recognition: The identification of novel inhibitors of arylamine *N*-acetyltransferases. *J. R. Soc. Interface* **2010**, *7*, 335–342.

(47) Cassioli, A.; Locatelli, M.; Schoen, F. Global optimization of binary Lennard-Jones clusters. *Optim. Methods Software* **2009**, *24*, 819–835.

(48) Aquilanti, V.; Lombardi, A.; Yurtsever, E. Global view of classical clusters: The hyperspherical approach to structure and dynamics. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5040–5051.

(49) Aquilanti, V.; Lombardi, A.; Sevryuk, M. B. Phase-space invariants for aggregates of particles: Hyperangular momenta and partitions of the classical kinetic energy. *J. Chem. Phys.* **2004**, *121*, 5579–5589.

(50) Johnston, R. L. Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries. *Dalton Trans.* **2003**, 4193–4207.

(51) Lee, J.; Lee, I.-H.; Lee, J. Unbiased global optimization of Lennard-Jones clusters for $N \leq 201$ by conformational space annealing method. *Phys. Rev. Lett.* **2003**, *91*, 080201.1080201.4.

(52) Liu, D.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program. B* **1989**, *45*, 503–528.

(53) Soh, H.; Ong, Y.-S.; Nguyen, Q. C.; Nguyen, Q. H.; Habibullah, M. S.; Hung, T.; Kuo, J.-L. Discovering unique, low-energy pure water isomers: Memetic exploration, optimization, and landscape analysis. *IEEE Trans. Evol. Comput.* **2010**, *14*, 419–437.

CI100219F