

# Geometrical eigen-subspace framework based molecular conformation representation for efficient structure recognition and comparison

Xiao-Tian Li,<sup>1</sup> Xiao-Bao Yang,<sup>1,2</sup> and Yu-Jun Zhao<sup>1,2,a)</sup>

<sup>1</sup>Department of Physics and School of Materials Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China

<sup>2</sup>Key Laboratory of Advanced Energy Storage Materials of Guangdong Province, South China University of Technology, Guangzhou, Guangdong 510640, China

(Received 24 February 2017; accepted 5 April 2017; published online 20 April 2017)

We have developed an extended distance matrix approach to study the molecular geometric configuration through spectral decomposition. It is shown that the positions of all atoms in the eigen-space can be specified precisely by their eigen-coordinates, while the refined atomic eigen-subspace projection array adopted in our approach is demonstrated to be a competent invariant in structure comparison. Furthermore, a visual eigen-subspace projection function (EPF) is derived to characterize the surrounding configuration of an atom naturally. A complete set of atomic EPFs constitute an intrinsic representation of molecular conformation, based on which the interatomic EPF distance and intermolecular EPF distance can be reasonably defined. Exemplified with a few cases, the intermolecular EPF distance shows exceptional rationality and efficiency in structure recognition and comparison. *Published by AIP Publishing.* [<http://dx.doi.org/10.1063/1.4981212>]

## I. INTRODUCTION

Molecular conformation plays an important role in solid state physics and materials science for its intrinsic connection with various physical properties. A reasonable configuration space in which the vectors characterize structures uniquely provides us not only with an insight of the intrinsic structures but also with a novel approach to material design, in particular in the age of Big Data. Recently, a flurry of global optimization methods were proposed for structure prediction,<sup>1–10</sup> where a technique to eliminate equivalent configurations from a huge number of candidates is of great demand for efficiency. Meanwhile, the high efficiency is also imperative in chemical/structural similarity searching,<sup>11,12</sup> which has been a basic retrieval mechanism in structure database. It is thus highly desirable to have a proper representation of molecular structures, based on which the intermolecular similarity can be reasonably defined and obtained. In chemistry and biology, while most of the research objects are valence-dominated organics, a traditional structural formula or a simplified notation such as SMILES<sup>13,14</sup> is sufficient to describe the molecular structure accurately. Moreover, by summarizing multiple atoms or functional groups into several types of nodes, one can get simpler molecular representations, such as a pharmacophoric pattern<sup>15–17</sup> and reduced molecular graph.<sup>18</sup> The description and comparison of macromolecules based on these representations turn out to be easy to carry out. In solid state physics, however, it is hard to find common structural features in most nanoparticles,<sup>19–23</sup> where an unbiased description of the molecular structure (robust enough against noise) and a related scheme

to quantify the difference between configurations are certainly helpful in the research community.

A complete set of atomic Cartesian coordinates can describe the detailed structure of a molecule accurately, but it is not convenient in structure comparison for its dependence on the coordinate frame and atomic ordering. It turns out that identical molecules with different coordinate frames and atomic orderings correspond to a set of equivalent atomic coordination descriptions. As a result, a reasonable intermolecular distance quantifying their structural difference should be a distance between different sets of equivalent descriptions. For instance, Sadeghi *et al.*<sup>24</sup> proposed a distance between molecules  $p$  and  $q$  by

$$\text{RMSD}(p, q) = \frac{1}{\sqrt{n}} \min_{\mathbf{P}, \mathbf{U}} \|\mathbf{R}^p - \mathbf{U}\mathbf{R}^q\mathbf{P}\|, \quad (1)$$

where  $\mathbf{R} \equiv (r_1, r_2, \dots, r_n) \in R^{3 \times n}$  is a detailed atomic coordination description of a molecule, the absolute symbol represents taking the Frobenius norm for coordinates of  $p$  and  $q$ , while  $\mathbf{U}$  and  $\mathbf{P}$  stand for a rotation (and reflection) and a permutation operation, respectively. Since the effect of translation has been eliminated beforehand by taking the centroid of each molecule to be the origin, the distance quantifies the molecular structural difference intrinsically, irrespective of the coordinate frame and the atomic ordering. Unfortunately, nowadays there is no algorithm to find the global root-mean-square distance (RMSD) within polynomial time.

A distance matrix is another complete representation of molecular conformation (except overall chirality). Containing all the interatomic distances, the distance matrix is sufficient to reconstruct the detailed structure of a molecule.<sup>25</sup> Despite the independence of coordinate frame, the distance matrix still relies on the atomic ordering, i.e., identical molecules

<sup>a)</sup> Author to whom correspondence should be addressed. Electronic mail: zhaoyj@scut.edu.cn. Tel.: +86-20-87110426. Fax: +86-20-87112837.

with different atomic orderings correspond to a set of distance matrices with permutation similarity to each other. As a result, it encounters the same problem as the atomic coordination description in structure comparison. While the atoms have been arranged in order, the corresponding distance matrix would be unique and easily managed. For instance, the distance matrix is widely utilized in protein for structure comparison,<sup>26,27</sup> benefited from their polypeptide chain structures. In general cases, however, it is not easy to obtain a unified atomic ordering for a molecule or nanocluster. Call *et al.*<sup>1</sup> proposed a coarse approach to arrange the atoms according to the atomic numbers and the distances to the center of mass of the molecule. The arrangement, however, is ambiguous and hard to adopt in practice, since the atomic azimuths have not been taken into account.

Impeded by the problems in structure comparison, most of the structure prediction methods<sup>3,6,7</sup> prefer fingerprints/descriptors to characterize configurations rather than complete representations. Constructed by structural invariants, fingerprints/descriptors are convenient in structure comparison for the independence of a coordinate frame and atomic ordering. A lot of molecular fingerprints/descriptors<sup>3,7,28</sup> were proposed based on the interatomic distances, the vital invariants under translation, reflection, and rotation in physics. Besides, the eigenvalues of the distance matrix are also very suitable for structure recognition<sup>24</sup> and analysis.<sup>29,30</sup> Despite the advantages in structure comparison, fingerprints/descriptors encounter a problem that they are hard to guarantee a unique configuration with incomplete structural information. For instance, there may be two distinct molecules with identical eigenvalues, an isomorphism problem.<sup>31</sup> As a result, a highly-qualified representation designed for structure recognition and comparison needs to balance accuracy and efficiency, i.e., it had better contain as much structural information as possible on the premise that the intermolecular distance can be easily carried out.

In this paper, we analyze the atomic projection information in the eigen-space based on an extended distance matrix, whose diagonal elements are assigned with the atomic numbers (or any other characters) of corresponding atoms. The derived eigen-subspace projection array (EPA) and eigen-subspace projection function (EPF) can characterize the surrounding configuration of an atom naturally, leading to the definition of the interatomic EPF distance and intermolecular EPF distance, which quantify the differences of atoms and molecules in a geometric structure, respectively. With the discussion of several cases, the intermolecular EPF distance shows an excellent performance in structure recognition and comparison.

## II. MODEL AND METHODS

We define the elements of the extended distance matrix of a molecule (or cluster) as

$$\begin{aligned} D_{ij} &= z_i, & \text{if } i = j, \\ &= d_{ij}, & \text{if } i \neq j, \end{aligned} \quad (2)$$

where  $z_i$  is the atomic number (or any other character) of atom  $i$  and  $d_{ij}$  is the Cartesian distance between atoms  $i$  and  $j$  in

arbitrary units. The matrix differs from the conventional defined one in the diagonal elements by replacing zeros with the atomic numbers, crucial to distinguish elements. Such a matrix contains all the structural information of a molecule except overall chirality.<sup>25</sup>

The extended distance matrix can be spectral factorized into a canonical form

$$\mathbf{D} = \sum_{k=1}^n \lambda_k \mathbf{u}_k \mathbf{u}_k^T, \quad (3)$$

where  $\lambda_1, \lambda_2, \dots, \lambda_n$  are all the eigenvalues of  $\mathbf{D}$  in the ascending order and  $\mathbf{u}_k$  are the corresponding eigenvectors. The  $i$ th component of eigenvector  $\mathbf{u}_k$  represents the projection of atom  $i$  on it. Typically, we can utilize the projections of an atom on the complete set of eigenvectors to characterize it in the eigen-space, i.e., a complete set of projections constitute the atomic eigen-coordinates with respect to the eigenvector basis. It is obvious that the eigenvector basis (associated with the eigenvalues) together with the atomic eigen-coordinates reserves all the information of the extended distance matrix.

Unfortunately, the eigenvector basis of identical molecule is not unique, especially in the degenerate cases that one can arbitrarily choose mutually orthogonal eigenvectors associated with the degenerate eigenvalues. It turns out that the atomic eigen-coordinates vary with the eigenvector basis, which discourages their application in structure recognition and comparison. To overcome the randomness of eigenvector basis, we suggest replacing it with a framework constructed by a complete set of eigen-subspaces. Accordingly, the atomic projections on eigenvectors are replaced by the norm of projections on eigen-subspaces. Here we define the eigen-subspace projection array (EPA) of atom  $i$  as

$$\mathbf{s}_i = (s_i^{\lambda_1}, s_i^{\lambda_2}, \dots, s_i^{\lambda_s}), \quad (4)$$

where  $\lambda_1, \lambda_2, \dots, \lambda_s$  are all the distinct eigenvalues of  $\mathbf{D}$  in the ascending order and  $s_i^{\lambda_k}$  is the norm of orthogonal projection of atom  $i$  on the eigen-subspace associated with  $\lambda_k$ . The eigen-subspace framework (associated with the distinct eigenvalues) together with the atomic EPAs constitutes a general representation of molecular conformation.

Admittedly, the eigen-subspace framework and atomic EPAs overlook some structural details with respect to the eigenvector basis and atomic eigen-coordinates. In fact, all the molecular projection information on the eigen-subspace associated with  $\lambda_k$  is contained in the projection matrix,

$$\mathbf{P}^{\lambda_k} = \sum_m \mathbf{u}_m^{\lambda_k} \mathbf{u}_m^{\lambda_k T}, \quad (5)$$

where  $\mathbf{u}_m^{\lambda_k}$  constitute a complete set of eigenvectors associated with  $\lambda_k$ . It turns out that the element  $P_{ij}^{\lambda_k}$  of  $\mathbf{P}^{\lambda_k}$  represents the dot product of projections of atoms  $i$  and  $j$ , i.e., the diagonal elements reserve the individual projection information of the atoms, while the non-diagonal ones for their correlations. From this perspective, the EPAs discard the tangled correlations between atomic projections, crucial in structure recognition and comparison. With refined information, the EPAs become easy and efficient to operate, providing an intrinsic insight of atomic geometric positions.

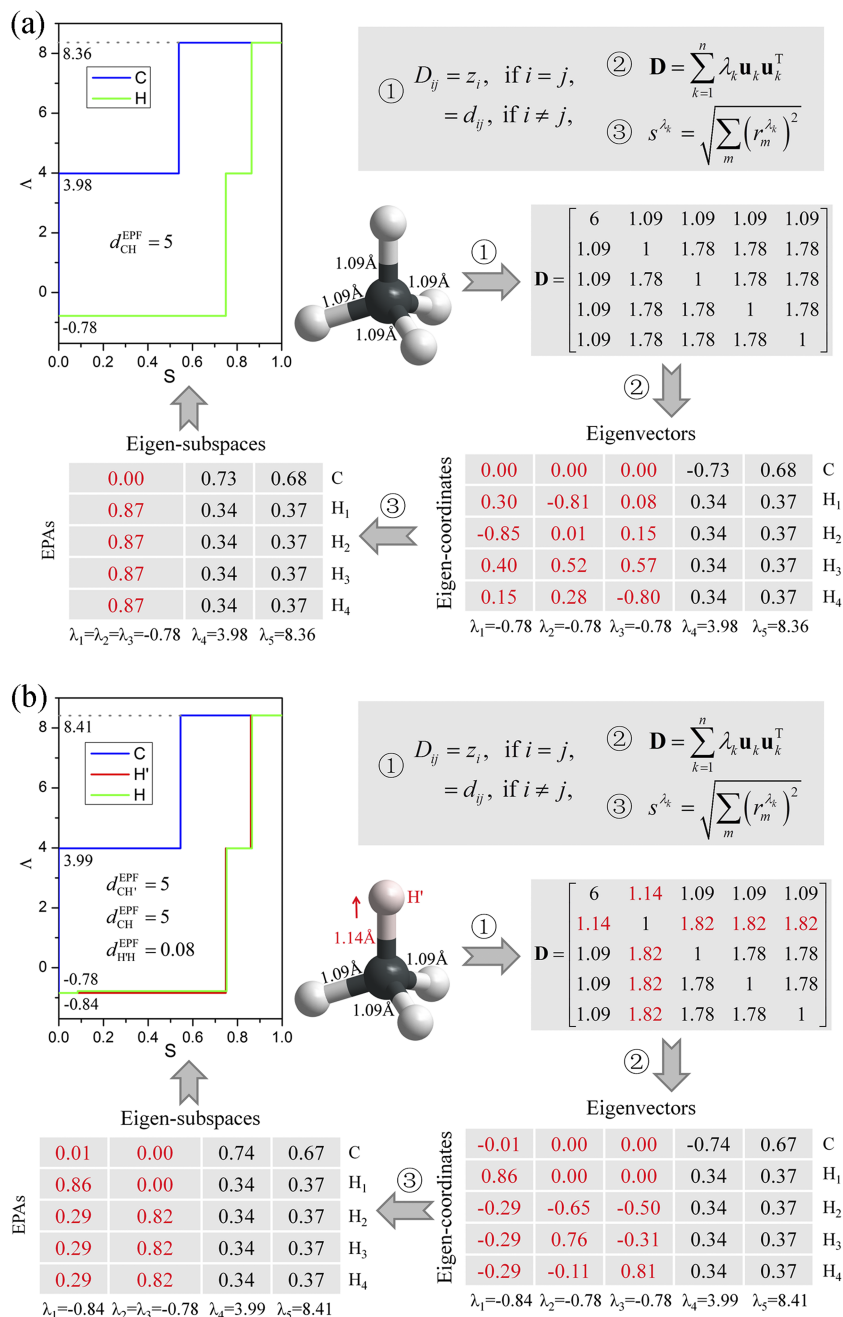


FIG. 1. The detailed structures, extended distance matrices, atomic eigen-coordinates, EPAs, and EPFs of (a) methane and (b) variant methane with one of its C–H bonds stretched from 1.09 Å to 1.14 Å. The columns highlighted in red in the eigen-coordinates and EPAs are corresponding to the triply degenerate eigen-subspace, requiring further treatment as the degenerate eigenvalues split. The notations for Equations (1) and (2) are explained in the main text. In Equation (3), the summation runs over the complete set of eigenvectors associated with eigenvalue  $\lambda_k$ . For instance, the first column of EPA in (a) is simply the square root of sum of the squares of the first three columns of the eigen-coordinates correspondingly.

For instance, Figure 1(a) shows the detailed structure and corresponding extended distance matrix of methane, from which we can get the atomic eigen-coordinates and EPAs. Due to the symmetry of methane, there are triply degenerate eigenvalues  $\lambda_1 = \lambda_2 = \lambda_3 = -0.78$ , associated with a three-dimensional eigen-subspace. The atomic eigen-coordinates specify the precise positions of all atoms in the eigen-space (especially in the triply degenerate eigen-subspace), but they rely on the eigenvector basis. In contrast, the atomic EPAs are very effective in structure recognition and comparison for their independence of eigenvector basis, although they can only specify the individual positions of atoms in the eigen-space (i.e., ignoring their correlation). Besides, equivalent atoms in a molecule would have identical EPAs naturally (such as the four equivalent H atoms in methane), critical in the structure analysis.

It is worth noting that the atomic EPA is not just a series of numbers but associates with the eigen-subspaces closely. Suppose we stretch one of the C–H bonds of methane from 1.09 Å to 1.14 Å, as shown in Fig. 1(b). As the symmetry broken, the triple eigenvalues split into a single one and a double one, leading to the splitting of the atomic EPAs. Note that the deviated H atom has a quite different EPA compared with the other three, although they have similar positions in the molecule. Besides, it is hard to determine the structural similarity for the ideal and slightly variant methane molecules by their atomic EPAs directly (Fig. 1). These irregularities would be easy to understand while the eigen-subspaces are taken into account. In fact, since  $\sum_{k=1}^s (s^{\lambda_k})^2 = 1$ , each atom can be viewed as a unit vector in the eigen-space, whose segments are projected on the complete set of eigen-subspaces. From this perspective, the

deviated H atom in the variant methane has a segment of  $0.86^2$  projected on the eigen-subspace associated with  $\lambda_1 = -0.84$ , while the other three H atoms have segments of  $0.29^2$  and  $0.82^2$  on  $\lambda_1 = -0.84$  and  $\lambda_2 = \lambda_3 = -0.78$  correspondingly, as seen in Fig. 1(b). Since similar segments of atomic vectors are projected on the split eigen-subspaces associated with numerically similar eigenvalues, the structural similarity between the H atoms in variant methane can be perceived.

To determine the structural similarity directly and more conveniently, one can get assisted from the atomic eigen-subspace projection function (EPF), which illustrates the relationship between the segments of a unit atomic vector in the eigen-space ( $S \in [0, 1]$ ) and the eigenvalues that the projected eigen-subspaces are associated with. The atomic EPFs of the variant methane are shown in Fig. 1(b), from which we easily observe that the red line (corresponding to the deviated H atom) is no remarkable different to the green ones (the other three H atoms), while the blue line (the C atom) is significantly different, consistent with the atomic positions in the molecule. Note that all EPFs increase monotonously since we have sorted the eigenvalues in the ascending order beforehand. As the EPF summarizes the surrounding configuration of an atom finely, a complete set of atomic EPFs constitute an intrinsic representation of molecular conformation.

Based on the atomic EPF, we define the EPF distance between atoms  $i$  and  $j$  to be

$$d_{ij}^{\text{EPF}} = \int_0^1 |\Lambda_i - \Lambda_j| dS, \quad (6)$$

where  $\Lambda_i$  and  $\Lambda_j$  represent the EPFs of atoms  $i$  and  $j$ , respectively. Note that the EPF distance quantifies the atomic difference in their surrounding configurations, different from their geometric distance in real space. For instance, in the variant methane,  $d_{\text{HH}}^{\text{EPF}} = 0$ , while  $d_{\text{H'H}}^{\text{EPF}} = 0.08$  (Fig. 1(b), where H' stands for the deviated H atom), revealing the equivalence of the three fixed H atoms and the degree of deviation of the stretched one.

Moreover, we point out here that the atomic EPF depends on the atomic number (or any other atomic character assigned in the diagonal element of  $\mathbf{D}$ ) to a great extent besides its surrounding configuration, since  $\int_0^1 \Lambda_i dS = z_i$ . As a result,

$$d_{ij}^{\text{EPF}} = \int_0^1 |\Lambda_i - \Lambda_j| dS \geq \left| \int_0^1 \Lambda_i dS - \int_0^1 \Lambda_j dS \right| = |z_i - z_j|. \quad (7)$$

It turns out that the EPF distance between atoms of different elements is always greater than the difference between their atomic numbers, critical to distinguish elements. For instance, the C atom in the variant methane has an exact EPF distance of 5 to all the H atoms as seen in Fig. 1(b), since its EPF is always larger than the ones of H atoms.

Following the interatomic EPF distance, we define the EPF distance between molecules  $p$  and  $q$  (both with  $n$  atoms) to be

$$d_{p,q}^{\text{EPF}} = \frac{1}{n} \min_{\{i,j\}} \sum_{i=j=1}^n d_{ij}^{\text{EPF}}, \quad (8)$$

where atoms  $i$  and  $j$  belong to molecules  $p$  and  $q$ , respectively, and the summation represents taking the accumulated difference of all atoms. Here each atom  $i \in p$  is associated with an atom  $j \in q$ , and we take the atoms of two molecules matched in such a way that the accumulated distance is minimized. This is actually an assignment problem, which can be easily approached by the Hungarian algorithm.<sup>32</sup> With the atoms properly matched, the equivalence of two molecules can be easily judged by their corresponding distance matrices. The EPAs and EPFs turn out to be subtle fingerprints of atoms reflecting their surrounding configurations and are expected to be applicable to periodic systems with a specific treatment of the translational symmetry, save the chiral structures due to the intrinsic liability of the distance matrix. To be clear, the method does not distinguish chiral molecules and is still not suitable for periodic structures so far.

### III. DISCUSSIONS

As mentioned above, the eigenvalues of the distance matrix encounter the isomorphism problem in structure recognition that there may be two distinct configurations with identical eigenvalues. For instance, Figure 2(a) shows two twenty-one-atom triangular fragments (TFs) with identical eigenvalues despite their structural nonequivalence, clearly reflected by the shadowed outlines. While the eigenvalues are incapable of distinguishing the two TFs, we can get assisted from the atomic projection information on the eigen-subspaces. Here we derive the atomic eigen-spectra from the EPAs for a better view by the Lorentz expansion,

$$I_i(\lambda) = \sum_{k=1}^s s_i^{\lambda_k} \frac{\sigma/\pi}{(\lambda - \lambda_k)^2 + \sigma^2}, \quad (9)$$

where  $\lambda_1, \lambda_2, \dots, \lambda_s$  are all the distinct eigenvalues of  $\mathbf{D}$  in the ascending order,  $s_i^{\lambda_k}$  is the EPA of atom  $i$  on the eigen-subspace associated with  $\lambda_k$ , and  $\sigma$  is the Lorentzian width parameter (set to be 0.2 here). In Fig. 2, the atoms of the two TFs are associated with each other according to their interatomic EPF distances in the ascending order by the Hungarian algorithm. Although the atomic eigen-spectra of the two TFs have peaks at the same positions in general due to their identical eigenvalues, the structural nonequivalence can be clearly indicated, especially from the inset of their difference (Fig. 2(b)). Their intermolecular EPF distance corresponds to be 0.331 (see Eq. (8)). Of note, the spectra of atoms 1, 2, and 3 of the two TFs are identical, consistent with the fact that they have the same distances with other 20 atoms in both TFs, reflecting a reasonable association between the atoms in the molecules.

In order to test the accuracy and efficiency of the intermolecular EPF distance, we compare it with the RMSD distance. We perform a molecular dynamics (MD) simulation at 300 K for the ground state of  $\text{B}_{36}$  clusters,<sup>33,34</sup> a quasi-planar configuration with a central hexagonal hole. The MD simulation is carried out based on the density functional theory (DFT) implemented in the Vienna *ab initio* simulation package (VASP).<sup>35,36</sup> The projector augmented wave (PAW) and the Perdew-Burke-Ernzerhof (PBE) of Generalized Gradient Approximation (GGA) functional<sup>37</sup> are employed for the

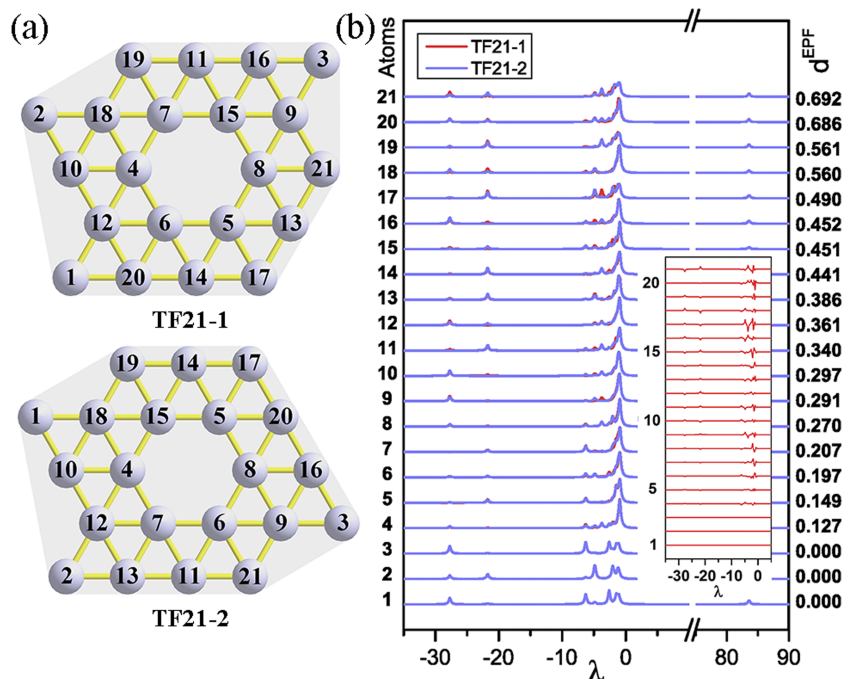


FIG. 2. (a) Two TFs with distance matrices of identical eigenvalues and (b) their atomic eigen-spectra with an inset of their difference. The atoms of the two TFs are matched and arranged according to their interatomic EPF distances, which are marked on the right side of the eigen-spectra.

calculations. The cutoff energy is 500 eV and the vacuum distance is set to be 20 Å. Using the Nosé thermostat<sup>38,39</sup> approach, we performed the constant-temperature MD simulation for  $B_{36}$  with a time step of 1 fs.

The RMSD and EPF distances between each pair of 50 configurations, taken from the MD simulation after equilibrium, are shown in Fig. 3. It is obvious that there is a strong positive correlation between the RMSD and EPF distances with a correlation coefficient of 0.986. Of note, the RMSD distances, difficult to carry out in general, are obtained here for  $B_{36}$  clusters benefited from their quasiplanar structures. We rotate (and reverse/reflect) one of the two given configurations along the normal of the quasiplane for every  $\frac{2\pi}{500}$  rad and perform the permutation and the rotation (and reflection) operation alternately until the RMSD distance reaches a local minimum. Since the two configurations have been matched

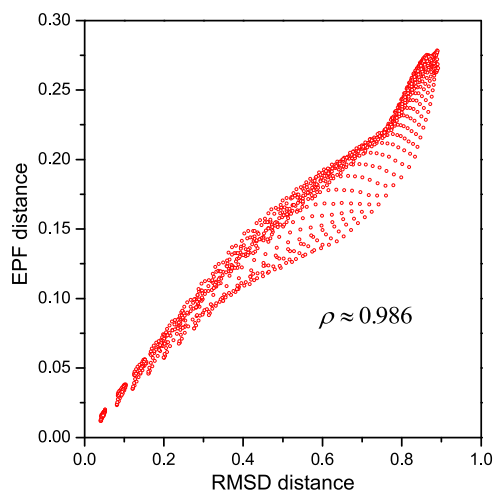


FIG. 3. Correlation of the RMSD and EPF distances of various pairs (each point in the plot corresponding to a pair) of 50  $B_{36}$  configurations taken from the MD simulation. The correlation coefficient is shown to be 0.986.

considering a lot of potential azimuths, we can get the global minimum from all the local minima. However, this is a time-consuming process, expending about 500 times as much as that by the EPF distance approach. Moreover, the RMSD approach is hard to generalize to three dimensional configurations, while our EPF distance has no such limit. It turns out that our intermolecular EPF distance shows a good consistency with the RMSD distance with much higher efficiency.

While the surrounding configuration of each atom can be depicted by an EPF, molecules turn out to locate in a configuration space constructed by a series of EPFs. Such a configuration space provides us not only a visual knowledge of the geometric correlations between configurations but also some specific approaches to practical issues. For instance, we explore the

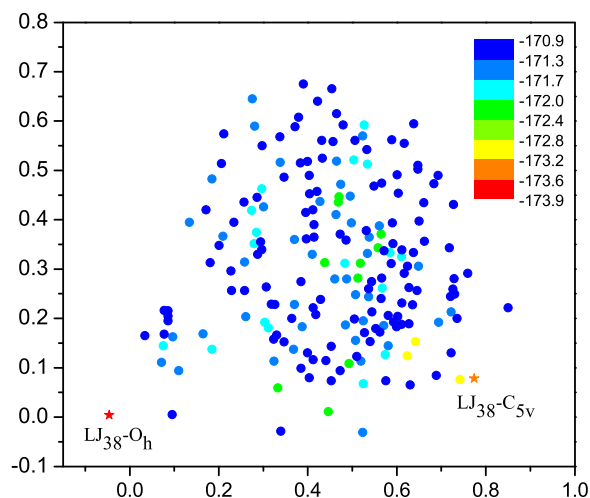


FIG. 4. 2D landscape of the potential energy surface of  $LJ_{38}$  cluster. Each point here represents a local minimum configuration, whose position on the 2D map is determined by minimizing  $\sum_{p < q} (d_{pq}^{2D} - d_{pq}^{EPF})^2$ . The energies of the configurations are indicated by their colors, while the global minimum  $LJ_{38}-O_h$  and the second-minimum  $LJ_{38}-C_{5v}$  are marked by stars.

potential energy surface of LJ<sub>38</sub> cluster (LJ cluster is consisted of identical atoms interacting by a pair Lennard-Jones potential) and project 200 local minima on a 2D map (Fig. 4), from which we can get an intuitive impression of the configurations. In particular, the global minimum LJ<sub>38</sub>-O<sub>h</sub> is far from the other local minima in contrast to the second-minimum LJ<sub>38</sub>-C<sub>5v</sub>, illustrating clearly why LJ<sub>38</sub>-O<sub>h</sub> is very difficult to obtain in most global optimization methods. Besides, we can conjecture the potential transition paths between configurations with the knowledge of their geometric correlations, critical to determine the transition states.

#### IV. SUMMARY

Based on the extended distance matrix, a representation has been proposed to describe the molecular conformation by projecting the atomic geometric information on the eigen-subspaces. Neglecting the correlations between atomic projections, the refined EPA and EPF reflect the surrounding configuration of an atom nicely. Moreover, the EPF based approach is demonstrated to be *competent* in the isomorphism problem and very *intuitive* and *efficient* in structure comparison.

#### ACKNOWLEDGMENTS

This work is financially supported by NSFC (Grant Nos. 11574088 and 51431001), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (Grant No. 51621001), and Natural Science Foundation of Guangdong Province of China (Grant No. 2016A030312011).

<sup>1</sup>S. T. Call, D. Y. Zubarev, and A. I. Boldyrev, *J. Comput. Chem.* **28**, 1177 (2007).

<sup>2</sup>A. R. Oganov, A. O. Lyakhov, and M. Valle, *Acc. Chem. Res.* **44**, 227 (2011).

<sup>3</sup>A. O. Lyakhov *et al.*, *Comput. Phys. Commun.* **184**, 1172 (2013).

<sup>4</sup>Y. Wang *et al.*, *Phys. Rev. B* **82**, 094116 (2010).

<sup>5</sup>Y. Wang *et al.*, *Comput. Phys. Commun.* **183**, 2063 (2012).

<sup>6</sup>J. Lv *et al.*, *J. Chem. Phys.* **137**, 084104 (2012).

<sup>7</sup>L. B. Vilhelmsen and B. Hammer, *J. Chem. Phys.* **141**, 044711 (2014).

<sup>8</sup>R. Martoňák, A. Laio, and M. Parrinello, *Phys. Rev. Lett.* **90**, 075503 (2003).

<sup>9</sup>S. Goedecker, *J. Chem. Phys.* **120**, 9911 (2004).

<sup>10</sup>Q. Zhu, A. R. Oganov, and A. O. Lyakhov, *CrystEngComm* **14**, 3596 (2012).

<sup>11</sup>P. Willett, J. M. Barnard, and G. M. Downs, *J. Chem. Inf. Model.* **38**, 983 (1998).

<sup>12</sup>D. Stumpfe and J. Bajorath, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **1**, 260 (2011).

<sup>13</sup>D. Weininger, *J. Chem. Inf. Model.* **28**, 31 (1988).

<sup>14</sup>D. Weininger, A. Weininger, and J. L. Weininger, *J. Chem. Inf. Model.* **29**, 97 (1989).

<sup>15</sup>S. E. Jakes and P. Willett, *J. Mol. Graphics* **4**, 12 (1986).

<sup>16</sup>S. E. Jakes *et al.*, *J. Mol. Graphics* **5**, 41 (1987).

<sup>17</sup>A. T. Brint and P. Willett, *J. Mol. Graphics* **5**, 49 (1987).

<sup>18</sup>Y. Takahashi, M. Sukekawa, and S. Sasaki, *J. Chem. Inf. Model.* **32**, 639 (1992).

<sup>19</sup>H.-J. Zhai *et al.*, *Nat. Mater.* **2**, 827 (2003).

<sup>20</sup>J. E. Dahl, S. G. Liu, and R. M. K. Carlson, *Science* **299**, 96 (2003).

<sup>21</sup>D. P. Kosimov, A. A. Dzhrakhalov, and F. M. Peeters, *Phys. Rev. B* **81**, 195414 (2010).

<sup>22</sup>H. Sun *et al.*, *Comput. Theor. Chem.* **1062**, 74 (2015).

<sup>23</sup>H. Häkkinen *et al.*, *J. Phys. Chem. A* **107**, 6168 (2003).

<sup>24</sup>A. Sadeghi *et al.*, *J. Chem. Phys.* **139**, 184118 (2013).

<sup>25</sup>T. Havel, I. Kuntz, and G. Crippen, *Bull. Math. Biol.* **45**, 665 (1983).

<sup>26</sup>L. Holm and C. Sander, *J. Mol. Biol.* **233**, 123 (1993).

<sup>27</sup>I.-G. Choi, J. Kwon, and S.-H. Kim, *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3797 (2004).

<sup>28</sup>A. R. Oganov and M. Valle, *J. Chem. Phys.* **130**, 104504 (2009).

<sup>29</sup>K. Balasubramanian, *Chem. Phys. Lett.* **239**, 117 (1995).

<sup>30</sup>A. Kloczkowski *et al.*, *J. Struct. Funct. Genomics* **10**, 67 (2009).

<sup>31</sup>K. Balasubramanian, *Chem. Phys. Lett.* **169**, 224 (1990).

<sup>32</sup>H. W. Kuhn, *Nav. Res. Logist. Q.* **2**, 83 (1955).

<sup>33</sup>Z. A. Piazza *et al.*, *Nat. Commun.* **5**, 3113 (2014).

<sup>34</sup>S.-G. Xu *et al.*, *J. Chem. Phys.* **142**, 214307 (2015).

<sup>35</sup>G. Kresse and J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).

<sup>36</sup>G. Kresse and D. Joubert, *Phys. Rev. B* **59**, 1758 (1999).

<sup>37</sup>J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).

<sup>38</sup>S. Nosé, *J. Chem. Phys.* **81**, 511 (1984).

<sup>39</sup>S. Nosé, *Mol. Phys.* **100**, 191 (2002).