Analysis of trajectory similarity and configuration similarity in on-the-fly surface-hopping simulation on multi-channel nonadiabatic photoisomerization dynamics

Xusong Li,[1,2,3] Deping Hu,[1,3] Yu Xie,[4] and Zhenggang Lan[1,2,4,a)]

[1]*CAS Key Laboratory of Biobased Materials, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao 266101, China*

[2]*Sino-Danish Center for Education and Research/Sino-Danish College, University of Chinese Academy of Sciences, Beijing 100049, China*

[3]*University of Chinese Academy of Sciences, Beijing 100049, China*

[4]*The Environmental Research Institute, MOE Key Laboratory of Theoretical Chemistry of Environment, South China Normal University, Guangzhou 510006, China*

(Received 11 July 2018; accepted 26 November 2018; published online 26 December 2018)

We propose an "automatic" approach to analyze the results of the on-the-fly trajectory surface hopping simulation on the multi-channel nonadiabatic photoisomerization dynamics by considering the trajectory similarity and the configuration similarity. We choose a representative system phytochromobilin (PΦB) chromophore model to illustrate the analysis protocol. After a large number of trajectories are obtained, it is possible to define the similarity of different trajectories by the Fréchet distance and to employ the trajectory clustering analysis to divide all trajectories into several clusters. Each cluster in principle represents a photoinduced isomerization reaction channel. This idea provides an effective approach to understand the branching ratio of the multi-channel photoisomerization dynamics. For each cluster, the dimensionality reduction is employed to understand the configuration similarity in the trajectory propagation, which provides the understanding of the major geometry evolution features in each reaction channel. The results show that this analysis protocol not only assigns all trajectories into different photoisomerization reaction channels but also extracts the major molecular motion without the requirement of the pre-known knowledge of the active photoisomerization site. As a side product of this analysis tool, it is also easy to find the so-called "typical" or "representative" trajectory for each reaction channel. *Published by AIP Publishing.* https://doi.org/10.1063/1.5048049

I. INTRODUCTION

Photoinduced isomerization reactions via the double-bond twisting motions on molecular excited states widely exist in photochemistry.[1–4] For instance, the photoisomerization processes of the chromophores in photoreceptor proteins are the primary steps in the solar-to-mechanical energy conversions, which trigger important photoinduced biological functions.[1,2,4–6] The photoisomerization mechanism received considerable research interests in the last decades.[1–5,7,8] Among these studies, theoretical calculations clarified that nonadiabatic dynamics at conical intersections are essential for photoisomerization processes.[1,2] The simulation of nonadiabatic dynamics needs to take the coupled electron-nucleus motion into account, in which Born-Oppenheimer approximation breaks down.[9,10] Although many theoretical approaches were proposed to solve nonadiabatic dynamics,[2,3,9–41] trajectory surface hopping (TSH) approaches become popular due to their simplicity and easy implementation.[32,34,40–54] With the development of computational facilities, the on-the-fly TSH dynamics provides us a reasonable way to simulate the nonadiabatic dynamics of polyatomic molecules by the inclusion of all degrees of freedom.[7,34,41–47,51,52,55–68] Nowadays, the combination of the on-the-fly dynamics and TSH (or other theoretical approaches) becomes a promising tool to understand the photoisomerization mechanism at the atomic level.[1–3,7,8,41–43,45–47,51,60,63,65,69–76]

The on-the-fly TSH dynamics often requires the computation of a large number of trajectories. The statistical analysis over all trajectories gives various dynamical features, for instance, the excited-state population decay, the structure evolution, and the geometrical features at potential energy surface (PES) crossings. In the typical analysis of the TSH results, the active reaction coordinates are normally identified by the eye view of many trajectories and the results are discussed by the explanation of a few "representative" trajectories.[42,43] This approach also largely relies on the preliminary understanding of the nonadiabatic dynamics, such as the reaction pathways and the relevant conical intersections. This "eye-view" analysis routine becomes not easy when the system size becomes large, the complicated molecular motions are involved, many trajectories are concerned, or the pre-known knowledge on the reaction channels is missing. Thus, the

a)Author to whom correspondence should be addressed: zhenggang.lan@m.scnu.edu.cn and zhenggang.lan@gmail.com. Tel.: +86-532-80662630. Fax: +86-532-80662778.

novel analysis tool should be developed to examine the TSH simulation results, particularly because more and more studies take the on-the-fly TSH calculations to treat the different nonadiabatic dynamics of various complicated systems.[41] As a typical example, the analysis of the TSH simulation on the photoisomerization dynamics is not trivial because the twisting motions may happen at different twisting sites, the major motion may involve the strong couplings between different nuclear degrees of freedom, and several reaction channels may result in different photoproducts.

Unsupervised Machine Learning (ML) algorithms, particularly dimensionality reduction approaches, such as principle component analysis (PCA),[77–79] multidimensional scaling (MDS),[80,81] isometric feature mapping (ISOMAP),[82,83] diffusion map,[84,85] autoencoder,[86] etc., were employed to examine the main feature of the geometrical evolution in the ground-state molecular dynamics simulation.[87–94] In recent years, some groups tried to use such tools in the analysis of nonadiabatic dynamics,[95–100] which tried to automatically extract the main geometrical feature of the trajectory evolution. The underlining idea is as follows. A single geometry in a trajectory is represented by a point in a high-dimensional coordinate space. After the collection of a large number of geometries generated by the trajectory propagation in the nonadiabatic dynamics, these unsupervised ML approaches construct a mapping from the high-dimensional space to a low-dimensional space, which tries to conserve the pattern feature of data point distribution. The active motion responsible for the nonadiabatic dynamics was then examined in the low-dimensional space. These efforts help us to understand the geometric evolution in the nonadiabatic dynamics. However, the application of these approaches in real analysis tasks may not be fully straightforward. For instance, such an idea may not work properly in the multi-channel situations because different reactive coordinates may be responsible for different channels. Most importantly, the analysis in the configuration space does not directly take an important dynamic feature, namely, "time evolution," into account. Instead the time feature is indirectly included afterwards, through monitoring the movement of the dataset in the low-dimensional space constructed by the dimensionality reduction.

In this paper, we propose an improved "automatic" approach to analyze the on-the-fly TSH results by reconsidering the concept of "trajectory evolution with time being." Instead of only performing the dimensionality reduction in coordinate space, we also examine the trajectory evolution in the so-called "trajectory space," in which we measure the "distance" or "dissimilarity" between different trajectories. The estimation of the trajectory similarity is widely employed in various scientific fields.[101–111] In the current work, the so-called Fréchet distance[108,112–114] was taken to evaluate the "dissimilarity" between two trajectories. After the construction of the pair-wise dissimilarity matrix for all trajectories, the clustering method is employed to assign the trajectories into different groups. In this trajectory clustering analysis, each group in principle should represent a reaction channel. The reactive coordinate responsible for each channel is further identified by the dimensionality reduction approaches in the coordinate space, suggested in our previous work.[95] Overall, this analysis considers first the trajectory similarity and second the configuration similarity, which makes the analysis procedure more transparent and automatic. This provides us a powerful tool to analyze the nonadiabatic dynamics with many reactive channels.

As the first attempt, we wish to know whether the above idea can clearly identify distinguishing channels and clarify their active motion in the photoisomerization dynamics. The reason is that the photoisomerization serves a kind of prototype reactions, in which the twisting motions at different sites give rather different reaction channels and several distinguishing photoproducts are formed as a result.[1–4] Thus, in principle, this type of the nonadiabatic dynamics provided us a very good model to examine our idea on the estimation of the trajectory similarity and the geometrical similarity. In this work, we take the phytochromobilin (ZaZsZa PΦB in Fig. 1) model as an example to check the performance of the above proposed analysis method. As widely existing plant's photoreceptors, the PΦB and other phytochromes were studied extensively.[1,6,70,115–124] The PΦB system decays to the ground state via different conical intersections, and finally, several photo-products are formed.[1,70,124] Thus, the PΦB model is an ideal system to test our new approach. The results show that the analysis approach with the combination of the trajectory similarity and the configuration similarity is a very powerful protocol that can perform the automatic and efficient analysis of nonadiabatic photoisomerization dynamics with several channels and different products. Although the current work is based on the TSH calculations of photoisomerization, it is also possible to use the similar idea to understand other types of trajectory-based nonadiabatic dynamics simulation.

This work is organized as follows. Section II outlines theoretical methods, implementation, and computational details. Section III shows the results, and Sec. IV provides the discussions. Section V gives the conclusion of the current work.

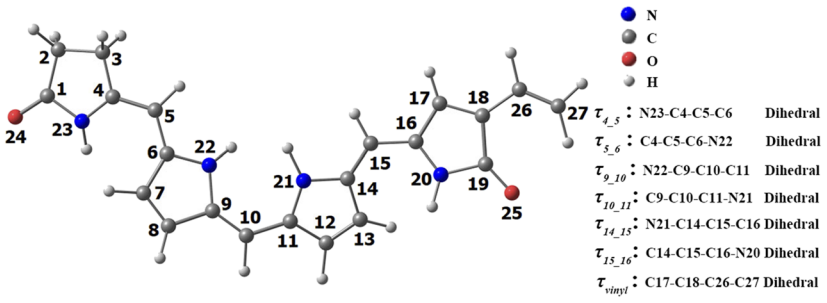

FIG. 1. The model of the ZaZsZa isomer of PΦB and some key coordinates.

II. THEORETICAL METHODS AND COMPUTATIONAL DETAILS

A. Theoretical methods

1. Trajectory surface hopping dynamics

Many previous studies have provided the detailed discussion on Tully's TSH approaches,[32,42,43,46,47,52] so we outline the main concept here. In Tully's TSH framework, the nuclear motion was treated by the classical Newtonian mechanics, while the electronic motion was described by the quantum evolution. The nonadiabatic transitions were described by the trajectory hops between different electronic states, and the hopping probability was determined by Tully's fewest switches algorithm.[32] The initial conditions (such as geometries and velocities) were sampled by the Wigner distribution of the ground vibrational level of the normal modes on the electronic ground state.

2. Configuration similarity definition

Mathematically, a single geometry is represented by a point in a high-dimensional coordinate space, which is characterized by a high-dimensional vector. Thus, the similarity/dissimilarity between two geometry snapshots is measured by the distance between the two corresponding points in the metric view. This provides us the basic idea on the definition of the dissimilarity matrix **D** over all snapshots. Following previous work,[95–97] the elements d_{ij} in the **D** matrix were defined by the root mean square distance (RMSD) of two configurations. In the RMSD calculations, it is necessary to remove the contribution of translational and rotational motions.[95,125,126]

3. Trajectory similarity definition

One of central ideas in this work involves the definition of the similarity between two trajectories. Different numerical approaches were proposed to compute the trajectory similarity,[101–114] such as the Hausdorff distance, Fréchet distance, and so on.[102,113] Among them, the Fréchet distance is a good candidate to conduct the analysis of trajectory evolution because the chronological order is taken into account explicitly and this analysis approach also does not require the same propagation duration for all trajectories.[102,113,127]

Roughly speaking, it is possible to understand the Fréchet distance in an intuitive way. Let us assume that a man is walking along a path **P** and his dog is running along another path **Q**. They are connected by a leash in the whole walking procedure. Both starting and ending points are known for path **P** and path **Q**. The man and his dog move along their own pathways independently under constrain that their motion must follow the monotonic chronological way from the starting point to the ending point, and no backward movement is allowed. When the dog changes speed to make the leash as slack as possible, the length of the shortest leash sufficient for both the man and his dog moving along their own paths defines the Fréchet distance between two curves **P** and **Q**.

Next, we discussed the formal mathematical view of the Fréchet distance.[108,112–114] Suppose that **P** and **Q** are two given curves in the metric space $\boldsymbol{V}_s$, which are represented by the continuous mappings as follows:

$$\begin{aligned} \mathbf{P}: &\quad [p_0,\, p_1] \to \boldsymbol{V}_{\mathrm{s}} \quad [p_0, p_1 \in \boldsymbol{R}_{\mathrm{s}}, \quad p_0 \le p_1], \\ \mathbf{Q}: &\quad [q_0,\, q_1] \to \boldsymbol{V}_{\mathrm{s}} \quad [q_0, q_1 \in \boldsymbol{R}_{\mathrm{s}}, \quad q_0 \le q_1], \end{aligned} \tag{1}$$

where p_0 and p_1 (or q_0 and q_1) are the starting and ending points of the curve **P** (or **Q**) in the space of $\boldsymbol{R}_{\mathrm{s}}$. The Fréchet distance between **P** and **Q** is defined as

$$\delta_F(\mathbf{P}, \mathbf{Q}) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \{dist[\mathbf{P}(\alpha(t)), \mathbf{Q}(\beta(t))]\}, \tag{2}$$

where α(t) [or β(t)] is an arbitrary continuous non-decreasing function that maps the unit interval [0,1] onto $[p_0, p_1]$ (or $[q_0, q_1]$), namely, $\alpha(0) = p_0$ and $\alpha(1) = p_1$ [or $\beta(0) = q_0$ and $\beta(1) = q_1$].

For computational practices, an arbitrary continuous curve is typically approximated by a polygonal curve, and thus the discrete Fréchet distance,[108,112–114] instead of its continuous counterpart, is often used to examine the dissimilarity of two polygonal curves.

Two trajectories **P** and **Q** are approximated by polygonal curves represented by two sequences **S(P)** $(p_1, \ldots p_{\mathrm{i}} \ldots, p_{\mathrm{n}})$ and **S(Q)** $(q_1, \ldots, q_{\mathrm{j}}, \ldots, q_{\mathrm{m}})$, where p_{i} is the i-th snapshot of the trajectory **P** and q_{j} is the j-th snapshot of the trajectory **Q**. The coupling **C** between **P** and **Q** in the production space **S(P)** × **S(Q)** is given by a sequence

$$\mathbf{C}(\mathbf{P}, \mathbf{Q}) \equiv (p_{a_1}, q_{b_1}), (p_{a_2}, q_{b_2}), \ldots, (p_{a_k}, q_{b_k}), \ldots, (p_{a_T}, q_{b_T}), \tag{3}$$

with correct starting and ending conditions $\mathrm{a}_1 = \mathrm{b}_1 = 1$, $\mathrm{a}_T = \mathrm{n}$, and $\mathrm{b}_T = \mathrm{m}$.

Notice that here the lengths of **S(P)** and **S(Q),** namely, n and m, may not be the same. However, it is always possible to construct **C(P, Q)** because two successive elements, for instance, p_{a_k} and $p_{a_{k+1}}$, may be the same. More precisely, starting from a point pair (p_{a_k}, q_{b_k}), one point (or both points) should move to its next position (or their next positions) at each step. This means that one of the following three conditions should be satisfied:

$$\begin{aligned} &(I) \quad a_{k+1} = a_k + 1 \quad b_{k+1} = b_k, \\ &(II) \quad a_{k+1} = a_k \quad b_{k+1} = b_k + 1, \\ &(III) \quad a_{k+1} = a_k + 1 \quad b_{k+1} = b_k + 1. \end{aligned} \tag{4}$$

When the coupling **C** is calculated, the corresponding coupling distance is defined as the largest distance between p_{a_k} and q_{b_k},

$$\|\mathbf{D}_{\mathbf{C}}\| \equiv \max_{k=1,2,\ldots,T} dist(p_{a_k}, q_{b_k}). \tag{5}$$

Because the coupling between two given trajectories **P** and **Q** is not uniquely defined, all possible couplings **C** form a space $\boldsymbol{R}_{\mathrm{s}}$(**C**). The discrete Fréchet distance between **P** and **Q** is defined as the minimum coupling distance over all possible couplings in the space $\boldsymbol{R}_{\mathrm{s}}$(**C**), namely,

$$\delta_{dF} \equiv \min\{ \quad \|\mathbf{D}_{\mathbf{c}}\| \mid \quad \mathbf{C} \in \boldsymbol{R}_{\mathrm{s}}(\mathbf{C}) \quad \}. \tag{6}$$

It is not easy to get all possible couplings in a space $\boldsymbol{R}_{\mathbf{s}}$**(C)** by the straightforward way. One possible solution to perform such calculations was proposed by Alt and Godau,[113] while the mathematical implementation is difficult. However, Eiter and Mannila[112] once proved that it is possible to calculate the

discrete Fréchet distance between two trajectories (**P** and **Q**) by the dynamical programing way. This provides the simple way for the computational implementation. This approach was clearly discussed in several previous studies,[108,112,114] even including the pseudocodes. According to this idea, it is possible to compute the discrete Fréchet distance by the dynamical programming algorithm.[108,112,114,128] This allows us to compute the pair-wise dissimilarity matrix over all trajectories, giving the possibility to employ various machine learning algorithms in further analysis. More discussions on Fréchet distance and computational details are found in the Appendix, Subsections 1–3.

4. Multi-dimensional scaling

As a widely used dimensionality reduction method, the classical MDS constructs the low-dimension space, in which the pair-wise dissimilarities between all data points under study are preserved.[80] The MDS algorithm starts from the construction of the pair-wise dissimilarity matrix **D** with the dimension n × n, where n is the number of objects. d_{ij} represents the "distance" between two data points, and then, it is possible to define the scalar product matrix **B** as

$$\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J}, \tag{7}$$

where $\mathbf{D}^{(2)}$ is the squared proximity matrix with elements ${d_{ij}}^2$, namely,

$$\mathbf{D}^{(2)} = \left[{d_{ij}}^2\right], \tag{8}$$

and **J** is the center matrix defined as

$$\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{9}$$

where **I** is a unit matrix. **1** is a column vector with all elements equal to 1, and $\mathbf{1}^{\mathrm{T}}$ is the corresponding row vector. Thus, the product of these two matrices $\mathbf{1}\mathbf{1}^{\mathrm{T}}$ gives a matrix with all elements equal to 1.

Next, we diagonalize the **B** matrix and reorder all eigenvalues from largest to smallest. The larger eigenvalue corresponds to a more important dimension. For instance, if a reduced space with *m*-dimension is considered, we need to take the *m* largest positive eigenvalues $\lambda_1, \ldots, \lambda_m$ and their corresponding eigenvectors $e_1, \ldots, e_m$. The coordinates of all data points in the low-dimensional space are computed by

$$\mathbf{L} = (e_1 \cdots e_m)\begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_m} \end{pmatrix}, \tag{10}$$

The relative embedding error is computed by the stress function; see the MDS textbook.[80]

In current work, two kinds of pair-wise dissimilarity matrices **D** are involved. We first employ the MDS to analyze the trajectory similarity. In this step, the "pair-wise dissimilarity matrix" describes the dissimilarity between different trajectories, which is named as $\mathbf{D}_{\mathrm{traj}}$. The distance between two trajectories is defined as their Fréchet distance. In the second step, we try to select all trajectories belonging to the same cluster, collect geometries from these chosen trajectories (similar trajectories), and then define the pair-wise dissimilarity matrix between all chosen geometries. This dissimilarity matrix is then named as $\mathbf{D}_{\mathrm{geom}}$, in which all elements are the RMSD between two aligned geometries. All details can be found in the below discussions on the implementation details.

5. DBSCAN clustering algorithm

Here we selected the DBSCAN (density-based spatial clustering of applications with noise)[129,130] algorithm to perform the trajectory clustering analysis after the construction of the pair-wise dissimilarity matrix of all trajectories. The DBSCAN is a density-based clustering algorithm. The basic assumption of the DBSCAN method is that all data points form the high-density and low-density areas. Then it is possible to put all points belonging to the same high-density area together to define a cluster, while different clusters are separated by the low-density areas. The data points located in the low-density areas are labeled as outliers. Because a large number of trajectories are computed in the TSH calculation, it is possible to get a few abnormal trajectories, which give a few so-called outer data points that do not belong to any cluster in the trajectory clustering analysis. As a density-based method, the DBSCAN cluster algorithm is robust to outlier points relevant to these abnormal trajectories.

B. Implementation details

1. PΦB model and on-the-fly TSH dynamics simulation

Phytochromobilin (PΦB) acts as the chromophore of plant phytochromes. Two conformers (ZaZsZa and ZsZsZa) are important during the P_r and P_{fr} conversion, as summarized in Ref. 1. The current paper is mainly for the examination of the novel analysis approach of the trajectory surface hopping results. We choose the ZaZsZa configuration (Fig. 1) as the initial isomer because our previous work discussed the dynamical details of this isomer from the traditional analysis way (distribution of the key coordinates, typical trajectory, and so on).[70] Thus, it is more transparent to compare the results obtained from the current analysis tool and the previous calculation. In addition, the resonance Raman (RR) spectroscopy work has demonstrated that this isomer should be an important isomer in phytochromes.[115]

The nonadiabatic photoisomerization dynamics of the PΦB model is investigated by the TSH method at the semiempirical OM2/MRCI level (the orthogonalization model 2 semiempirical method combined with multi-reference configuration interaction algorithm).[131–133] The Wigner sampling of the lowest vibrational level of the electronic ground state is performed. All trajectories start from the S_1 state, and the propagation lasts up to 1 ps. We use the same computational setups, such as the active space, discussed in previous studies.[70,95,96] To analyze the simulation data easily, the TSH calculations are performed by the JADE code[46] by calling the OM2/MRCI[131–133] calculations with MNDO code.[134] In the current work, the trajectory clustering analysis requires a large number of trajectories (see Sec. III).

2. Analysis of excited-state photoisomerization dynamics before S_1-S_0 hops

Normally, the analysis of the multi-channel nonadiabatic dynamics should identify which conical intersection is responsible for the internal conversion and which molecular motion is relevant to the excited-state dynamics. For the current PΦB model, this task becomes the identification of the different isomerization channels via different conical intersections. To address these key questions, the following protocols are employed.

(a) We selected the geometries at every 10 fs for each trajectory before the S_1-S_0 hops. In this sense, the excited-state dynamics before the S_1 decay is fully characterized by these trajectories containing a large number of snapshots.

(b) For two trajectories **P** and **Q**, we computed the distances between any two geometries p_i ($p_i \in \mathbf{P}$) and q_j ($q_j \in \mathbf{Q}$). In this step, the distance between two geometries is defined by their RMSD by neglecting hydrogen atoms. We performed the alignment of each snapshot with respect to the reference geometry (ground-state minimum) to remove the contribution of translational and rotational motions. This alignment approach, instead of the pair-wise alignments for all snapshots, confirms that a correct metric space is formed in the Fréchet distance calculations.[108,128,135]

(c) The dissimilarity of each pair of trajectories is defined by the discrete Fréchet distance. Finally, we got the pairwise distance matrix $\mathbf{D}_{traj}$ of all trajectories, whose dimension is $N_{traj} \times N_{traj}$.

(d) The MDS analysis was performed in the basis of the pairwise distance matrix $\mathbf{D}_{traj}$ of all trajectories. Then in the two-dimensional space, each trajectory is represented by a point and the basic feature of the data distribution is easily examined. When two data points are closer, two corresponding trajectories are more "similar."

(e) The trajectory clustering analysis was performed with the DBSCAN clustering algorithm, which divide all data points into different groups in the two-dimensional space. In the trajectory clustering, the trajectories with high similarity in principle should be assigned into the same group.

(f) In the ideal case, each cluster corresponds to a decay channel in the nonadiabatic photoisomerization dynamics after the trajectory clustering analysis. For this purpose, we performed the additional check. The clustering analysis divided all trajectories into different groups. Next based on the trajectories belonging to the same cluster (for instance, cluster **A**), we took their Fréchet distances to construct the pair-wise distance matrix (labeled as $\mathbf{D}_{traj\_A}$), which is the submatrix of the full pair-wise dissimilarity matrix $\mathbf{D}_{traj}$ of all trajectories. Based on $\mathbf{D}_{traj\_A}$, the MDS dimensionality reduction and the DBSCAN clustering algorithm were performed again, to see whether it is possible to divide cluster **A** into several smaller sub-clusters. Notice that the different reduced spaces were formed at two successive runs because the different distance matrices were employed in the dimensionality reduction before the clustering analysis. This procedure should be repeated until each generated small cluster only gives a single dense dataset. Until now, we wish that each cluster in principle corresponds to a single nonadiabatic decay channel.

(g) The next task is to identify which reactive coordinates are responsible for a single nonadiabatic decay channel. In this step, we simply took the dimensionality reduction analysis discussed in our previous work.[95] The trajectories belonging to the same cluster were collected. All snapshots belonging to the selected trajectories were used to calculate the pair-wise dissimilarity matrix $\mathbf{D}_{geom}$. Then the MDS analysis based on $\mathbf{D}_{geom}$ was performed to construct the low-dimensional space, and each point now refers to a configuration. For the data points located in the same grid area, we overlapped their configurations together and examined the characteristic geometric feature. In this way, it is possible to identify the major reactive coordinate responsible for a particular channel in the nonadiabatic decay dynamics.

3. Analysis of full nonadiabatic photoisomerization dynamics towards different photoproducts

We collected trajectories belonging to the same cluster (for instance, cluster **A**) generated from the analysis of excited-state dynamics before S_1-S_0 hops, after making sure that each cluster should not be divided again. These trajectories in principle pass the same conical intersections, while different products may be formed after internal conversion. Next, we wish to understand their full nonadiabatic dynamics towards photoproducts. We expect that cluster **A** can be divided into several smaller clusters again after the ground-state dynamics is considered.

The geometry re-sampling for these trajectories is performed with a larger time step (40 fs) and a longer time duration (1 ps). The employment of the longer time duration confirms that photoproducts are formed by the successive ground-state dynamics after the internal conversion. The use of the larger time step is mainly for reducing computational cost. We performed the trajectory clustering analysis again by taking the ground-state dynamics into account for the trajectories passing the same conical intersection. Several clusters were formed, and we hope that each cluster includes trajectories with high similarity, namely, passing the same conical intersection and forming the same photoproduct. The analysis of the geometry similarity with the dimensionality reduction approach is again employed for each group of trajectories, to clarify the major molecular motions in a channel.

4. The definition of the "typical trajectory"

As a side product of the trajectory clustering process, it is easy to find the "typical trajectory" for each reaction channel. As discussed in Sec. II B 2, all trajectories belonging to the same non-dividable cluster should be "similar" in trajectory clustering analysis. Thus, if one trajectory shows the highest similarity with rest of the trajectories within a cluster, this one can be assigned as the "typical" trajectory.

Starting from all trajectories belonging to the same cluster, we estimated their similarity via the pair-wise Fréchet distance matrix. Among all trajectories, it is always possible to find a trajectory which gives the minimum value of the sum of the Fréchet distances between this selected trajectory and all other trajectories. In this situation, we can assign this trajectory as the "typical" or "representative" one that characterizes the important geometrical evolution of this group of trajectories.

C. Coding issues

In this work, the dynamics simulation was performed within the developing version of the JADE package,[46,47,136] which contains a module to interface with several quantum chemistry packages (including the interface with the MNDO package[134]). A simple homemade FORTRAN code was developed to calculate the RMSD between two geometries.[95] Most analysis scripts were written with Python language, and the Scikit-learn Python toolkit[137,138] was used for the data analysis, such as the DBSCAN clustering.

III. RESULTS

A. Clustering analysis of trajectory similarity before S_1-S_0 hops

In the analysis of the nonadiabatic dynamics of photoisomerization, an important task is to understand the excited-state dynamics before the S_1-S_0 decay. Thus, we cut the trajectories until their hops, defined the pair-wise distance matrix among all trajectories by invoking the Fréchet distance calculations, used the dimensionality reduction approach by the MDS, and performed the trajectory clustering analysis with DBSCAN methods. Two clusters appear clearly [Fig. 2(a)], which are labeled as cluster **A** and cluster **B**.

Cluster **A** contains 142 trajectories and cluster **B** contains 303 trajectories, while a few trajectories (~2.8%) were ignored according to the noise reduction principle of the DBSCAN algorithm. Although only two clusters exist, it is necessary to check whether each cluster can be divided again. For this purpose, at the second step, we took all trajectories belonging to cluster **A**, defined the pair-wise Fréchet distance matrix, and performed the dimensionality reduction approach and the trajectory clustering analysis again. Figure 2(b) shows that it is not possible to divide cluster **A** into small groups. The same operation on cluster **B** was performed and the results are given in Fig. 2(c). We wish to point out that at each step, different reduced spaces are formed because different distance matrixes were employed in the MDS analysis. When each cluster could not be divided anymore after several iterative steps of trajectory clustering analysis, two clusters are finally obtained. In principle, this indicates that two nonadiabatic decay channels may be involved. The next task is to check the trajectory feature of each group and to understand the dynamical evolution in each channel.

1. Geometrical evolution of trajectories belonging to cluster A

To get the geometrical features of trajectories belonging to cluster **A**, we first collected all trajectories belonging to such clusters. The snapshot was taken before the hopping events, and totally 2827 geometries were collected to form a dataset.

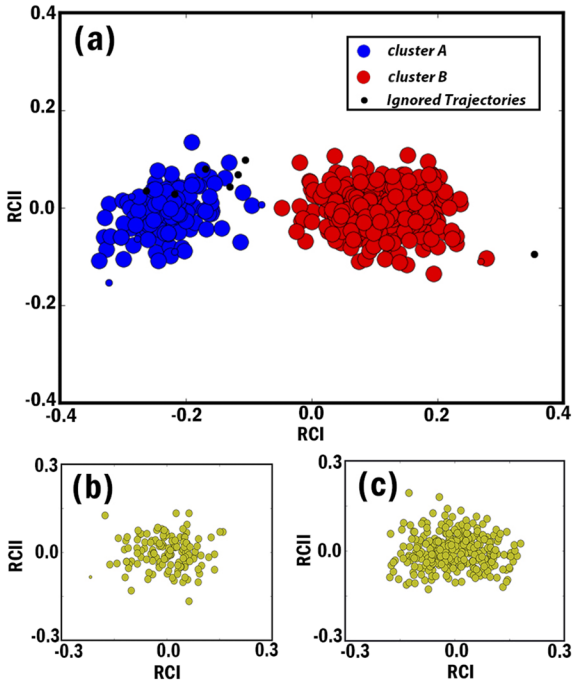

FIG. 2. The clustering analysis of trajectory similarity before S_1-S_0 hops. (a) In the first run, we simply collected all trajectories, defined the pair-wise distance matrix for all trajectories, employed the dimensionality reduction approach, and performed the clustering analysis. Two clusters appear, labeled as cluster **A** and cluster **B**. (b) In the second run, we took all trajectories belonging to cluster **A** and repeated the above analysis as the first step. (c) The similarity analysis was also performed for trajectories belonging to cluster **B**.

This classical MDS analysis of the pair-wise distance matrix among all collected geometries gives a clear distribution pattern in the low-dimensional space spanned by two reduced coordinates, as shown in Fig. 3(a), in which each point represents a geometry snapshot. It is obvious that the snapshots evolve from the Franck-Condon (FC) region to the S_1-S_0 conical intersection region, corresponding to the changing of RCI values from ~−0.1 to ~0.2, as shown in Fig. 3(b). We selected three representative local domains along the RCI axis and stacked all snapshots in each selected domain. It turns out that the RCI was governed by the torsional angle at the C_{10}-C_{11} bond, namely, τ_{10_11}, as shown in Figs. 3(c) and 3(d). Overall, the torsional motion of τ_{10_11} is observed and the hops take place near the S_0-S_1 conical intersection region with τ_{10_11} ~ 70°-90°, as shown in Figs. 3(b) and 3(d).

2. Geometrical evolution of trajectories belonging to cluster B

Similar analysis was also performed for cluster **B**. We totally collected 6379 geometries for MDS analysis. As shown in Fig. 4(a), the dominant reaction coordinates of the trajectories in cluster **B** can also be represented by a one-dimensional reduced coordinate (RCI); see Fig. 4(b). The torsional motion at the C_9-C_{10} bond (τ_{9_10}) is responsible for RCI, as shown in Figs. 4(c) and 4(d). This indicates that the strong torsional

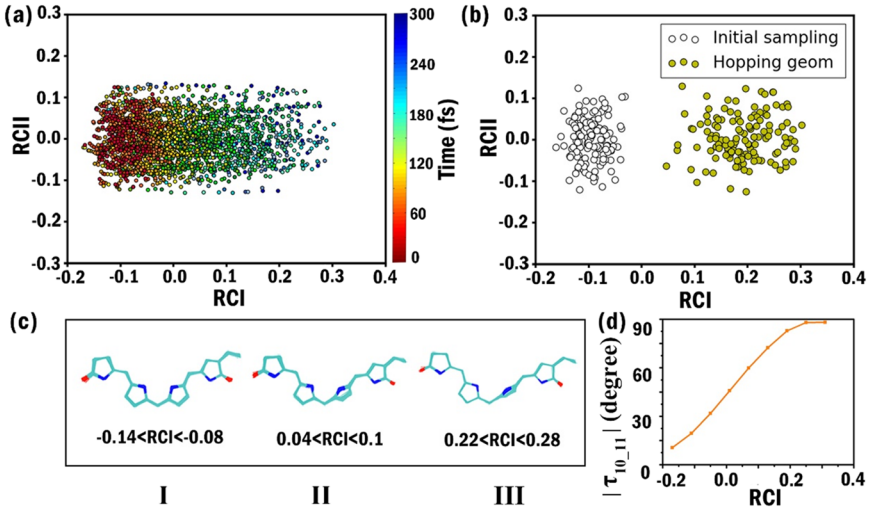

FIG. 3. Classical MDS analysis of the geometrical evolution for the trajectories belonging to cluster **A**. (a) Locations of sampled geometries in the low-dimensional space spanned by two leading reduced coordinates RCI and RCII. Colour codes indicate the time evolution. (b) Locations of the initial geometries and the hopping geometries in the two-dimensional reduced space. (c) Geometrical aggregations in three representative local domains. (d) The values of τ_{10_11} *vs* RCI.

motion τ_{9_10} takes place in the excited-state decay pathway towards conical intersections.

In the above protocol, two clusters are formed in the clustering analysis of the trajectory similarity, and the further MDS analysis of the geometry similarity shows that each cluster corresponds to a single reaction channel. It is obvious that cluster **A** is relevant to the torsional motion along τ_{10_11} while cluster **B** is governed by the torsional motion of τ_{9_10}. The above detailed analysis gives a clear description of the dynamics process from the initial sampling to two S_1-S_0 intersection regions. These observations are consistent with our previous studies.[70]

If we wish to get a full dynamical picture of photoinduced reactions, it is also important to know photoproducts. We repeated the above analyses of trajectory similarity and geometry similarity again, while this time all trajectories stop at 1 ps. For all trajectories belonging to either cluster **A** or **B**, we computed the trajectory similarity and perform the clustering analysis again, while the ground-state dynamics after the internal conversion is also included. This way clearly shows that the trajectories of cluster **A** (or **B**) can be distinguished by their different photoproducts. Because the analysis procedures of cluster A and B are very similar, we tried to mainly focus on the analysis of cluster B containing more trajectories in the below discussion. To avoid redundancy, we gave the main results relevant to cluster **A** in the Appendix, Subsection 4.

B. Clustering analysis of trajectory similarity with the inclusion of photoproducts

After the inclusion of the ground state dynamics, cluster **B** is divided into two sub-clusters as shown in Fig. 5(a), which are labeled as clusters **B1** and **B2**. Cluster **B1** contains 191 trajectories, while cluster **B2** contains 106 trajectories, while a few trajectories are treated as noise in the DBSCAN algorithm. The second round of trajectories' clustering results as shown in Figs. 5(b) and 5(c) proves that these two sub-clusters cannot be divided anymore. The appearance of two clusters, **B1** and **B2**, indicates that two different

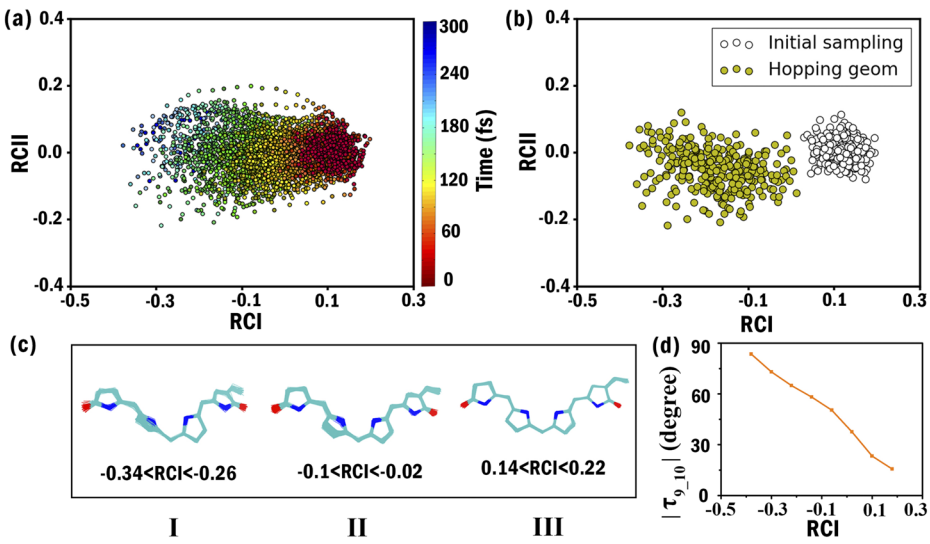

FIG. 4. Classical MDS analysis of the geometrical evolution for the trajectories belonging to cluster **B**. (a) Locations of sampled geometries in the low-dimensional space spanned by two leading reduced coordinates RCI and RCII. Colour codes indicate the time evolution. (b) Locations of the initial geometries and the hopping geometries in the two-dimensional reduced space. (c) Geometrical aggregations in three representative local domains. (d) The values of τ_{9_10} *vs* RCI.

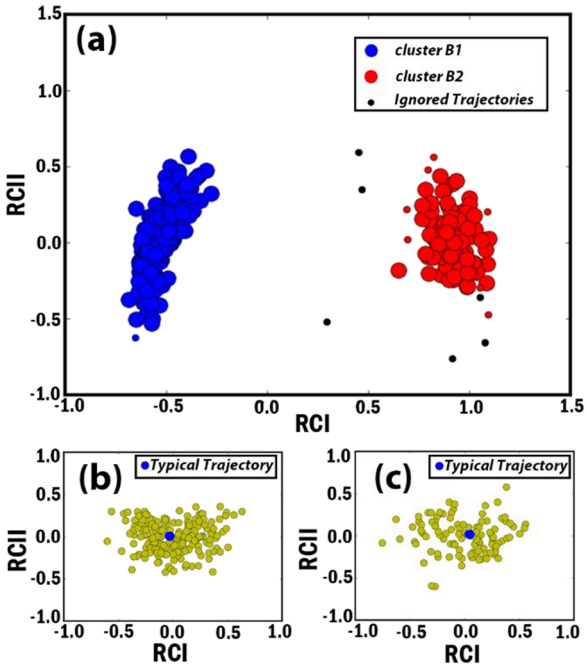

FIG. 5. The further clustering analysis of all trajectories belonging to cluster **B** when the propagation lasts up to 1 ps. (a) In the first run, we collected all trajectories in cluster **B**, defined their pair-wise distance matrix, employed the dimensionality reduction approach, and performed the trajectory clustering analysis. Two clusters appear, labeled as cluster **B1** and cluster **B2**. (b) In the second run, we took all trajectories belonging to cluster **B1** and repeated the above analysis as the first step. (c) The similar analysis was performed for trajectories belonging to cluster **B2**. The blue dots in (b) and (c) represent the typical trajectories of each cluster.

products are formed for the trajectories passing the same conical intersection.

1. Geometrical evolution of trajectories belonging to cluster B1

For all trajectories in cluster **B1**, we checked their geometrical evolutions by the analysis of the geometry similarity. We took a snapshot at every 40 fs for each trajectory belonging to cluster **B1,** and 4775 geometries were collected.

The classical MDS analysis of the geometry similarity over the trajectories belonging to cluster **B1** gives a very interesting distribution pattern in the low-dimensional space shown in Fig. 6(a). Before ~500 fs, the first two leading dimensions control the propagations. After ~500 fs, the third reduced coordinate starts to play an important role. We also show the data distribution and evolution in the two-dimensional reduced space in Fig. 6(b). Starting from the FC region (RCI ~ 0.15 and RCII ~ −0.15), the whole propagation seems to follow a circle. We selected three key blocks (I, II, and III) in the representative regions (FC region, hopping region, and photoproduct region) and aggregate all the snapshots in each block, shown in Fig. 6(c).

From block I to block II, the τ_{9_10} twisting angle clearly experiences rotational motion accompanied by the weak torsional motion along τ_{5_6}. From block II to block III, both τ_{9_10} and τ_{5_6} angles return to the initial values while the "hot" geometries appear due to excessive energies. The large-amplitude vibrational motions not only include the τ_{9_10} and τ_{5_6} torsions but also the vibrations of other coordinates such as the deformation of the side vinyl group. This explains that the third reduced coordinate RCIII is involved after the system goes back to the ground state. Also due to the same reason, the data ensemble does not seem to finally go back to the starting region, while all photoproduct geometries look very similar to the initial reactant ones.

To confirm the above analysis results, we gave the time-dependent distribution diagram of the τ_{9_10}, τ_{5_6}, and τ_{vinyl} group in Figs. 7(a)–7(c), respectively. Before 300 fs, the strong torsional motion of τ_{9_10} is observed, accompanied with the weak change of τ_{5_6}. This observation is consistent with our previous theoretical results.[70] From 300 fs to 500 fs, both torsional angles return back to the initial configurations. After 500 fs, both τ_{9_10} and τ_{5_6} show the rather board distributions, also indicating their large-amplitude vibrational motions. The deformation of the side vinyl group starts to be very important after 400 fs, reflected by the evolution of τ_{vinyl}. Most importantly, such motion is highly excited because the distribution of τ_{vinyl} covers a very board angular range. Overall, we can still assign cluster **B1** to be the channel, in which the system assesses the conical intersection by the strong τ_{9_10} torsional motion and the weak τ_{5_6} torsional motion, and then trajectories go back to the reactant region, if we only consider the backbone motion and neglect the side-chain motion.

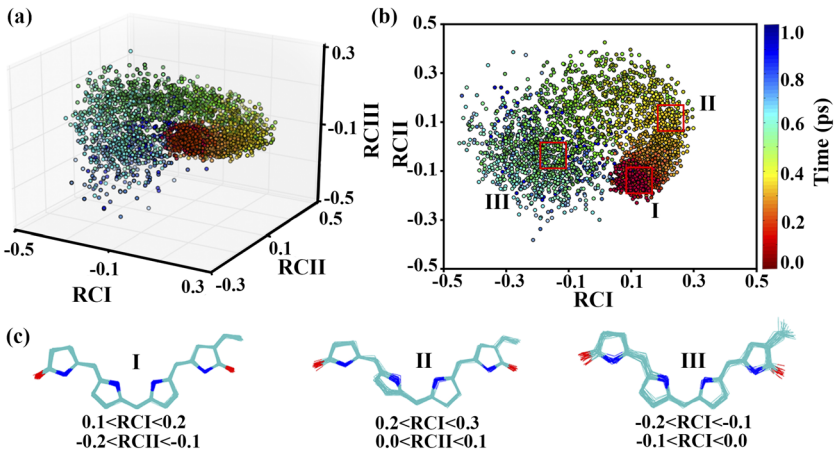

FIG. 6. Classical MDS analysis of the geometrical evolution for the trajectories belonging to cluster **B1**. (a) Locations of sampled geometries in the low-dimensional space spanned by three leading reduced coordinates and colour codes indicate the time. (b) Locations of sampled geometries in the low-dimensional space spanned by two leading reduced coordinates and three representative blocks. (c) Geometrical aggregations in three representative locations.

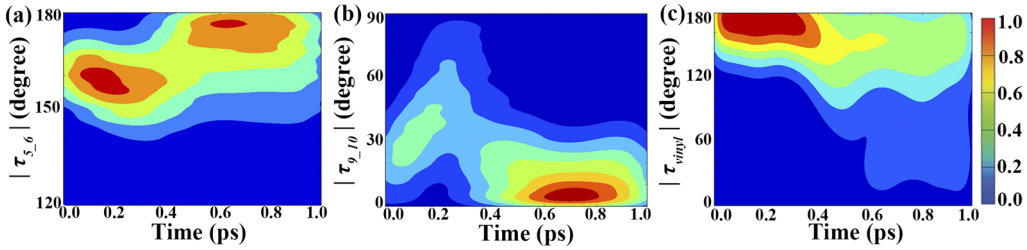

FIG. 7. Time-dependent distribution diagram of three key coordinates for the trajectories belonging to cluster **B1**. In the colour bar, 0 means the minimum value while 1 means the maximum value. (a) Distribution of τ_{5_6} *vs* time. (b) Distribution of τ_{9_10} *vs* time. (c) Distribution of torsion of vinyl group (τ_{vinyl}) *vs* time.

2. Geometrical evolution of trajectories belonging to cluster B2

Cluster **B2** contains 106 trajectories, which are examined by the same analysis protocol (based on 2650 geometries), as used in cluster **B1**.

Compared to cluster **B1**, the geometrical evolution of trajectories belonging to cluster **B2** has a very clear propagation pattern, as seen in Figs. 8(a) and 8(b), both in a three-dimensional space and a two-dimensional space. Before 600 fs, the propagation is dominated by the first two key coordinates RCI and RCII. After that, the third dimension RCIII starts to be involved. In the two-dimensional space spanned by RCI and RCII, as shown in Fig. 8(b), the geometry evolution basically follows a semi-circle, starting from the FC region. Then, we selected four representational blocks (I, II, III, and IV) to examine the features of the geometry evolution in Fig. 8(c). From block I (close to the FC region) to block II, τ_{9_10} increases from ~0° to ~90°, accompanied by the change of τ_{5_6}. From block II to block III, the τ_{9_10} tends to be ~180° while the τ_{5_6} returns to the initial values. In the later stage, τ_{14_15} starts to play roles from block III to block IV. At the same time, we also observe the large distribution of the τ_{vinyl} angle.

We made the time-dependent distribution diagram of the four key torsion angles (τ_{5_6}, τ_{9_10}, τ_{14_15}, and τ_{vinyl}) as shown in Fig. 9. It is almost the same with our discussion on the geometry evolution. The τ_{9_10} angle goes from ~0° to ~90° and then continuously moves to ~180° to give the photoproducts before 500 fs. The τ_{5_6} angle also displays the visible changes and then returns in the dynamics. Please notice that the τ_{14_15} angle may also show some torsional motion. However, such motion starts to take place only on the ground-state dynamics, even after the final products are almost formed, and the τ_{14_15} angle quickly goes back to the initial configuration, as shown in Figs. 9(b) and 9(c). Thus, it is safe to believe that this angle is not relevant to the current analysis and no other isomer is formed by such motion. Similar to the cases in cluster **B1**, we observe the large amplitude motion of the side vinyl group in Fig. 9(d). Overall, cluster **B2** corresponds to the channel in which the system assesses the conical intersection by the τ_{9_10} torsional motion and the weak τ_{5_6} torsional motion, and then trajectories move towards the photoproducts with τ_{9_10} ~ 160°–180°.

3. Typical trajectory

In the above analysis, we clearly demonstrated that it is possible to divide all trajectories into different clusters, while each cluster represents a reaction channel. In this situation, we can define the "typical trajectory" in each cluster. For cluster **B1** and cluster **B2**, their typical trajectories are given in Figs. 10(a1)–10(c1) and 10(a2)–10(c2), respectively. For illustration, we show a few important key coordinates *vs* time and give the evolution of other coordinates in the Appendix, Subsection 5. As shown in Figs. 10(a1)–10(c1), in the typical trajectory that represents the evolution of cluster **B1**, τ_{9_10} increases from ~0° to ~100° and then returns to ~0° during the dynamics. At 487 fs, the hop takes place with τ_{9_10} ~ 90° in the vicinity of the conical intersection seam. It is also clear

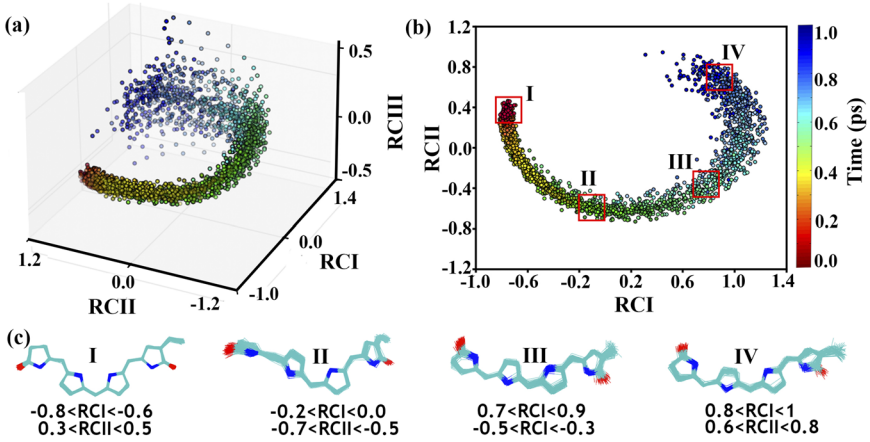

FIG. 8. Classical MDS analysis of the geometrical evolution for the trajectories belonging to cluster **B2**. (a) Locations of sampled geometries in the low-dimensional space spanned by three leading reduced coordinates and colour codes indicate the time. (b) Locations of sampled geometries in the low-dimensional space spanned by two leading reduced coordinates and four representative blocks. (c) Geometrical aggregations in four representative locations.

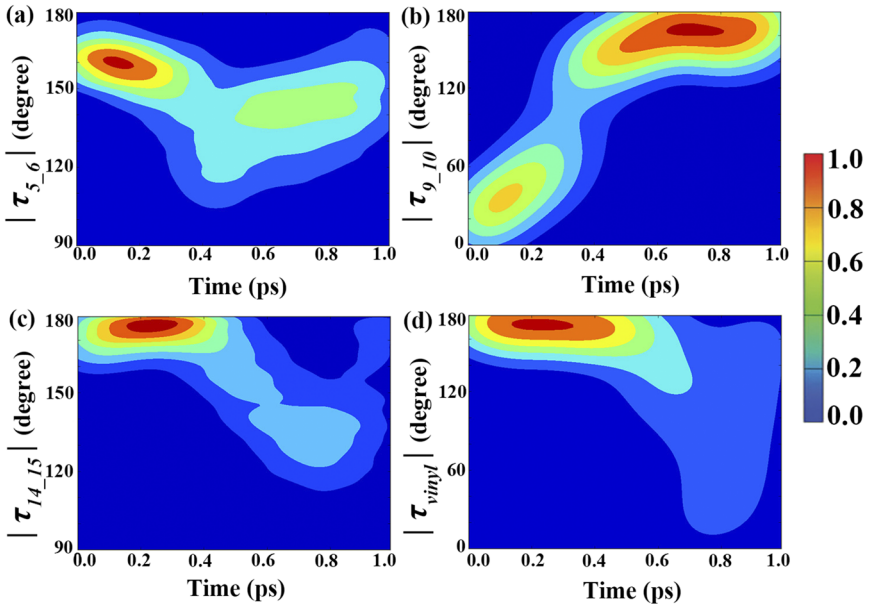

FIG. 9. Time-dependent distribution diagram of four key coordinates for the trajectories belonging to cluster **B**2. In the colour bar, 0 means the minimum value while 1 means the maximum value. (a) Distribution of τ_{5_6} *vs* time. (b) Distribution of τ_{9_10} *vs* time. (c) Distribution of τ_{14_15} *vs* time. (d) Distribution of τ_{vinyl} *vs* time.

that the C_9-C_{10} distance becomes longer in the early state of dynamics. All these features, including the time scale and the geometry evolution, are consistent with the above discussions. This strongly implies that a reasonable "representative" trajectory is selected. The same way can also be applied to select the typical trajectory for cluster **B2**; see Figs. 10(a2)–10(c2).

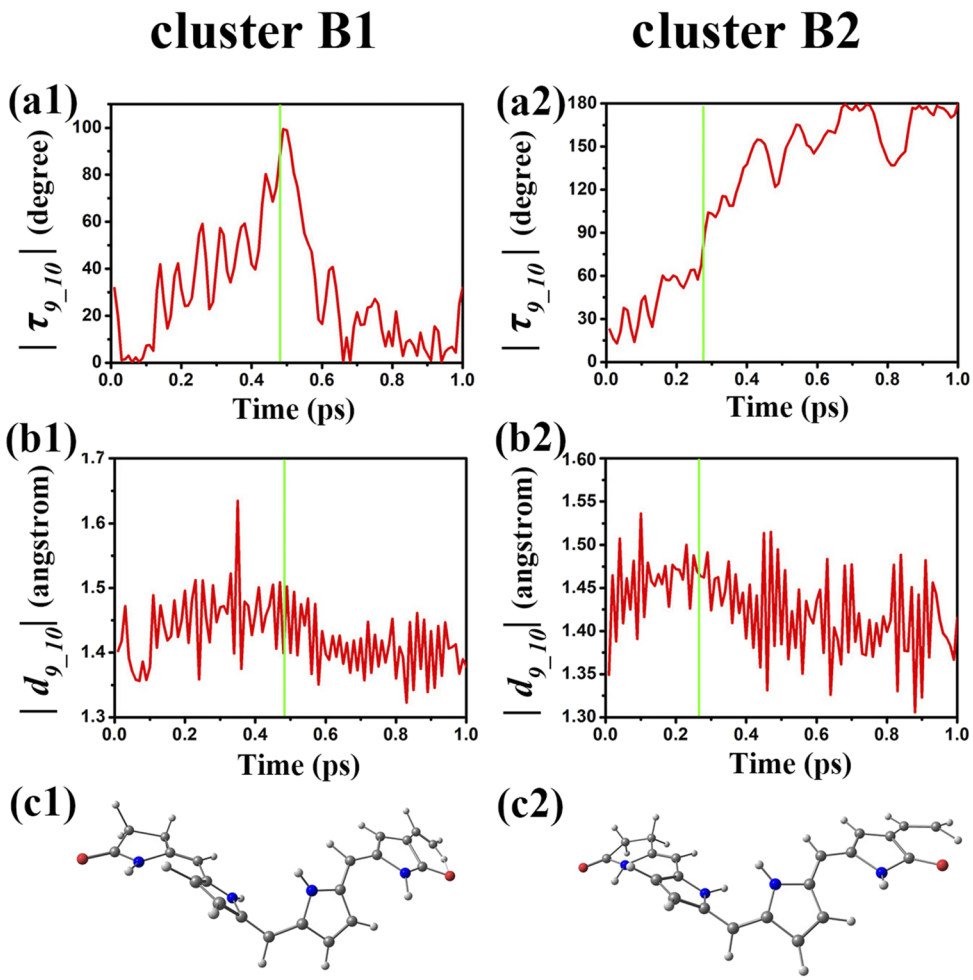

FIG. 10. Geometrical evolution in typical trajectories for clusters **B1** and **B2**. The time-dependent torsional angle τ_{9_10}, bond distance d_{9_10}, and hopping geometries of the typical trajectory in the **B1** cluster are given in (a1), (b1), and (c1), respectively. The corresponding results in the **B2** cluster are shown in (a2), (b2), and (c2), respectively. The vertical green lines represent the hopping events.

C. Photoisomerization mechanism, reaction channels, and branching ratio

Up to now, we employed the clustering analysis on the trajectory similarity to distinguish different channels in the nonadiabatic dynamics of the PΦB multi-channel photoisomerization. The configurational similarity analysis further gives us the geometrical evolution feature of each channel. Most interestingly, the trajectory clustering analysis automatically provides a way to define the so-called typical trajectory.

Starting from the FC region, all trajectories are first grouped into two clusters, **A** and **B**, with the branching ratio around 0.31:0.66 (**A**:**B**). Then each cluster is divided again according to their final products. At the end, four clusters are given, which are **A1**, **A2**, **B1,** and **B2** with branching ratio 0.16:0.15:0.42:0.23. The sum of the total probability is not an exact 1 because some trajectories are neglected in clustering analysis. Clearly each cluster corresponds to a different reaction channel.

For cluster **A1**, the system tends to follow the torsional motion along τ_{10_11}, performs the S_1-S_0 hops with $\tau_{10\_11} \sim 70°$-$90°$ near conical intersection, and then returns to reactants. Although some vibrational motions, such as the geometrical deformation of the side vinyl group, are excited, we still can attribute that this channel finally gives the reactants by checking the backbone motion. For cluster **A2**, the torsional motion of τ_{10_11} is also responsible for the excited-state dynamics towards the conical intersections. After internal conversion, the trajectories tend to move forwards and to give the photoproducts with $\tau_{10\_11} \sim 160°$-$180°$.

For cluster **B**, the system moves towards the conical intersection along the τ_{9_10} torsional motion. After hopping back to the ground state, the trajectories belonging to cluster **B1** return to the reactants, while trajectories belonging to cluster **B2** continuously move to the photoproducts.

After all trajectories are clearly assigned into different clusters, we plot the important geometrical features in a few key events in the trajectory evolution associating with each cluster. For example, for each cluster, **A1**, **A2**, **B1**, and **B2**, we show their hopping geometries and final products in Fig. 11. Each cluster defines a reaction channel. This means that the current analysis can successively distinguish different reaction channels. Overall, all current results on the PΦB photoisomerization are highly consistent with our previous studies.[70]

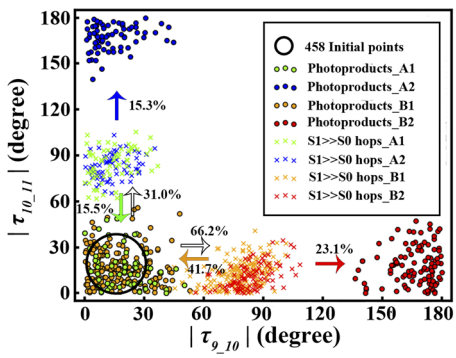

FIG. 11. The branching ratios towards different channels in the TSH simulation of the PΦB's photoisomerization.

IV. DISCUSSIONS

Here we emphasize again why we develop the novel analysis tool of the on-the-fly trajectory surface hopping dynamics. In the straightforward way to examine the on-the-fly TSH nonadiabatic dynamics, the evolution of each trajectory is examined one after another by eye view. It is also necessary to plot the hopping geometries, the final products, and the time-dependent evolution of relevant internal coordinates to perform a meaningful analysis. Some preliminary knowledge on the possible reaction pathways and active coordinates is also necessary for the analysis task. Thus, when a large number of trajectories are employed, the analysis task becomes tedious even under the help of computational scripts. In addition, it is not enough to employ the so-called "typical trajectory" to perform the analysis because it is not trivial to define the typical trajectory over a huge number of trajectories. When the system becomes complex, it is also not so easy to obtain the active molecular motion responsible for the nonadiabatic dynamics. Thus, some novel analysis tools should be developed, which allows us to understand the trajectory surface hopping results based on a large number of trajectories. Recently, Tully also pointed out a similar idea.[139] The current analysis approach automatically finds the reaction channels and branching ratio by the trajectory clustering analysis based on trajectory similarity. Then for each cluster, it is rather easy to extract the major geometry evolution responsible for the corresponding reaction channel.

It is also possible to perform the geometry similarity analysis in the configurational space directly, as shown in our previous studies.[95–100] However, the current analysis based on both the trajectory similarity and the geometry similarity is somehow more powerful due to several reasons. First the dimensionality reduction approaches purely based on geometry similarity basically give a few of leading coordinates. This may not work well in the multi-channel situations. In the current approach, each cluster corresponds to a single reaction channel; thus, all trajectories belonging to such a cluster experiences a similar molecular motion. In this case, it is easier to get meaningful results because it is possible to perform the dimensionality reduction analysis for each single channel. This explains why sometimes a single reduced coordinate (even derived from the linear dimensionality reduction algorithm) may be good enough for the analysis of the geometry evolution. Second when a large number of trajectories are involved, the dimensionality reduction purely based on the geometry similarity requires the linear algebra operations on the extremely huge pair-wise dissimilarity matrix formed by a large number of geometries. This task may become very challenging because the calculation may require an extremely huge amount of computer memory to store and treat the very huge matrices. The current approach, on the other hand, requires smaller computer memory in the estimation of the trajectory similarity, although the total computational time should be slightly longer. When each cluster is identified, we only need to perform the dimensionality reduction analysis based on all trajectories

belonging to the single cluster. Because the much smaller pair-wise dissimilarity matrix is considered in the dimensionality reduction approaches, the memory issue is largely alleviated. Third, we can also easily find the so-called "representative" trajectory for each channel from the current analysis. Overall, the current proposed analysis protocol is more powerful to analyze the multi-channel nonadiabatic dynamics.

In this work, we performed the trajectory clustering analysis in the two-step manner, namely, first checking the responsible conical intersection and second examining the final products. In principle, it is always recommended to examine the excited-state dynamics before hops because it is very important to understand the reaction channels of the excited-state dynamics and relevant conical intersections in the analysis of multi-channel nonadiabatic dynamics. In some cases, after the internal conversion, the system may become highly vibrationally excited and the "hot" ground-state dynamics may not be very relevant to the nonadiabatic dynamics. In this case, only the first step in the current analysis protocol is necessary. Although it is possible to plot the hopping geometries in the examination of the reaction channels, the current analysis displays many advantages, for instance, taking the time evolution into account directly and giving us the representative trajectory for each channel. In addition, the excited-state motion is normally driven by a few of reactive coordinates in a single channel in the ultrafast nonadiabatic dynamics and the Fréchet distance may well capture the main geometrical evolution. By contrast, sometimes the hot motion on the ground state may create many highly distorted snapshots even if the ground-state dynamics may follow some common pathways. In this case, the distance between two trajectories may be determined by these highly distorted geometries, instead of their different reaction channels via different conical intersections. Thus, it is more transparent to first check the relevant conical intersections and then the final products in more general cases.

V. CONCLUSION

We propose a powerful approach to analyze the TSH simulation results of the multi-channel nonadiabatic photoisomerization dynamics by considering both the trajectory similarity and the geometry similarity. In this approach, the clustering analysis of the trajectory similarity is first employed to find how many reaction channels are involved, while the active reaction coordinates responsible for each channel are then identified by the geometry similarity analysis in the configuration space without the requirement of the pre-known knowledge.

In practice, the analysis protocol starts from many trajectories obtained from TSH simulation. The trajectory similarity is estimated by their Fréchet distance. After the pair-wise Fréchet distance matrix was built for all trajectories, the DBSCAN clustering analysis is performed to assign trajectories into different groups. When each group cannot be divided any more, all trajectories belonging to the single non-separable cluster in principle are governed by the same individual reaction channel. To identify the major geometrical evolution feature in each reaction channel, we collect the geometries from the trajectories belonging to the same cluster and compute their pair-wise dissimilarity matrix. Then the MDS dimensionality reduction approach is performed to extract the major coordinates responsible for each channel. As a side product, it is very easy to find the so-called "representative" trajectory from this analysis protocol.

The multi-channel PΦB photoisomerization dynamics is used to explain this novel approach in this work. We first consider the excited-state dynamics and set the cutoff of trajectory propagation at hops. The clustering analysis of the trajectory similarity shows that two clusters are formed, which correspond to two decay channels via their individual conical intersections. In the second step, we start from each cluster, take the photoreaction products into account, and perform the same analysis again. At this step, we notice that the single cluster, obtained at the first step, can be divided into two clusters again. This means that after passing the same conical intersection, the trajectories may go forwards to form the photoproduct or return to the reactant. Totally, four groups of trajectories can be clearly identified and each of them corresponds to a reaction channel. For all four reaction channels, it is possible to extract the active torsional motion and find the typical trajectory. All these results are consistent with our previous studies.[70]

This work demonstrates that the current analysis protocol can extract the main features of multi-channel nonadiabatic photoisomerization dynamics, such as the reaction channels, the branching ratio, and relevant molecular motions, in a more automatic and intelligent way. This analysis approach should be very powerful, which can also be employed in other trajectory-based dynamics approaches.[11,17,23,33] The current work only focuses on the photoisomerization, while, in principle, the same approach can also be employed to treat more general types of nonadiabatic dynamics.[33,41,68] In more realistic cases, this analysis task may face additional problems, such as the trajectory clustering analysis may not give the clearly distinguishable cluster structure or the estimation of geometry similarity may require more advanced molecular descriptors.[140–142] In addition, we also fully know that the distance between two trajectories is only a rather approximated approach. It is also necessary to check different ways to define such similarity by using other distance measurement. All these tasks represent interesting challenging topics in future.

ACKNOWLEDGMENTS

This work was supported by the NSFC Project (Nos. 21873112, 21673266, and 21503248). The authors thank the Supercomputing Centre, Computer Network Information Center, CAS, National Supercomputing Center in Shenzhen, National Supercomputing Center in Guangzhou, and Super Computational Centre of CAS-QIBEBT for providing computational resources.

The authors declare no competing financial interest.

APPENDIX: ADDITIONAL DISCUSSIONS ON METHOD DETAILS AND SUPPORTING FIGURES

1. Geometry alignment

The geometry alignment is important to form the correct metricity in the measurement of the Fréchet distance during

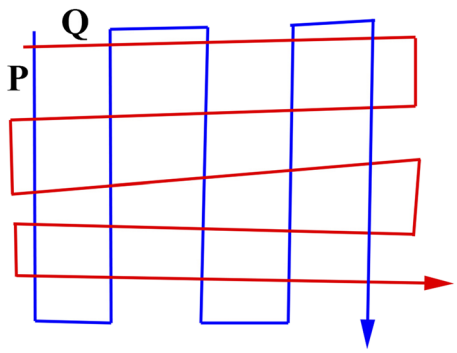

FIG. 12. Situation in which the distance between two trajectories (**P** and **Q**) is described correctly by the Fréchet distance, not by the Hausdorff distance.

the trajectory clustering procedure. Let us assume that we have three points, A, B and C, and their distances are AB, BC, and CA. The correct metricity requires

$$\mathrm{AB} \geqslant 0, \tag{A1}$$

$$\mathrm{AB} = 0 \Leftrightarrow \mathrm{A} = \mathrm{B}, \tag{A2}$$

$$\mathrm{AB} = \mathrm{BA}, \tag{A3}$$

$$\mathrm{AB} + \mathrm{BC} \geqslant \mathrm{AC}. \tag{A4}$$

Previous discussions figured out that it is necessary to select a single suitable reference structure to align all geometries,[135] instead of performing the pair-wise alignment way. In this work, we chose the ground state minimum as reference to perform the alignment of all snapshots universally.

The key point is how to define the distance between two snapshots. Normally it is possible to define their distance by using their RMSD in the Cartesian coordinate, while the contribution of translational and rotational motions must be removed in the alignment procedure. For this purpose, we moved the selected geometry in the translational and rotational ways, and the relevant matrix transformation is determined by the minimization of the RMSD between two geometries. The treatment of the translation motion is rather trivial, while the removing of the rotational motion follows the algorithm proposed in the previous studies.[125,126]

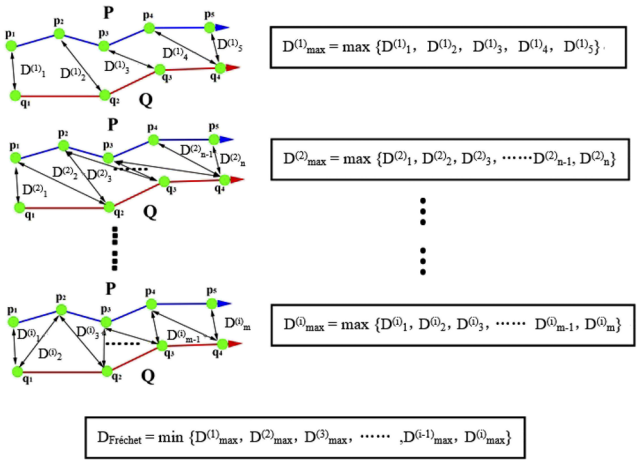

FIG. 13. The definition of the Fréchet distance in a sample model.

```
def _dis(dis,i,j,dis_matrix):
    if dis[i,j] > -1:
        return dis[i,j]
    elif i == 0 and j == 0:
        dis[i,j] = dis_matrix[0,0]
    elif i > 0 and j == 0:
        dis[i,j] = max(_dis(dis,i-1,0,dis_matrix),dis_matrix[i,0])
    elif i == 0 and j > 0:
        dis[i,j] = max(_dis(dis,0,j-1,dis_matrix),dis_matrix[0,j])
    elif i > 0 and j > 0:
        dis[i,j] = max(min(_dis(dis,i-1,j,dis_matrix),_dis(dis,  &
         i-1,j-1,dis_matrix),_dis(dis,i,j-1,dis_matrix)),dis_matrix[i,j])
    else:
        dis[i,j] = float("inf")
    return dis[i,j]

def frechetDist(dis_matrix):
    len_p = np.shape(dis_matrix)[0]
    len_q = np.shape(dis_matrix)[1]
    dis = np.ones((len_p,len_q))
    dis = np.multiply(dis,-1)
    return _dis(dis,len_p-1,len_q-1,dis_matrix)
```

FIG. 14. The pseudocode for the Fréchet distance calculation, where dis_matrix denotes the $\mathbf{D}_{snap}$ matrix.

2. Trajectory similarity

There are several ways to define "distance" between different trajectories. For example, it is possible to define the Euclidian distance between two trajectories, while such distance requires that two trajectories have the same number of the points. In addition, only when two trajectories pass the similar points at the same time during the whole trajectory propagation, they are similar in the Euclidian distance measurement. Thus, this is not suitable for the current analysis purpose.

Hausdorff distance and Fréchet distance are two popular choices to define the "similarity" of two trajectories. It is well known that the Hausdorff distance would face problems in the

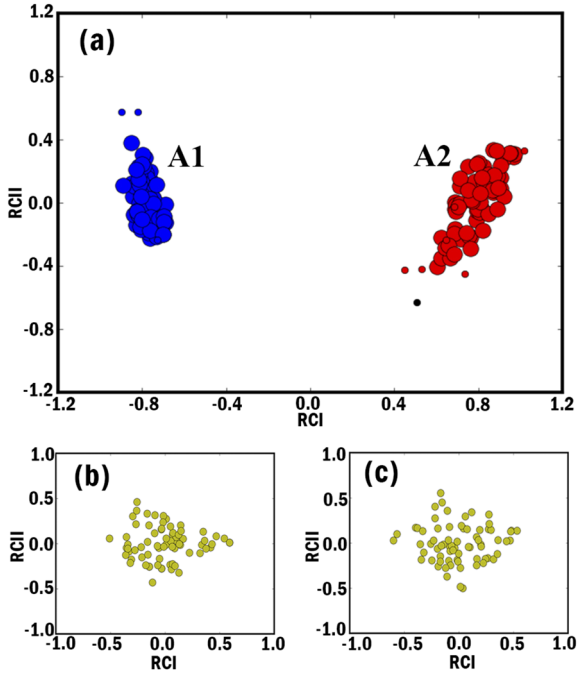

FIG. 15. The further clustering analysis of all trajectories belonging to cluster **A** when the propagation lasts up to 1 ps. (a) In the first run, we collected all trajectories in cluster **A**, defined their pair-wise distance matrix, employed the dimensionality reduction approach, and performed the trajectory clustering analysis. Two clusters appear, labeled as cluster **A1** and cluster **A2**. (b) In the second run, we took all trajectories belonging to cluster **A1**, defined the pair-wise distance matrix, employed the dimensionality reduction approach, and performed the trajectory clustering analysis. (c) A similar analysis was performed for trajectories belonging to cluster **A2**.

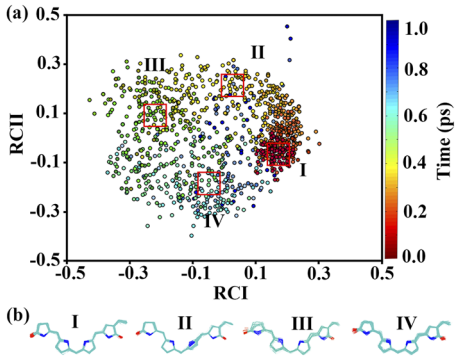

FIG. 16. Classical MDS analysis of the nonadiabatic dynamics results of the trajectories belonging to cluster **A1**. (a) Locations of sampled geometries in the low-dimensional space spanned by two leading reduced coordinates and four representative blocks. (b) Geometrical aggregations in four representative locations.

situation shown Fig. 12. The evolution of two trajectories (**P** and **Q**) is quite different, while their Haussdoff distance is very small. However, the Fréchet distance can distinguish these two trajectories which stay in the similar coordinate space region while their evolutions are different. So, in the current work, we choose the Fréchet distance to define the "similarity" of two trajectories.

3. The definition of the discrete Fréchet distance

We take the below simple model to explain the definition of the Fréchet distance. As shown in Fig. 13, we have two trajectories **P** and **Q**, which are represented by discrete points $\{P_i\}$ and $\{Q_j\}$. Then it is possible to use a different way to link the geometry points in each trajectory; see Fig. 13.

For each connectivity way (labeled as $D^{(k)}$), it is possible to find the largest distance (labeled as $D^{(k)}_{max}$) between all connections. Then the smallest value of all $D^{(k)}_{max}$ (for all k) finally gives the Fréchet distance.

Previous work demonstrated that it is possible to use the dynamical programming algorithm to calculate the Fréchet distance; see Refs. 108, 112, 114, and 143. Imagine we have m snapshots in the **P** trajectory and n snapshots in the **Q** trajectory, and it is easy to get an m-times-n pairwise distance matrix $\mathbf{D}_{snap}$ between all geometries. Then the Fréchet distance can be calculated by the pseudocode; see Fig. 14.

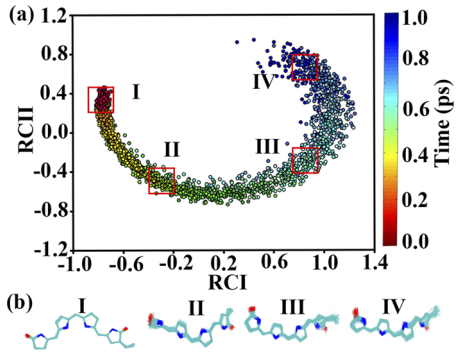

FIG. 17. Classical MDS analysis of the nonadiabatic dynamics results of the trajectories belonging to cluster **A2**. (a) Locations of sampled geometries in the low-dimensional space spanned by two leading reduced coordinates and four representative blocks. (b) Geometrical aggregations in four representative locations.

4. Trajectory similarity in cluster A

We checked the trajectory similarity in cluster A, and the results are shown in Fig. 15. Next, we collect trajectories belonging to the A1 and A2 clusters individually, examine the geometry evolution, and give the results in Figs. 16 and 17.

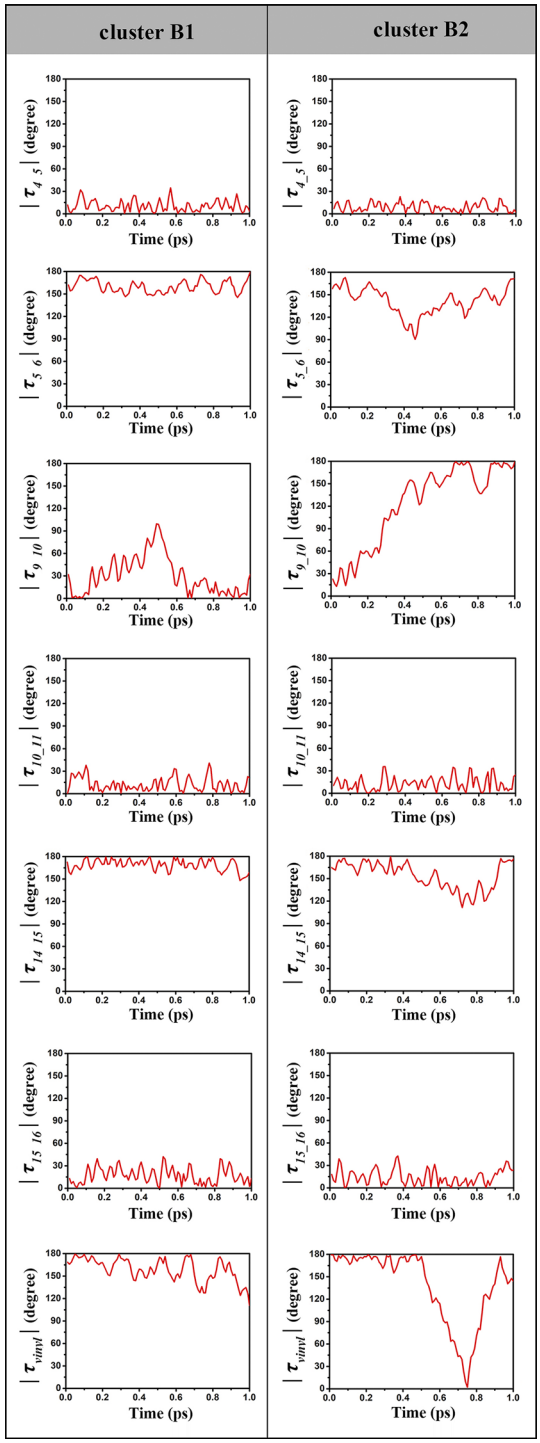

FIG. 18. The propagation of seven key torsion angles *vs* time in the typical trajectories in clusters **B1** and **B2**. Please notice that τ_{14_15} may also show some torsional motion. However, such motion starts to take place only on the ground-state dynamics, even after the final products are almost formed. In addition, the angle τ_{14_15} quickly goes back to the initial configuration. Thus, it is safe to believe that this angle is not relevant to the current analysis and no other isomer is formed by such motion.

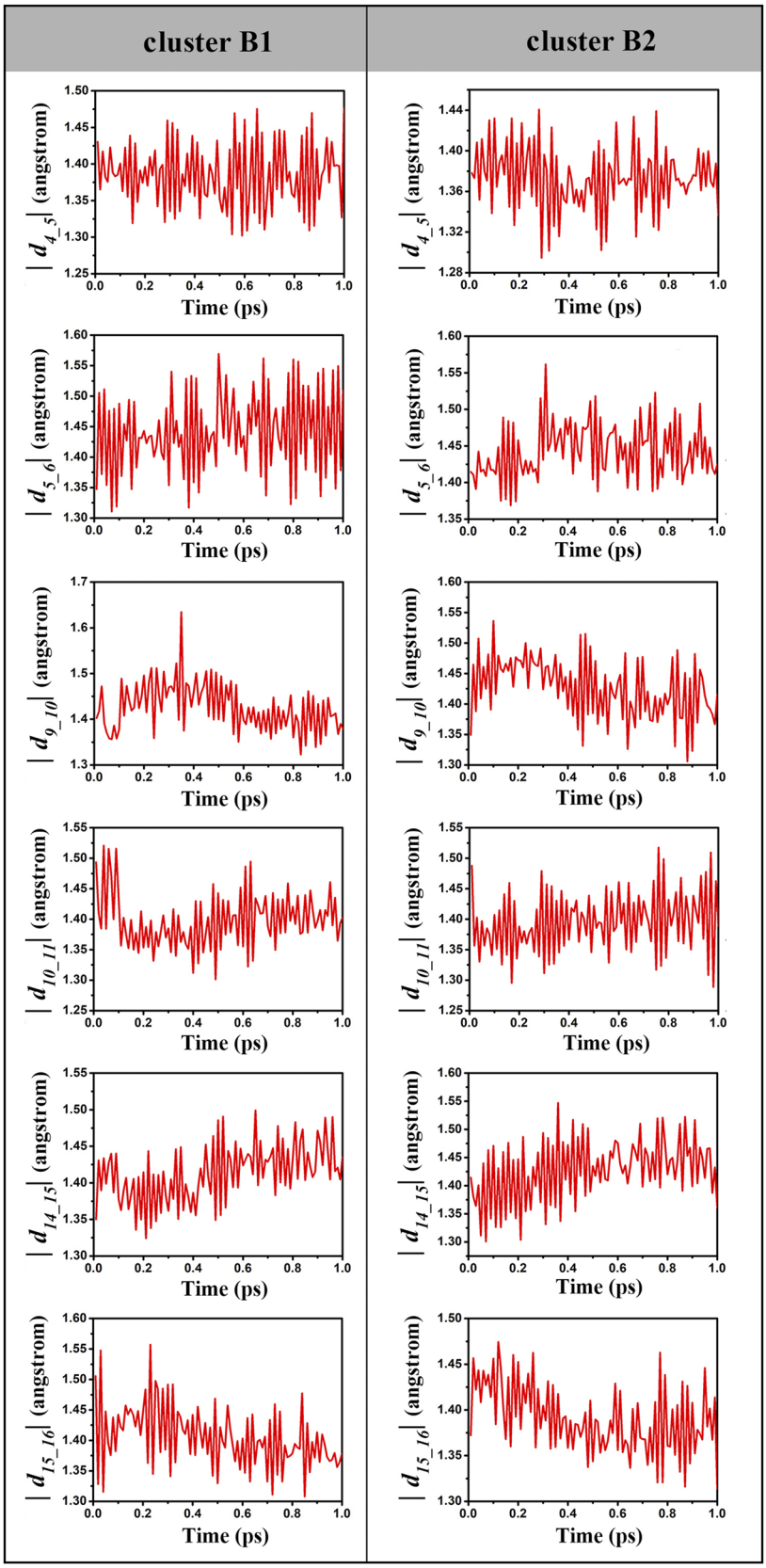

FIG. 19. The propagation of six key bond distances *vs* time in the typical trajectories in clusters **B1** and **B2**.

5. Geometrical evolution of typical trajectory

For clusters B1 and B2, we locate the typical trajectories. The evolutions of the key internal coordinates are given in Figs. 18 and 19.

[1]S. Gozem, H. L. Luk, I. Schapiro, and M. Olivucci, Chem. Rev. **117**(22), 13502–13565 (2017).
[2]B. G. Levine and T. J. Martínez, Annu. Rev. Phys. Chem. **58**, 613–634 (2007).
[3]M. Ben-Nun and T. J. Martínez, Adv. Chem. Phys. **121**, 439–512 (2002).
[4]T. J. Martínez, Acc. Chem. Res. **39**(2), 119–126 (2006).

[5]N. C. Rockwell, Y. S. Su, and J. C. Lagarias, Annu. Rev. Plant Biol. **57**, 837–858 (2006).
[6]M. A. Mroginski, D. H. Murgida, and P. Hildebrandt, Acc. Chem. Res. **40**(4), 258–266 (2007).
[7]D. Polli, P. Altoè, O. Weingart, K. M. Spillane, C. Manzoni, D. Brida, G. Tomasello, G. Orlandi, P. Kukura, R. A. Mathies, M. Garavelli, and G. Cerullo, Nature **467**(7314), 440–446 (2010).
[8]P. B. Coto, A. Strambi, N. Ferré, and M. Olivucci, Proc. Natl. Acad. Sci. U. S. A. **103**(46), 17154–17159 (2006).
[9]W. Domcke, D. R. Yarkony, and H. Köppel, *Conical Intersections: Electronic Structure, Dynamics & Spectroscopy* (World Scientific, Singapore, 2004).
[10]W. Domcke, D. R. Yarkony, and H. Köppel, *Conical Intersections II: Theory, Computation and Experiment* (World Scientific, Singapore, 2011).
[11]T. Yonehara, K. Hanasaki, and K. Takatsuka, Chem. Rev. **112**(1), 499–542 (2012).
[12]M. F. Herman, J. Phys. Chem. A **109**(41), 9196–9205 (2005).
[13]G. W. Richings and G. A. Worth, J. Phys. Chem. A **119**(50), 12457–12470 (2015).
[14]G. H. Tao, J. Phys. Chem. Lett. **7**(21), 4335–4339 (2016).
[15]H. D. Meyer and W. H. Miller, J. Chem. Phys. **70**(7), 3214–3223 (1979).
[16]G. Stock and M. Thoss, Phys. Rev. Lett. **78**(4), 578–581 (1997).
[17]S. J. Cotton and W. H. Miller, J. Phys. Chem. A **117**(32), 7190–7194 (2013).
[18]I. Horenko, C. Salzmann, B. Schmidt, and C. Schütte, J. Chem. Phys. **117**(24), 11075–11088 (2002).
[19]Q. Shi and E. Geva, J. Chem. Phys. **121**(8), 3393–3404 (2004).
[20]C. C. Martens and J. Y. Fang, J. Chem. Phys. **106**(12), 4918–4930 (1997).
[21]K. Ando and M. Santer, J. Chem. Phys. **118**(23), 10399–10406 (2003).
[22]R. Kapral and G. Ciccotti, J. Chem. Phys. **110**(18), 8919–8929 (1999).
[23]K. Saita and D. V. Shalashilin, J. Chem. Phys. **137**(22), 22A506 (2012).
[24]X. S. Li, J. C. Tully, H. B. Schlegel, and M. J. Frisch, J. Chem. Phys. **123**(8), 084106 (2005).
[25]S. C. Cheng, C. Y. Zhu, K. K. Liang, S. H. Lin, and D. G. Truhlar, J. Chem. Phys. **129**(2), 024112 (2008).
[26]S. Meng and E. Kaxiras, J. Chem. Phys. **129**(5), 054110 (2008).
[27]M. J. Bedard-Hearn, R. E. Larsen, and B. J. Schwartz, J. Chem. Phys. **123**(23), 234106 (2005).
[28]C. Y. Zhu, A. W. Jasper, and D. G. Truhlar, J. Chem. Theory Comput. **1**(4), 527–540 (2005).
[29]M. Schroter, S. D. Ivanov, J. Schulze, S. P. Polyutov, Y. Yan, T. Pullerits, and O. Kühn, Phys. Rep. **567**, 1–78 (2015).
[30]H. B. Wang and M. Thoss, J. Chem. Phys. **119**(3), 1289–1299 (2003).
[31]H. D. Meyer, U. Manthe, and L. S. Cederbaum, Chem. Phys. Lett. **165**(1), 73–78 (1990).
[32]J. C. Tully, J. Chem. Phys. **93**(2), 1061–1071 (1990).
[33]B. F. E. Curchod and T. J. Martínez, Chem. Rev. **118**(7), 3305–3336 (2018).
[34]A. V. Akimov, A. J. Neukirch, and O. V. Prezhdo, Chem. Rev. **113**(6), 4496–4565 (2013).
[35]L. J. Wang, A. Akimov, and O. V. Prezhdo, J. Phys. Chem. Lett. **7**(11), 2100–2112 (2016).
[36]S. K. Min, F. Agostini, I. Tavernelli, and E. K. U. Gross, J. Phys. Chem. Lett. **8**(13), 3048–3055 (2017).
[37]J. E. Subotnik, J. Chem. Phys. **132**(13), 134112 (2010).
[38]Y. Yao, K. W. Sun, Z. Luo, and H. B. Ma, J. Phys. Chem. Lett. **9**(2), 413–419 (2018).
[39]V. N. Gorshkov, S. Tretiak, and D. Mozyrsky, Nat. Commun. **4**, 2144 (2013).
[40]F. Webster, E. T. Wang, P. J. Rossky, and R. A. Friesner, J. Chem. Phys. **100**(7), 4835–4847 (1994).
[41]R. Crespo-Otero and M. Barbatti, Chem. Rev. **118**(15), 7026–7068 (2018).
[42]M. Barbatti, G. Granucci, M. Persico, M. Ruckenbauer, M. Vazdar, M. Eckert-Maksić, and H. Lischka, J. Photochem. Photobiol., A **190**(2-3), 228–240 (2007).
[43]E. Fabiano, T. W. Keal, and W. Thiel, Chem. Phys. **349**(1-3), 334–347 (2008).
[44]M. Richter, P. Marquetand, J. González-Vázquez, I. Sola, and L. González, J. Chem. Theory Comput. **7**(5), 1253–1258 (2011).
[45]N. L. Doltsinis and D. Marx, Phys. Rev. Lett. **88**(16), 166402 (2002).
[46]L. K. Du and Z. G. Lan, J. Chem. Theory Comput. **11**(4), 1360–1374 (2015).
[47]L. K. Du and Z. G. Lan, J. Chem. Theory Comput. **11**(9), 4522–4523 (2015).
[48]C. Y. Zhu, K. Nobusada, and H. Nakamura, J. Chem. Phys. **115**(7), 3031–3044 (2001).
[49]A. K. Belyaev, C. Lasser, and G. Trigila, J. Chem. Phys. **140**(22), 224108 (2014).
[50]N. Shenvi, J. E. Subotnik, and W. T. Yang, J. Chem. Phys. **134**(14), 144102 (2011).
[51]E. Tapavicza, I. Tavernelli, and U. Rothlisberger, Phys. Rev. Lett. **98**(2), 023001 (2007).
[52]U. Werner, R. Mitrić, T. Suzuki, and V. Bonačić-Koutecký, Chem. Phys. **349**(1-3), 319–324 (2008).
[53]J. Y. Fang and S. Hammes-Schiffer, J. Phys. Chem. A **103**(47), 9399–9407 (1999).
[54]G. Granucci and M. Persico, J. Chem. Phys. **126**(13), 134114 (2007).
[55]B. F. E. Curchod, I. Tavernelli, and U. Rothlisberger, Phys. Chem. Chem. Phys. **13**(8), 3231–3236 (2011).
[56]H. Langer and N. L. Doltsinis, Phys. Chem. Chem. Phys. **6**(10), 2742–2748 (2004).
[57]C. F. Craig, W. R. Duncan, and O. V. Prezhdo, Phys. Rev. Lett. **95**(16), 163001 (2005).
[58]R. Mitrić, U. Werner, M. Wohlgemuth, G. Seifert, and V. Bonačić-Koutecký, J. Phys. Chem. A **113**(45), 12700–12705 (2009).
[59]S. Klein, M. J. Bearpark, B. R. Smith, M. A. Robb, M. Olivucci, and F. Bernardi, Chem. Phys. Lett. **292**(3), 259–266 (1998).
[60]L. Yu, C. Xu, Y. B. Lei, C. Y. Zhu, and Z. Y. Wen, Phys. Chem. Chem. Phys. **16**(47), 25883–25895 (2014).
[61]M. Barbatti, M. Ruckenbauer, F. Plasser, J. Pittner, G. Granucci, M. Persico, and H. Lischka, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **4**(1), 26–33 (2014).
[62]S. Kilina, D. Kilin, and S. Tretiak, Chem. Rev. **115**(12), 5929–5978 (2015).
[63]A. Toniolo, C. Ciminelli, M. Persico, and T. J. Martínez, J. Chem. Phys. **123**(23), 234308 (2005).
[64]G. Granucci, M. Persico, and A. Toniolo, J. Chem. Phys. **114**(24), 10608–10615 (2001).
[65]S. H. Xia, G. L. Cui, W. H. Fang, and W. Thiel, Angew. Chem., Int. Ed. **55**(6), 2067–2072 (2016).
[66]G. A. Worth, P. Hunt, and M. A. Robb, J. Phys. Chem. A **107**(5), 621–631 (2003).
[67]J. W. Park and T. Shiozaki, J. Chem. Theory Comput. **13**(8), 3676–3683 (2017).
[68]M. Sebastian, M. Philipp, and L. González, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **8**(6), e1370 (2018).
[69]L. M. Frutos, T. Andruniów, F. Santoro, N. Ferré, and M. Olivucci, Proc. Natl. Acad. Sci. U. S. A. **104**(19), 7764–7769 (2007).
[70]X. H. Zhuang, J. Wang, and Z. G. Lan, J. Phys. Chem. B **117**(50), 15976–15986 (2013).
[71]D. P. Hu, J. Huang, Y. Xie, L. Yue, X. H. Zhuang, and Z. G. Lan, Chem. Phys. **463**, 95–105 (2015).
[72]Z. G. Lan, Y. Lu, O. Weingart, and W. Thiel, J. Phys. Chem. A **116**(6), 1510–1518 (2012).
[73]O. Weingart, Z. G. Lan, A. Koslowski, and W. Thiel, J. Phys. Chem. Lett. **2**(13), 1506–1509 (2011).
[74]O. Weingart, P. Altoè, M. Stenta, A. Bottoni, G. Orlandi, and M. Garavelli, Phys. Chem. Chem. Phys. **13**(9), 3645–3648 (2011).
[75]M. Barbatti, A. J. A. Aquino, and H. Lischka, Mol. Phys. **104**(5-7), 1053–1060 (2006).
[76]I. Schapiro, O. Weingart, and V. Buss, J. Am. Chem. Soc. **131**(1), 16–17 (2009).
[77]C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer Science+Business Media, New York, 2006).
[78]A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins **17**(4), 412–425 (1993).
[79]I. Jolliffe, *Principal Component Analysis* (Wiley Online Library, 2002).
[80]I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer Science & Business Media, America, 2005).
[81]V. De Silva and J. B. Tenenbaum, *Sparse Multidimensional Scaling Using Landmark Points*, Technical Report, Stanford University, 2004.
[82]M. Balasubramanian and E. L. Schwartz, Science **295**(5552), 7a (2002).
[83]J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science **290**(5500), 2319–2323 (2000).
[84]R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U. S. A. **102**(21), 7426–7431 (2005).
[85]R. R. Coifman and S. Lafon, Appl. Comput. Harmon. Anal. **21**(1), 5–30 (2006).

[86]G. E. Hinton and R. R. Salakhutdinov, Science **313**(5786), 504–507 (2006).
[87]P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U. S. A. **103**(26), 9885–9890 (2006).
[88]M. Ceriotti, G. A. Tribello, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **108**(32), 13023–13028 (2011).
[89]M. A. Rohrdanz, W. W. Zheng, and C. Clementi, Annu. Rev. Phys. Chem. **64**, 295–316 (2013).
[90]M. A. Rohrdanz, W. W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**(12), 124116 (2011).
[91]W. W. Zheng, M. A. Rohrdanz, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**(14), 144109 (2011).
[92]G. A. Tribello, M. Ceriotti, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **109**(14), 5196–5201 (2012).
[93]J. P. P. Zauleck, S. Thallmair, M. Loipersberger, and R. de Vivie-Riedle, J. Chem. Theory Comput. **12**(12), 5698–5708 (2016).
[94]J. P. P. Zauleck and R. de Vivie-Riedle, J. Chem. Theory Comput. **14**(1), 55–62 (2018).
[95]X. S. Li, Y. Xie, D. P. Hu, and Z. G. Lan, J. Chem. Theory Comput. **13**(10), 4611–4623 (2017).
[96]X. S. Li, Y. Xie, D. P. Hu, and Z. G. Lan, J. Chem. Theory Comput. **13**(12), 6434 (2017).
[97]A. M. Virshup, J. H. Chen, and T. J. Martínez, J. Chem. Phys. **137**(22), 22A519 (2012).
[98]A. K. Belyaev, W. Domcke, C. Lasser, and G. Trigila, J. Chem. Phys. **142**(10), 104307 (2015).
[99]G. Capano, T. J. Penfold, I. Tavernelli, and M. Chergui, Phys. Chem. Chem. Phys. **19**, 19590–19600 (2017).
[100]A. J. Atkins and L. González, J. Phys. Chem. Lett. **8**(16), 3840–3845 (2017).
[101]A. Nayyeri and A. Sidiropoulos, Autom., Languages, Program. **9134**, 997–1009 (2015).
[102]P. C. Besse, B. Guillouet, J. M. Loubes, and F. Royer, IEEE Trans. Intell. Transp. Syst. **17**(11), 3306–3317 (2016).
[103]A. Efrat, L. J. Guibas, S. Har-Peled, J. S. B. Mitchell, and T. M. Murali, Discrete Comput. Geom. **28**(4), 535–569 (2002).
[104]J. Rydzewski and W. Nowak, J. Chem. Theory Comput. **12**(4), 2110–2120 (2016).
[105]X. He, Y. Shen, F. R. Hung, and E. E. Santiso, J. Chem. Phys. **143**(12), 124506 (2015).
[106]E. E. Santiso and B. L. Trout, J. Chem. Phys. **143**(17), 174109 (2015).
[107]N. Musolino and B. L. Trout, J. Chem. Phys. **138**(13), 134707 (2013).
[108]S. L. Seyler, A. Kumar, M. F. Thorpe, and O. Beckstein, PLoS Comput. Biol. **11**(10), e1004568 (2015).
[109]K. Toohey and M. Duckham, SIGSPATIAL Spec. **7**(1), 43–50 (2015).
[110]G. Yuan, P. Sun, J. Zhao, D. Li, and C. Wang, Artif. Intell. Rev. **47**(1), 123–144 (2017).
[111]J. Bian, D. Tian, Y. Tang, and D. Tao, preprint arXiv:1802.06971 (2018).
[112]T. Eiter and H. Mannila, Technical Report No. CD-TR 94/64, 1994.
[113]H. Alt and M. Godau, Int. J. Comput. Geom. Appl. **5**(1-2), 75–91 (1995).
[114]M. H. Jiang, Y. Xu, and B. H. Zhu, J. Bioinf. Comput. Biol. **06**(01), 51–64 (2008).
[115]M. A. Mroginski, D. H. Murgida, D. von Stetten, C. Kneip, F. Mark, and P. Hildebrandt, J. Am. Chem. Soc. **126**(51), 16734–16735 (2004).
[116]C. Kneip, P. Hildebrandt, W. Schlamann, S. E. Braslavsky, F. Mark, and K. Schaffner, Biochemistry **38**(46), 15185–15192 (1999).
[117]F. Andel, J. T. Murphy, J. A. Haas, M. T. McDowell, I. van der Hoef, J. Lugtenburg, J. C. Lagarias, and R. A. Mathies, Biochemistry **39**(10), 2667–2676 (2000).
[118]D. H. Murgida, D. von Stetten, P. Hildebrandt, P. Schwinté, F. Siebert, S. Sharda, W. Gärtner, and M. A. Mroginski, Biophys. J. **93**(7), 2410–2417 (2007).
[119]J.-y. Hasegawa, M. Isshiki, K. Fujimoto, and H. Nakatsuji, Chem. Phys. Lett. **410**(1), 90–93 (2005).
[120]F. Velazquez Escobar, D. von Stetten, M. Günther-Lütkens, A. Keidel, N. Michael, T. Lamparter, L.-O. Essen, J. Hughes, W. Gärtner, Y. Yang, K. Heyne, M. A. Mroginski, and P. Hildebrandt, Front. Mol. Biosci. **2**, 37 (2015).
[121]B. Durbeej, O. A. Borg, and L. A. Eriksson, Chem. Phys. Lett. **416**(1), 83–88 (2005).
[122]B. Durbeej and L. A. Eriksson, Phys. Chem. Chem. Phys. **8**(35), 4053–4071 (2006).
[123]B. Durbeej, Phys. Chem. Chem. Phys. **11**(9), 1354–1361 (2009).
[124]P. Altoè, T. Climent, G. C. De Fusco, M. Stenta, A. Bottoni, L. Serrano-Andrés, M. Merchán, G. Orlandi, and M. Garavelli, J. Phys. Chem. B **113**(45), 15067–15073 (2009).
[125]W. Kabsch, Acta Crystallogr., Sect. A **34**(5), 827–828 (1978).
[126]W. Kabsch, Acta Crystallogr., Sect. A **32**(5), 922–923 (1976).
[127]J. D. Mazimpaka and S. Timpf, J. Spat. Inf. Sci. **2016**(13), 61–99.
[128]N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, J. Comput. Chem. **32**(10), 2319–2327 (2011).
[129]H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery **1**(3), 231–240 (2011).
[130]M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (AAAI Press, Portland, Oregon, 1996), pp. 226–231.
[131]T. W. Keal, A. Koslowski, and W. Thiel, Theor. Chem. Acc. **118**(5-6), 837–844 (2007).
[132]A. Koslowski, M. E. Beck, and W. Thiel, J. Comput. Chem. **24**(6), 714–726 (2003).
[133]W. Weber and W. Thiel, Theor. Chem. Acc. **103**(6), 495–506 (2000).
[134]W. Thiel, MNDO Program, Version 7.0, Mülheim an der Ruhr, Germany, 2007.
[135]See https://www.mdanalysis.org/docs/documentation_pages/analysis/psa.html for the information about the geometry alignment.
[136]D. P. Hu, Y. F. Liu, A. L. Sobolewski, and Z. G. Lan, Phys. Chem. Chem. Phys. **19**(29), 19168–19177 (2017).
[137]L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, and R. Layton, "API design for machine learning software: Experiences from the Scikit-learn project," preprint arXiv:1309.0238 (2013).
[138]F. Pedregosa, G. Varoquaux, and A. Gramfort, J. Mach. Learn. Res. **12**, 2825–2830 (2011), available at http://www.jmlr.org/papers/v12/pedregosa11a.html.
[139]J. C. Tully, J. Chem. Phys. **137**(22), 22A301 (2012).
[140]J. Behler and M. Parrinello, Phys. Rev. Lett. **98**(14), 146401 (2007).
[141]L. F. Zhang, J. Q. Han, H. Wang, R. Car, and E. Weinan, Phys. Rev. Lett. **120**(14), 143001 (2018).
[142]A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**(18), 184115 (2013).
[143]E. Sriraghavendra, K. Karthik, and C. Bhattacharyya, *International Conference on Document Analysis and Recognition* (IEEE, 2007), pp. 461–465.