HAL
open science

Graph theory for automatic structural recognition in molecular dynamics simulations

S. Bougueroua, Riccardo Spezia, S. Pezzotti, S. Vial, F. Quessette, D. Barth, M.-P. Gaigeot

To cite this version:

S. Bougueroua, Riccardo Spezia, S. Pezzotti, S. Vial, F. Quessette, et al.. Graph theory for automatic structural recognition in molecular dynamics simulations. The Journal of Chemical Physics, 2018, 149 (18), pp.184102. 10.1063/1.5045818 . hal-03967341

HAL Id: hal-03967341

https://hal.science/hal-03967341v1

Submitted on 1 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Graph theory for automatic structural recognition in molecular dynamics simulations

S. Bougueroua, R. Spezia, S. Pezzotti, S. Vial, F. Quessette, D. Barth, and M.-P. Gaigeot

Citation: J. Chem. Phys. **149**, 184102 (2018); doi: 10.1063/1.5045818
View online: https://doi.org/10.1063/1.5045818
View Table of Contents: http://aip.scitation.org/toc/jcp/149/18
Published by the American Institute of Physics

PHYSICS TODAY
WHITEPAPERS

ADVANCED LIGHT CURE ADHESIVES

Take a closer look at what these environmentally friendly adhesive systems can do

READ NOW

PRESENTED BY
MASTERBOND
ADHESIVES | SEALANTS | COATINGS

THE JOURNAL OF CHEMICAL PHYSICS **149**, 184102 (2018)

Graph theory for automatic structural recognition in molecular dynamics simulations

S. Bougueroua,[1,a)] R. Spezia,[1] S. Pezzotti,[1] S. Vial,[2] F. Quessette,[2] D. Barth,[2] and M.-P. Gaigeot[1,b)]

[1]*LAMBE UMR8587, Univ. Evry, Université d'Evry Val d'Essonne, CNRS, CEA, Université Paris-Saclay, Laboratoire Analyse et Modélisation pour la Biologie et l'Environnement, 91025 Evry, France*

[2]*DAVID, Université de Versailles Saint-Quentin-en-Yvelines, Université Paris-Saclay, Données et Algorithmes pour une Ville Intelligente et Durable, 78035 Versailles, France*

(Received 25 June 2018; accepted 25 October 2018; published online 9 November 2018)

Graph theory algorithms have been proposed in order to identify, follow in time, and statistically analyze the changes in conformations that occur along molecular dynamics (MD) simulations. The atomistic granularity level of the MD simulations is maintained within the graph theoric algorithms proposed here, isomorphism is a key component together with keeping the chemical nature of the atoms. Isomorphism is used to recognize conformations and construct the graphs of transitions, and the reduction in complexity of the isomorphism has been achieved by the introduction of "orbits" and "reference snapshots." The proposed algorithms are applied to MD trajectories of gas phase molecules and clusters as well as condensed matter. The changes in conformations followed over time are hydrogen bond(s), proton transfer(s), coordination number(s), covalent bond(s), multiple fragmentation(s), and H-bonded membered rings. The algorithms provide an automatic analysis of multiple trajectories in parallel, and can be applied to *ab initio* and classical MD trajectories alike, and to more coarse grain representations. *Published by AIP Publishing.* https://doi.org/10.1063/1.5045818

I. INTRODUCTION

Molecular dynamics (MD) simulations are certainly an essential method in theoretical and computational chemistry for characterizing structures, dynamics, chemical reactions, and spectroscopy of matter at the atomic and sub-atomic level of representation.[1–3] The trajectories obtained by MD simulations generate large data that have to be analyzed in order to extract microscopic and macroscopic properties. MD trajectories can be visualised through software such as Visual Molecular Dynamics (VMD)[4] to cite only one popular software, where the time evolution of the molecular system is animated snapshot by snapshot. While visualizing trajectory movies are straightforward for gas phase systems and might be directly used for extracting structural properties from the trajectory, it is not sufficient for condensed matter trajectories related to liquids and/or inhomogeneous solid/liquid interfaces for instance, where data analyses are overwhelmed by the environment. Beside direct visualisation, these softwares also allow to easily extract the time evolution of structural properties such as chosen covalent bond distances/H-bond distances/angles, distance/angular distributions or radial pair distribution functions for a more averaged picture, and global measures of changes in conformations by means of RMSD (Root Mean Square Displacement).

Such analyses however do not provide a direct knowledge of the global changes in the 3D-structures over time (whether one is interested in the whole molecular structure or only in part of it). Algorithms have to be proposed in order to achieve this goal. These algorithms have to be of low complexity and efficient, in order to be applied to the analysis of thousands of trajectories simultaneously. In many applications, 100′-1000′ MD trajectories are indeed generated and have to be analyzed in order to get a relevant statistical view of the structural properties. Graph theory is certainly one solution. Although graph representations have been introduced in chemistry for some time,[5,6] their use in extracting structural properties from MD trajectories is still scarce,[7–11] and we propose here graph theory algorithms in order to identify and follow in time 3D conformation(s) and extract statistical graphs of the transitions that occur in between conformers and isomers and/or in between chemically reactive species, including the statistical analysis of these transitions. As will be shown, these algorithms are applicable/transferable from gas phase molecules and clusters to condensed phase systems.

Recent implementations of graph theory in chemistry and statistical analysis of MD trajectories have shared a similar degree of granularity in the graph theory representation, leading to rather simple graphs. Clark *et al.*,[7,8,12,13] Pastor *et al.*,[9] Tenney and Cygan,[10] Choi *et al.*,[6] and Pietrucci *et al.*[11,14] have been mostly interested in recognizing atomic and molecular cluster geometries explored during MD simulations (gas phase and liquid phase trajectories), and/or exploring potential energy surfaces of gas phase molecules.[15,16] Presumably the more elaborated graph theory up-to-now developed in the chemistry and material communities are from Pietrucci *et al.*,[11,14] but with the purpose of driving biased MD simulations, which entirely justifies the use of simple graphs. To that end, each vertex of a graph is composed by one atom/one

a)Electronic mail: sana.bougueroua@univ-evry.fr
b)Electronic mail: mgaigeot@univ-evry.fr

molecule, and the chemical nature of the atoms/molecules is not taken into account. While these graphs are relatively easy to analyze, they however lack chemical information (e.g., covalent bonds, hydrogen bonds, exchange of atoms in homogeneous clusters, etc. . .) that might be relevant for a more detailed characterization of the structures. Many of the developed graph theory algorithms for chemistry in the literature furthermore use adjacency and/or geodesic matrices in order to compare structures sampled over the MD trajectory, which might not be the most efficient method for recognizing identical structures where chemically identical atoms have been swapped, while their ID number in the atoms list keep them non-identical.

We propose here graph theoric based algorithms where the granularity of the target graphs remains at the atomic level, i.e., one atom per graph vertex, and where the information on the hydrogen bond direction (i.e., acceptor/donor) is kept. As will be presented and discussed in Sec. II, one specificity is the atomic granularity level that is kept within the graphs, which is instrumental for the transferability of the method to very diverse molecular systems. The core of our method is based on graph isomorphism. While the McKay algorithm is a well-known algorithm for graph isomorphism,[23,24] the identified specific instances for our applications for the recognition of the changes in the conformations along MD trajectories led us to specialize this algorithm in order to reduce its complexity level. We also introduce novel features such as orbits and reference snapshots in order to reduce further the algorithm complexity and save computational time. A coloration scheme according to the chemical nature of each atom has also been included in the representation, such that the comparison of graphs keeps the chemical atomic information needed to analyze the isomeric conformations of molecules and clusters. All these features make our graph theory algorithms unique and of low complexity. While the graphs might appear more complex at this granularity level at first glance, they are however more precise for their use in chemical reactions dynamics as well as in hydrogen bonds dynamics in liquids, which are two of our goals.

One more originality of our work consists in constructing graphs of transitions, which goes beyond the "simple" recognition of conformations along time. To our best knowledge, these graphs have not been presented in other graph theoric works in the context of MD trajectories analyses. These graphs provide a statistical view of the dynamics without keeping track of the time but rather providing the statistical analysis of the conformers explored (represented by the vertices of the graphs of transitions) and providing the transitions between these conformers (represented by the directed edges of the graphs). The latter keep the information on the direction in the change in conformation, and one further advantage is that the degree of dynamicity of the molecular system can be assessed at one single glance, i.e., $A \rightarrow B$ or $B \rightarrow A$ if the transition occurs between A/B molecular systems (isomers, conformers, and chemical reactive fragments alike). Such graphs of transitions provide, at one single glance, all conformational events that occurred during the trajectory. One should also note that when analyzing several trajectories of the same molecular system (e.g., at different temperatures, starting from various initial conformations), the "total" graph of transitions (extracted as a statistics over all trajectories) becomes even more relevant as it is a coarse-grain view of the potential energy surface sampled at given thermodynamics conditions.

Our graph theory based method is presented in Secs. II and III introducing the algorithm for the recognition of the changes in the molecular conformations, *without* using manual construction and/or reference structures, and is transferable in between molecular systems. The algorithms proposed in this paper are applied to trajectories displaying "simple" conformational and isomeric transitions over time and more "complex" chemically reactive trajectories; see Sec. IV. These trajectories are extracted from our previous studies and refer to *ab initio* trajectories of isolated peptides,[17,18] ionic clusters,[19] and direct chemical dynamics of peptides fragmentations.[20] The same algorithm is straightforwardly applied to analyze the more complex air/liquid water interface from our previous work[21] at the end of Sec. IV D and hence provides a direct and easy analysis of the 2D-H bonding network that exists at this particular interface. For conveniency, the present work is in relation with our previous *ab initio* MD simulations, but it is not restricted to these rather short time scale and small size scales, as the performance analyses demonstrate.

II. MODEL BASED ON GRAPH THEORY

A. Definition of a molecular conformation and relationship to graph theory

Our algorithm is based on geometrical criteria to define a molecular conformation coupled to graph theory in order to extract the knowledge of isomorphism and possible change in conformations along the trajectory. One requirement is that the algorithm has to be of low complexity, such that the method can be fast. For the present work, our algorithm is developed for analyzing molecular dynamics trajectories in terms of intermolecular bonds (through hydrogen bonding or electrostatic intermolecular interactions) and/or covalent bonds. These are the changes in conformations tracked along time using graph theory. In order to analyze the molecular dynamics trajectories in terms of conformations explored over time, one crucial step is to set up a model that defines a molecular conformation. Such model should represent a conformation with the right level of granularity, which next allows us to identify the conformational changes occuring over time.

We have chosen to define a molecular conformation in terms of covalent and hydrogen bonds formed between the atoms; the definitions are based on geometrical criteria as described below (usual definitions in the chemistry community):

- **A covalent bond** is formed between a pair of atoms $[a, b]$ with respective Cartesian coordinates (x_a, y_a, z_a) and (x_b, y_b, z_b), if $\sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}$ (the Euclidean distance) is less than a cut-off distance D_C (D_C is chosen as the sum of covalent radii of atoms a and b in our work).
- **A hydrogen bond (H-bond)** is defined between a hydrogen atom (donor of H-bond) with Cartesian

coordinates (x_h, y_h, z_h) [the Cartesian coordinates of the heavy atom covalently bonded to this hydrogen are (x_d, y_d, z_d)] and an acceptor atom with Cartesian coordinates (x_a, y_a, z_a), if the Euclidean distance $\sqrt{(x_h - x_a)^2 + (y_h - y_a)^2 + (z_h - z_a)^2}$ is less than a cut-off distance D_H (2.3 Å), together with the angle $\widehat{dha}$ between the three atoms, being in the interval $[\pi - \alpha_h, \pi + \alpha_h]$ where α_h is an input parameter ($\frac{\pi}{3}$). Our convention is to represent an H-bond by the couple (heavy atom, acceptor).

- In the case of ionic clusters of the type $M^{\pm}(H_2O)_n$, where M is a metal ion (e.g., Li^+ in Sec. IV C) and H_2O is water, on top of the H-bonds formed in-between the water molecules (and defined as above), we also define and use M–O (and/or M–M) intermolecular distances in order to define the cluster conformations. One can see $M^{\pm}$ interactions with the water molecules through "intermolecular electrostatic interactions," although water is neutral. We therefore keep this name afterwards to identify such interactions.

In graph theory, a molecular conformation can be viewed as a molecular graph: in our work, atoms are the vertices of the graph, while covalent bonds are the edges. The originality of our model consists in using a mixed graph, i.e., containing edges and arcs. Hydrogen atoms are thus not represented in the graph, i.e., the covalent bonds involving these hydrogen atoms are not represented, and in order to identify the donor and acceptor atoms in any H-bond, a directed edge going from the donor to the acceptor is used.

In a formal way, a molecular conformation is a **mixed graph** $G = (V, E_C, A_H, E_I)$, with the following definitions:

- V is the set of all **atoms** of the system except the hydrogen atoms, where each atom is a **vertex** of G,
- $E_C = \{[a, b], a \in V, b \in V$: $[a, b]$ is a **covalent bond**$\}$, where each covalent bond represents an **edge** in G ($[a, b] = [b, a]$),
- $A_H = \{(a, b), a \in V, b \in V$: (a, b) is a **H-bond**$\}$, where each H-bond represents an **arc** in G ($(a, b) \neq (b, a)$), and
- $E_I = \{[a, b], a \in V, b \in V$: $[a, b]$ is an **electrostatic interaction**$\}$, where each electrostatic interaction represents an **edge** in G ($[a, b] = [b, a]$).

We also define a function ϕ:$V \rightarrow T = \{H, C, O, N, \ldots\}$ that provides the chemical type of atoms. For example, $\phi(a) = O$ if the atom a is an oxygen atom.

The Cartesian atomic positions taken from the trajectory are used *only* for the construction of the mixed graphs. Once the graphs are constructed, the changes in the conformations are analyzed through the comparison of these graphs as described in Sec. II B.

B. Isomorphism for conformational isomerism

A change in the molecular conformation from time t to $t + \Delta t$ is identified if and only if there has been at least one change in a bond set, i.e., having $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$ and $G_{i+1} = (V_{i+1}, E_{C_{i+1}}, A_{H_{i+1}}, E_{I_{i+1}})$ for the two consecutive conformations, the change can be related to one (or several) of the following instances:

- An **appearance** of a new covalent bond $[a, b]$ if $[a, b] \in E_{C_{i+1}}$ and $[\theta_{i,i+1}(a), \theta_{i,i+1}(b)] \notin E_{C_i}$.
- A **disappearance** of an existing covalent bond $[a, b]$ if $[a, b] \in E_{C_i}$ and $[\theta_{i,i+1}(a), \theta_{i,i+1}(b)] \notin E_{C_{i+1}}$.
- An **appearance** of a new H-bond (a, b) if $(a, b) \in A_{H_{i+1}}$ and $(\theta_{i,i+1}(a), \theta_{i,i+1}(b)) \notin A_{H_i}$.
- A **disappearance** of an existing H-bond (a, b) if $(a, b) \in A_{H_i}$ and $(\theta_{i,i+1}(a), \theta_{i,i+1}(b)) \notin A_{H_{i+1}}$.
- A **proton transfer** through a H-bond (a, b) if $(a, b) \in A_{H_i}$ and $(\theta_{i,i+1}(b), \theta_{i,i+1}(a)) \in A_{H_{i+1}}$.
- An **appearance** of a new electrostatic interaction $[a, b]$ if $[a, b] \in E_{I_{i+1}}$ and $[\theta_{i,i+1}(a), \theta_{i,i+1}(b)] \notin E_{I_i}$.
- A **disappearance** of an existing electrostatic interaction $[a, b]$ if $[a, b] \in E_{I_i}$ and $[\theta_{i,i+1}(a), \theta_{i,i+1}(b)] \notin E_{I_{i+1}}$.

From graph theory perspective, extracting the conformational isomers that have been explored along the trajectory can be seen as an exploration of the graph topologies. This can be done by graph isomorphism, which allows deciding if two graphs are identical.[22–25] We define the isomorphism between two conformations as follows: two conformations $G_i = (V_i, E_{C_i}, A_{H_i}, E_{I_i})$ and $G_j = (V_j, E_{C_j}, A_{H_j}, E_{I_j})$ are **isomorphic** if and only if there exists a bijection $\theta_{i,j}$:$V_i \rightarrow V_j$ such as:

1. $\forall a \in V_i$, $\phi(a) = \phi(\theta_{i,j}(a))$,
2. $[a, b] \in E_{C_i} \Leftrightarrow [\theta_{i,j}(a), \theta_{i,j}(b)] \in E_{C_j}$,
3. $(a, b) \in A_{H_i} \Leftrightarrow (\theta_{i,j}(a), \theta_{i,j}(b)) \in A_{H_j}$, and
4. $[a, b] \in E_{I_i} \Leftrightarrow [\theta_{i,j}(a), \theta_{i,j}(b)] \in E_{I_j}$.

In other words, two conformations are isomorphic if they have the same sets of covalent bonds, H-bonds, and electrostatic interactions by allowing interchanging atoms of the same chemical type. Figure 1 shows an example of two conformations that are found isomorphic (A and B) and one conformation (C) which is not isomorphic to any of them. Different methods have been developed in the literature for solving graph isomorphism.[22,26–29] We apply the **Mckay**

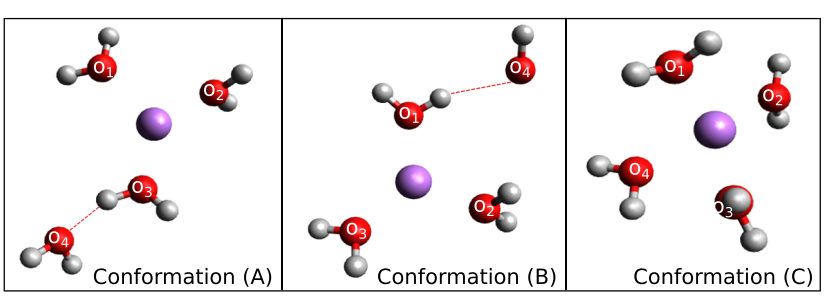

FIG. 1. Example of isomorphic conformations. Conformations (A) and (B) are isomorphic by interchanging oxygen atoms labeled (O_1) and (O_3), while conformation (C) is isomorphic to neither (A) nor (B).

method[23] that is considered as one of the most efficient methods in practice. It is based on the canonical labeling of graphs, which consists in placing the vertex labels in a way that does not depend on their initial labeling. Each graph has a unique canonical labeling. Two graphs are hence isomorphic if they have the same canonical labeling.

There are specific instances in the identification of the changes in the molecular structures that we can take advantage of, such as to reduce the number of times in the application of the isomorphism McKay algorithm, hence reducing the whole complexity of the algorithm. This is what has been done in the present work, introducing "orbits" and "reference snapshots," described in detail in Sec. III. This is a part of the original features of the algorithm developed in this work.

III. ALGORITHM

Our algorithm hence provides the set $\mathcal{G}$ of the conformational isomers $G_1, G_2, \ldots, G_k$ explored over time with their respective times of appearance and periods of existence, from a given sequence of snapshots $I_1, I_2, \ldots, I_S$ taken from a given molecular dynamics trajectory. The algorithm is based on two major steps:

1. **Initialization**: the algorithm reads the first snapshot I_1 and provides the first conformation G_1 by computing the different bond sets as described above. The set $\mathcal{G}$ is initialized to (G_1, I_1).
2. **Conformational dynamics analysis**: at each following snapshot of the trajectory:
   (a) Read a new snapshot I_i and define the associated conformation G_i.
   (b) Test if G_i is isomorphic to G_{i-1}:
      - If G_i is identical to G_{i-1} (comparison of adjacency matrices), conformation G_{i-1} is maintained at snapshot I_i, and the subsequent snapshot is next analyzed [return to step (a)].
      - If G_i is not isomorphic to G_{i-1}, a second test of isomorphism is applied [described hereafter in (c)].
   (c) Recognition to conformations already identified: test if there is a conformation G_j in the already analyzed set $\mathcal{G}$ that is isomorphic to G_i:
      - If such a conformation exists, there is therefore the appearance of an already existing conformation at snapshot I_i, and the information (G_j, I_i) is added to $\mathcal{G}$.
      - If there is no such conformation, a new conformation has hence been identified at snapshot I_i and the information (G_i, I_i) is added to $\mathcal{G}$.
   (d) Return to step (a) in order to read the subsequent snapshot.

At the end of the analysis, the set of conformations explored along the trajectory is obtained, as well as the time sequence of their having been visited and the mean residence time for each conformation.

Such an algorithm can be further refined and computational costs further decreased by taking advantage of specific instances. Such instances have been identified, leading to the following novel features "orbits" and "reference snapshots," and ultimately provide a reduction in the complexity of the isomorphism algorithm. These instances are now described.

The basic method to compute H-bonds requires to browse all hydrogen atoms, and for each one to check if there is one atom among *all* atoms of the system that could be an acceptor of H-bond. However, a hydrogen atom can form a H-bond only with a subset of atoms that are closely located. We hence define an **orbit** for each hydrogen atom. It is composed by a subset of atoms that are located at a given distance from the hydrogen atom lower than a cut-off distance $D_H \times \alpha$ (where D_H is the cut-off distance used for H-bonds and α is a coefficient with a default value of 3; this value was set after multiple tests on different trajectories and details have been presented in a related work[30]). Considering only the atoms within the orbits instead of *all* atoms of the system to compute H-bonds reduces the number of comparisons to be performed. Obviously, this method requires that the orbits are not recalculated at each snapshot of the trajectory, otherwise that would be computationally costly. One has to keep in mind, at least for the *ab initio* MD trajectories analyzed in this work, that two consecutive snapshots differ by 0.1-0.5 fs in time; therefore, no large modifications of the atoms that belong to one orbit should be expected from one time step to the next one.

We therefore define a subset of snapshots where orbits have to be recomputed. These snapshots are called **reference snapshots**. To decide if a snapshot I_i is such a reference snapshot, the displacements of atoms in snapshot I_i are analyzed according to the current reference snapshot. If the whole displacement is greater than a cut-off distance $D_H \times (\alpha - 1)$ (same parameters as for orbits), the reference snapshot is changed to I_i, and orbits are thus recomputed. The same philosophy is used to compute orbits and reference snapshots for covalent bonds and electrostatic interactions. The identification of the reference snapshots does not impact the complexity of the algorithm, as the necessary atomic displacement calculations are simultaneously done with the reading of the snapshots.

The computationally costly part of the algorithm is the isomorphism. Graph isomorphism is known as a non-polynomial mathematical problem.[31,32] The key component of our algorithm is to be able to reduce the number of isomorphism tests to be performed along the trajectory. Therefore, the isomorphism test at snapshot I_i is applied only if this snapshot is a reference snapshot, i.e., at least one orbit has changed. For the rest of the snapshots, a basic comparison of adjacency matrices is performed to decide if there has been a conformational change. One can easily understand that such instance reduces the algorithm complexity. If one were applying the isomorphism at each time-step of the dynamics, i.e., systematically comparing two successive graphs, the complexity of the algorithm would indeed depend on the time-step of the dynamics, and would therefore depend on the duration of the trajectory. If instead, one applies the isomorphism only at times when

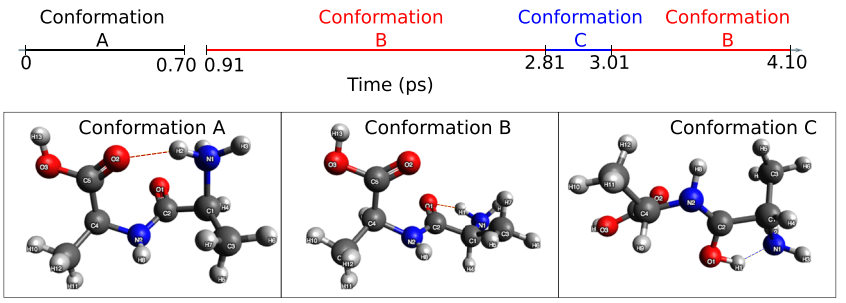

FIG. 2. Example of evolution of conformations along a trajectory identified by our algorithm using graph theory. Top: provides the time sequence of appearance of the conformations; Bottom: provides the 3D structures of the identified conformations.

one change in the conformation has been pre-identified, the algorithm thus only depends on the number of conformational changes, which can represent a huge decrease in the algorithm complexity. This is obviously relevant as long as the number in the conformational changes is not too large over the whole trajectory. This is indeed the case in *ab initio* MD trajectories where two subsequent snapshots are separated by 0.1-0.5 fs, without significant conformational changes expected within this interval. If however one applies the algorithm to classical MD trajectories with snapshots stored at rather distant intervals (e.g., 100-1000 time steps, that could amount up to 0.1-1.0 ps intervals for 1-10 fs time steps), one can imagine a far lower gain in the algorithm complexity, up to the extreme limit where the isomorphism would be applied at each time step. When this happens, our algorithm recovers the complexity of the standard McKay algorithm.

Once the whole trajectory has been analyzed, one has to sort out the identified conformations in terms of relevance for their time period of existence. Only the conformations existing for a total time T_r (5%) over the whole trajectory are sorted out.

Figure 2 shows an example of the evolution of conformations along a trajectory. In this illustration, there are three conformations (A), (B), and (C) identified along the dynamics, where conformation (B) appears twice and has the highest mean residence time, for a total of 2.85 ps. Figure 2 also shows a graphic representation of these conformations. We hence observe that there are changes in H-bonds when going from conformation (A) to conformation (B); there is a proton transfer when going from conformation (B) to conformation (C).

IV. EXPERIMENTAL VALIDATION OF THE ALGORITHM

Our algorithm has been primarily implemented in C-language, and it runs on Intel Core i7 processors. It has been tested and validated on *ab initio* MD trajectories of gas phase molecular systems at finite temperature. These trajectories are taken from our previous studies, either dedicated to the dynamics of gas phase protonated peptides of increasing size and complexity [$NH_3^+ - Ala_2 - COOH$ ($C_6H_{13}N_2O_3$) and Z-Ala_6—COOH ($C_{26}H_{39}N_7O_8$)], or dedicated to collisional induced fragmentation of peptides [here the gly2-NH_3^+ peptide colliding with an Ar atom ($C_4H_{10}N_3O_2Ar$)], or dedicated to clusters dynamics [here the $Li^+(H_2O)_4$ cluster]. These systems have been published elsewhere and analyzed in these papers for vibrational spectroscopy[17,19,34] and CID (Collision Induced Dissociation).[20] Cut-off values are 2.3 Å for H-bond distances (D_H), 60° for a H-bond angle ($\widehat{dha}$), and $\alpha = 3$ when defining orbits and reference snapshots.

A. Structural recognition of isolated protonated peptides trajectories

Two peptides from our past investigations have been chosen [all from density functional theory (DFT) based molecular dynamics simulations]: dipeptide $NH_3^+ - Ala_2 - COOH$ and peptide Z-Ala_6—COOH. Conformations for these systems are determined exclusively from the non-covalent interactions, more precisely from the evolution of H-bonds and proton transfers over time. No covalent bond has been broken/formed along the trajectory. In the analysis of conformational dynamics, we set the covalent bonds as they are found in the first snapshot and subsequently only observe the evolution of H-bonds over time.

The trajectory for $NH_3^+ - Ala_2 - COOH$ contains 10 200 snapshots (i.e., 4.0 ps of trajectory), the molecule is composed of 24 atoms, and the running time of the algorithm analysis was 0.88 s (see Table II). Four conformations have been sampled along the trajectory; see Fig. 3 for their 3D representations: conformation (1) has the N-terminal NH_3^+ that is located close to the C-terminal COOH, thus forming one

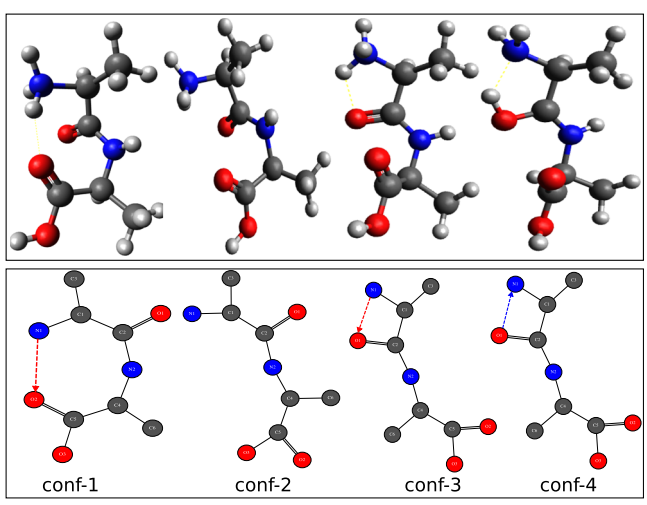

FIG. 3. Schematic representation of the conformations of $NH_3^+ - Ala_2 - COOH$ explored along the dynamics and analyzed by the present graph theory algorithm. Top of the figure shows the 3D representations of the conformations. Bottom of the figure represents 2D simplified graphs of the conformations. See the text for details.

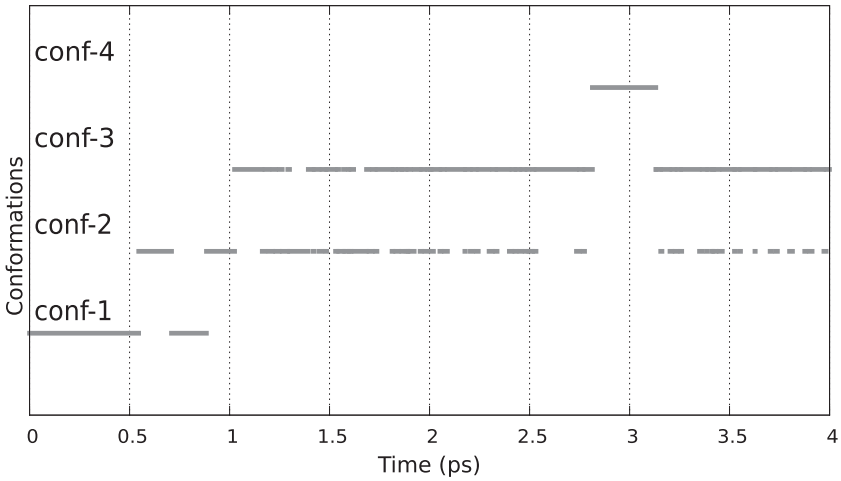

FIG. 4. Time evolution of the conformations explored along one trajectory of the $NH_3^+ - Ala_2 - COOH$. Each line represents one conformation.

$NH^+ \cdots O{=}C$ H-bond. Conformation (2) appears as a consequence of the rotational motion over the C—C—N—C backbone dihedral, which leads to the disappearance of the H-bond between NH_3^+ and COOH. This rotational motion also leads to the appearance of a new H-bond between the N-terminal NH_3^+ and the neighboring backbone C=O as shown in conformation (3). Conformation (4) is obtained from conformation (3) by one proton transfer between the N-terminal NH_3^+ and the neighboring backbone C=O.

The bottom of Fig. 3 presents the 2D simplified graphs of the conformations, as extracted from graph theory (see Sec. II). This graph does not possess the 3D spatial representation of the molecular system, as codes like VMD[4] or Avogadro[33] would provide, but instead it gives a direct view of covalent bonds and H-bonds in a simplified way. For all the graphs presented in this paper, we use the following conventions. The covalent bonds are represented by *black* edges. The H-bonds are represented by arcs directed from the heavy atom (tail of the arc) to the acceptor atom (head of the arc). The arc is *red* as long as no proton transfer occurs; it becomes *blue* when there is a proton transfer. The electrostatic interactions are represented by *blue* edges (used for the clusters).

Figure 4 presents the time evolution of the dipeptide structure. The X-axis represents time and the Y-axis represents the conformations (from Fig. 3). The figure shows that conformation (4) (top line) appears only over one single period of time, while conformations (2) and (3) coexist overtime (the two middle lines). These two conformations basically differ by the dynamics of the N-terminal NH_3^+ (rotation) and of the neighboring C=O carbonyl group. These results are confirmed by analyzing the frequencies showed on the graph of transitions in the left side of Fig. 5. For example, there are 60 transitions observed between the two conformations (2) and (3), which is related to the abovementioned dynamics but also reflects the variations in distances around the threshold values used in our method. Figure 5 shows not only conformations explored along the trajectory (white circles) but also shows the transitional states (gray circles). In this trajectory two transitional states have been identified. Analyzing the graph, we observe forth-and-backward isomerization between conformations (1)

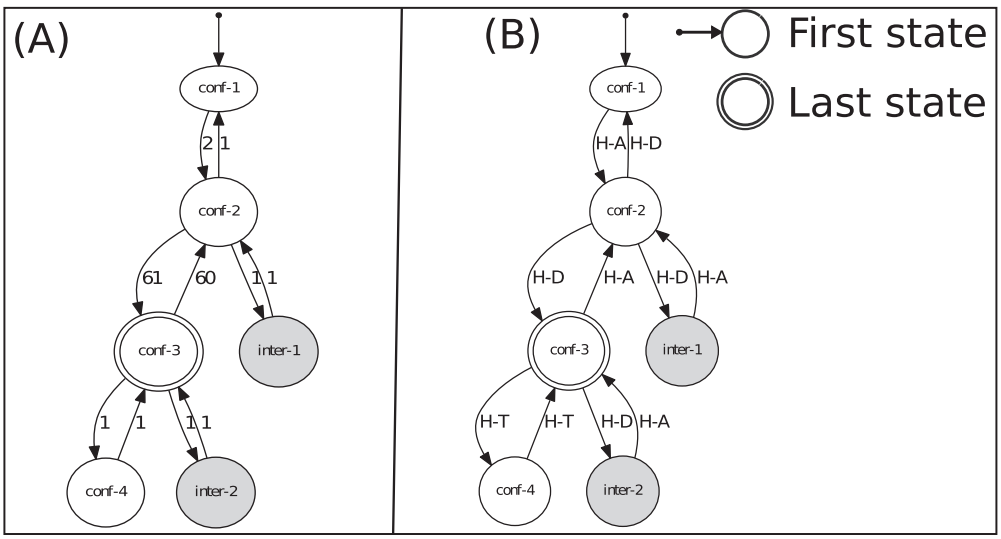

FIG. 5. Graph of transitions for $NH_3^+ - Ala_2 - COOH$ protonated dipeptide. (a) represents the graph of transitions with frequencies of occurrence and (b) represents the graph of transitions with the changes that occurred. Conformations are represented in white circles and transitional states in gray circles. See the text for definitions of labels.

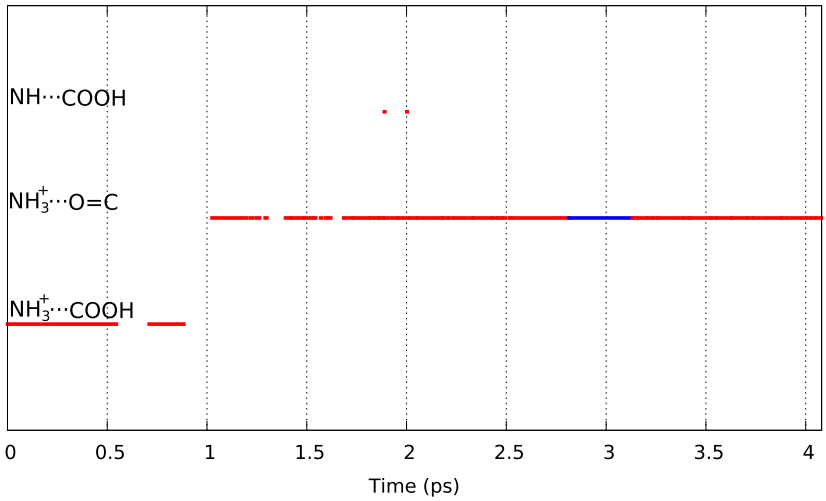

FIG. 6. Time evolution of the H-bonds formed in $NH_3^+ - Ala_2 - COOH$ along one trajectory. Each line represents one H-bond. The line is *red* when the H-bond is formed and it is *blue* when the H-bond is present but a proton transfer between donor and acceptor has occurred.

and (2), between conformations (2) and (3), and between conformations (3) and (4). The right side of Fig. 5 also provides the detailed changes occurring

- in going from conformation (1) to conformation (2), the appearance of one H-bond (H-A),
- from conformation (2) to conformation (3), the disappearance of one H-bond (H-D), and
- from conformation (3) to conformation (4), one proton transfer has occurred (H-T).

By analyzing the mean residence time for each observed conformation, conformation (3) is shown to be the most observed with a total residence time of 2.0 ps over 4.0 ps trajectory (50% over the dynamics).

Figure 6 presents the time evolution of the H-bond dynamics along the trajectory. The red line at the bottom shows the evolution of the H-bond formed between the N-terminal NH_3^+ and the C-terminal COOH: it is only observed within the first pico-second of the trajectory. Once this H-bond disappears, a new H-bond is formed between the N-terminal NH_3^+ and the backbone C=O (the middle line in Fig. 6). For this H-bond, there is furthermore proton transfer (the line in blue) occurring around 2.7 ps, which ends up at 3.2 ps when the proton goes back to its initial atom carrier.

The second analyzed trajectory concerns the Z-Ala_6—COOH peptide composed of 75 000 snapshots (30.0 ps of DFT-MD trajectory). The peptide is composed of 80 atoms, and the running time of the algorithm analysis was 20.87 s (see Table II). Comparing to the dipeptide trajectory, there are more H-bonds formed over time (seven H-bonds in total) for this slightly larger peptide. Figure 7 shows the time evolution of these H-bonds along the 30 ps trajectory, nicely showing co-existence in time of some of these H-bonds and disappearance/appearance of others over time.

Based on these H-bonds dynamics, three conformations have been identified by our graph theory method: conformation (1) has five strong H-bonds (the five red lines in the middle of Fig. 7) and has the C-terminal COOH that is located close to the backbone amide group NH, thus forming one NH···O=C H-bond (the bottom red line in Fig. 7). Conformation (2) appears as a consequence of the rotational motion over the C—C—N—C backbone dihedral, which leads to the disappearance of the H-bond between the C-terminal COOH and the backbone amide group NH and to the appearance of a new H-bond between this NH and a C=O backbone (the top red line in Fig. 7). A rotational motion is observed over another C—N—C—C backbone

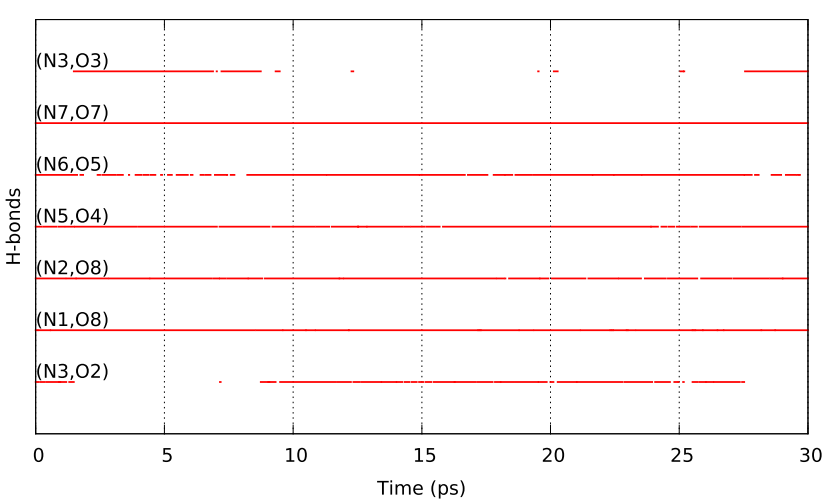

FIG. 7. Time evolution of the H-bonds formed in Z-Ala_6—COOH along one trajectory. Each line represents one H-bond. The line is *red* when the H-bond is formed. Labels represent the couple (heavy atom, acceptor) for each H-bond identified.

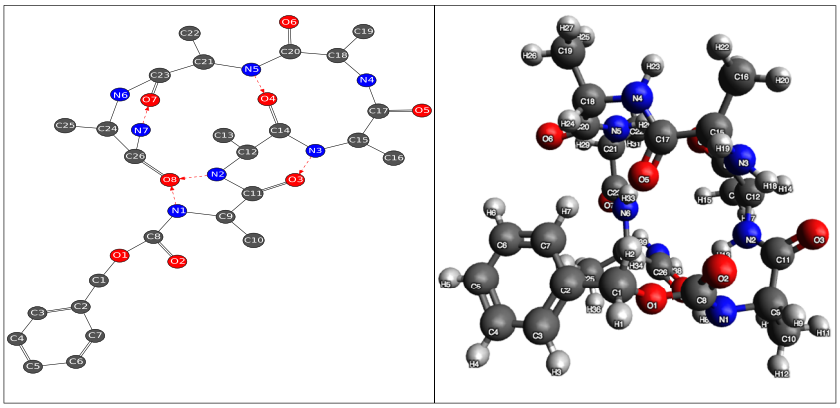

FIG. 8. Graph of one conformation of Z-Ala_6—COOH peptide. The left side of the figure represents the 2D simplified graph of the conformation, and the right side shows the 3D representation of the conformation. See the text for details.

dihedral; this leads to the disappearance of one H-bond between NH and C=O backbone groups [see the red line labeled (N6, O5)] and hence the appearance of conformation (3). Figure 8 shows the example of conformation (3) with the simplified 2D graph on the left side and the 3D structure on the right side.

The graph of transitions reported in Fig. 9 nicely illustrates the intricate H-bond conformational dynamics of the Z-Ala_6—COOH peptide and the transitional states (gray circles) that are sampled during these isomerisations.

With this series of examples taken from MD simulations of gas phase peptides for which conformational and isomeric dynamical events are observed over time, one can see that the use of graph theory approaches not only provides a direct and fast methodology to identify the various conformers and isomers explored during the dynamics but also allows to follow in time the interconversion between these conformers. While the easy automatic analysis of the trajectories provided by the graph algorithms developed here give all the time-dependent structural information contained in the trajectories, with the advantage of keeping the atomistic representation/knowledge of the trajectories, the graphs of transitions are the essential keys that provide the whole sequence of events occurring along the trajectories, at one glance, and without supplementary effort. As illustrated with this series of examples, graph theory becomes more and more relevant and useful as soon as the size and conformational flexibility and complexity of the molecular systems increase. Such algorithms can be readily applied to much larger and more complex systems, for instance in order to follow in time the folding of proteins, including the intermediate states sampled along the pathway. Furthermore, the graphs of transitions contain the general information on the potential energy surface, at least the part(s) sampled at the given temperatures of the dynamics. This includes general information on minima and transition states: coupled with minimization procedures, the graphs of transitions could hence immediately provide the more precise knowledge on these states.

B. Structural recognition in chemically reactive trajectories such as in collisional induced peptide fragmentations

Section IV A reported somehow on "easy analyses of H-bond dynamics." We now apply the same algorithm to

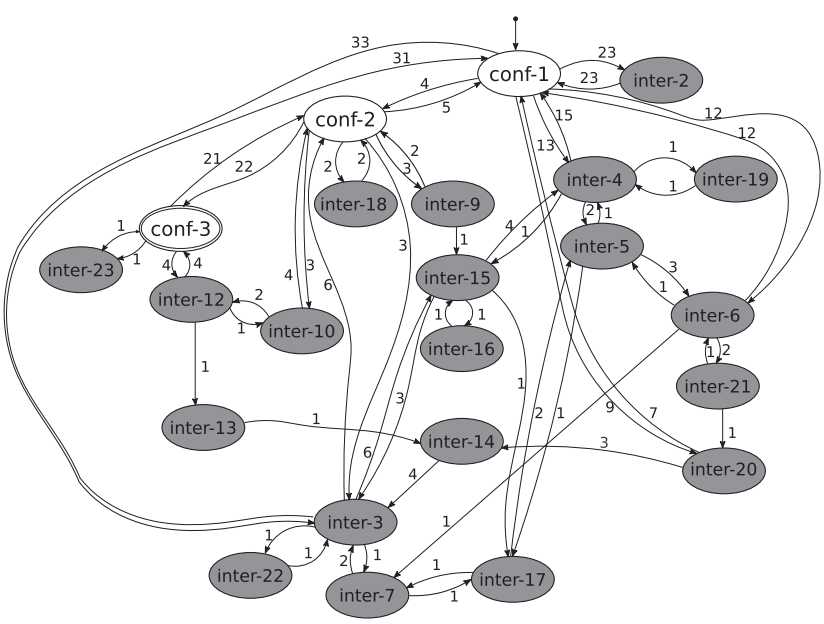

FIG. 9. Graph of transitions for Z-Ala_6—COOH protonated peptide trajectory. Conformations are represented in white circles and transitional states in gray circles. Labels represent the frequencies of transitions.

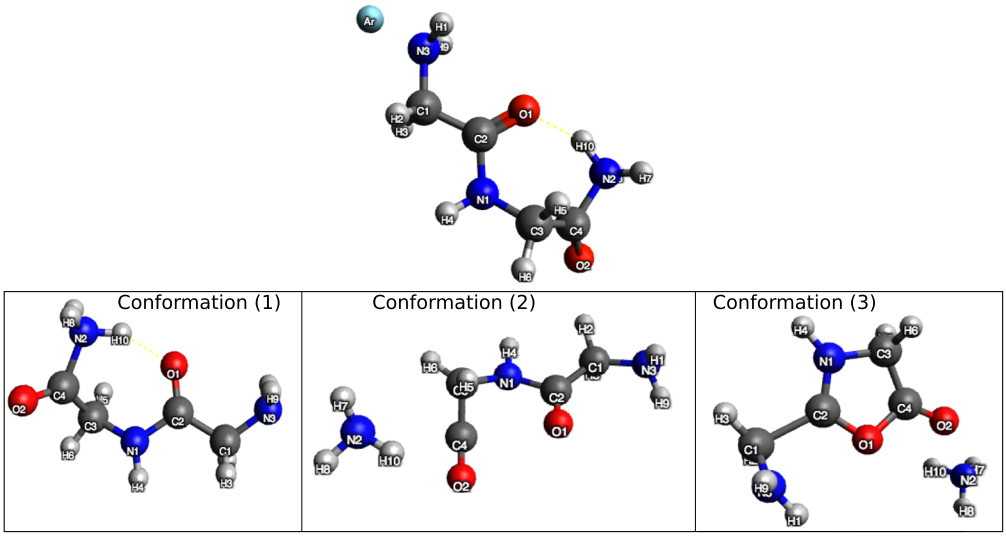

FIG. 10. Top of the figure shows a snapshot of one conformation before the collision of the argon atom with the protonated di-glycine gly2-NH_3^+ peptide. The bottom of the figure shows a schematic representation of the conformations of the fragments identified by the present graph theory algorithm along the dynamics of collision of the argon atom with the gly2-NH_3^+ peptide (argon has been removed from the bottom representations). Carbon atoms are represented in dark gray, nitrogen atoms in dark blue, oxygen atoms in red, hydrogen atoms in light gray, and the argon atom is in light blue.

chemical trajectories, where covalent bonds are broken and formed several times along the trajectory. To that end, we analyze CID (Collision Induced Dissociation) chemical dynamics trajectories, where one argon collides with a protonated glycine dipeptide (20 atoms). The CID results have been published in Ref. 20. For illustration, we analyze here a trajectory composed of 51 snapshots, representing an overall 5 ps. The running time of the analyses was 0.10 s (see Table II).

Three conformations were identified along the fragmentation pathway by our graph theory analysis. See Fig. 10 for the 3D representations. Conformation (1) is the initial structure of the peptide, which is seen in our graph analysis as one single fragment. Conformation (2) appears as a consequence of the collision between the argon atom and the peptide, leading to the breaking of the C—N covalent bond between the N-terminal NH_3^+ and the backbone C=O and hence leads to the appearance of two fragments. Conformation (3) results from the chemical reorganization of one fragment, with a new covalent bond being formed between the two backbone amide C=O groups.

Figure 11 presents the graph of transitions, showing how the conformations are related one to another and the time sequence. The graph on the right side of Fig. 11 provides, in particular, the sequence of changes occurring along the trajectory, starting with one fragment in conformation (1) (i.e., the initial peptide in that case), one covalent bond disappears (C-D) leading to conformation (2) which then evolves toward conformation (3) with one covalent bond formation (C-A), and one covalent bond breaking (C-D), without isomerising back. The frequency of occurrence of these events is reported in the left side of Fig. 11.

In chemical dynamics and CID trajectories, in particular, one has to launch a large number of collisional trajectories, sampling all possible collision impact parameters and all possible impacts of the argon onto the peptide. A thousand of trajectories could hence be generated, to be analyzed in terms of statistical relevance for the fragments to be generated. That is exactly what our method can do automatically. For example, by analyzing 40 trajectories for the $Ar—C_{15}H_{22}N_3O_4$ system, 22 conformations have been identified with a total of 31 different fragments being generated. Each trajectory analyzed contained 1000 snapshots and all of them have the same initial conformation of the peptide in terms of its covalent bonds. Based on covalent bonds dynamics analysis, we found that the most damageable collisions could lead up to 4 fragments.

This section has shown that the graph theory method proposed here allows to analyze multiple molecular trajectories in which fragmentation events occur over time (with several such events occurring per single trajectory, and with the added

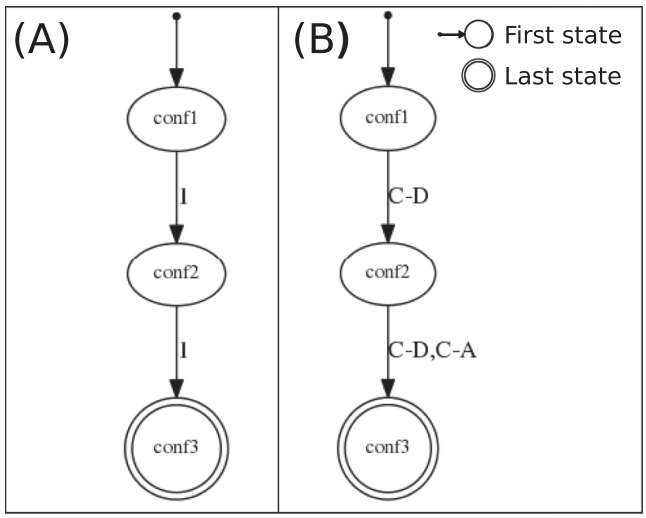

FIG. 11. Graph of transitions for the trajectory of the protonated di-glycine gly2-NH_3^+ peptide. (a) represents the graph of transitions with frequencies of occurrence of the different conformations and (b) represents the graph of transitions with the changes that occurred along the trajectory.

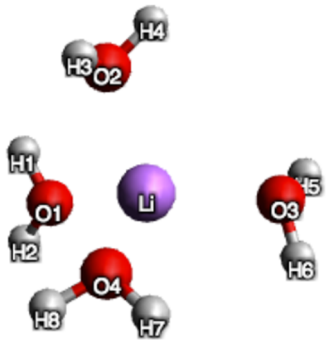

FIG. 12. Lowest energy conformation of the $Li^+(H_2O)_4$ cluster. Oxygen atoms are represented in red, hydrogen atoms in light gray, and the cation Li^+ is in light pink.

complexity that there are isomeric and chemical reorganisations of the molecular fragments over time), automatically and without effort. Graph theory has been shown to be essential to extract the information contained in the trajectories about the chemical reactions that occur, their time-sequence, to follow and analyze in time the chemical reaction mechanisms and subsequent chemical reorganisation of the molecular fragments, and automatically extract the statistical information of these processes. As shown here, the time scale analysis is rather low; the whole method can be applied in parallel (typically one trajectory analyzed per processor) in order to efficiently analyze multiple trajectories, giving us the confidence that such analyses can be performed routinely on any series of chemical trajectories of medium and much larger molecular systems. Again, the statistical analysis of all generated graphs to provide the final statistical overview of the chemical reactivity of a given molecular system is included automatically within the algorithms. We believe that our graph theory analysis of chemically reactive trajectories opens the path to a wide range of applications spanning numerous communities, not only the theoretical CID community but also, e.g., the catalysis community, the prebiotic and atmospheric communities, the heterogeneous community (see Sec. IV D for condensed phase applications, especially at interfaces relevant for heterogeneous catalysis at air-water and solid-water interfaces).

C. Structural recognition in trajectories of $Li^+(H_2O)_4$ clusters

Here, temperature effects on the isomerization of the $Li^+(H_2O)_4$ cluster are investigated [see Fig. 12 for the lowest energy conformer of $Li^+(H_2O)_4$]. See Ref. 19 for more details on the DFT-MD simulations performed in relation with vibrational spectroscopy. Our aim here is to show the capabilities of our algorithm in analyzing trajectories of clusters, where intermolecular interactions are the driving force. Here, they are electrostatic interactions in-between the Li^+ and water molecules as well as dispersion interactions in-between water molecules. Three trajectories have been analyzed, respectively, at 50 K, 300 K, and 400 K. Each trajectory contains 50 000 snapshots (20.0 ps). The total running time analysis for the three trajectories is 6.72 s.

Table I summarizes the number of conformations found for each trajectory, the coordination number of Li^+ (CN), and the number of H-bonds formed by the water molecules, all extracted from our graph theory analyses. The type of dynamics, i.e., in our present nomenclature, covalent bond dynamics, H-bond dynamics, and intermolecular electrostatic interactions dynamics related to the CN of the metal in $Li^+(H_2O)_4$, is also reported. The first comment from Table I is that temperature indeed influences the conformational dynamics of the $Li^+(H_2O)_4$ cluster. At the low temperature of 50 K, there is no conformational dynamics occurring (over the time scale of the dynamics), while at the higher temperatures of 300 K-400 K, there are H-bonds dynamics (i.e., dynamics of reorganization between the water molecules) and electrostatic interactions dynamics (i.e., dynamics of reorganization of the water molecules around Li^+ leading to a change in CN of the metal).

Figure 13 presents the graphs of transitions providing the conformational dynamics of the cluster as a function of temperature. At 50 K, there is no conformational dynamics, only one conformation of $Li^+(H_2O)_4$ is observed over all the trajectory. There is not enough energy within the system to overcome any energy barrier toward another conformer at this low temperature. At 300 K and 400 K, there is significant conformational dynamics as shown in the graphs (b) and (c) in Fig. 13. Labels on transitions in the graph (b) (300 K) show that there is a large amount of dynamics of reorganization between the water molecules, i.e., changes in H-bonds, the disappearance and appearance of H-bonds, but there is no change in the CN of the metal Li^+. Labels on transitions (arcs) in the graph (c) (400 K) show that both H-bonds dynamics and change in the CN of the metal have been observed along the trajectory. In addition, to go from one conformation to another, the cluster passes through transitional states (gray circles in Fig. 13) (300 K and 400 K).

Figure 14 presents the conformations explored at each temperature (50 K, 300 K, and 400 K) as analyzed by our graph theory algorithm. The left side of the figure shows the 3D representations of these conformations and on the right side their corresponding 2D simplified graphs that provide a direct

TABLE I. Summary of the conformational analyses for $Li^+(H_2O)_4$. See the text for details.

Temperature (K)	No. of Conformations	CN	No. of H-bonds	Dynamics found
50	1	2	3	None
300	1	3	0-2	H-bonds dynamics
400	2	4-3	0-2	H-bonds and electrostatic interactions dynamics

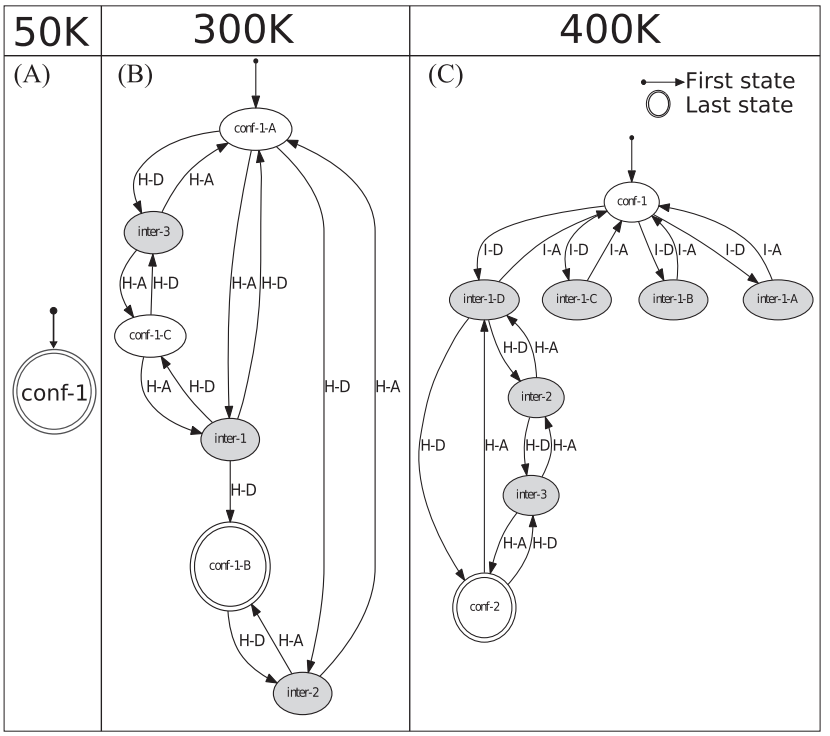

FIG. 13. Graph of transitions for the trajectory of the $Li^+(H_2O)_4$ cluster according to the temperature of the system. (a), (b), and (c) represent graphs at 50 K, 300 K, and 400 K, respectively. The graphs are presented with the detailed changes that occurred along the trajectories. Conformations are represented in white circles and transitional states in gray circles.

view of electrostatic interactions and H-bonds. In the conformation explored at the low temperature of 50 K, two strong electrostatic interactions are observed between Li^+ and water molecules (corresponding to a CN of 2) and three H-bonds are formed in-between water molecules. At 300 K, the Li^+ is surrounded by three water molecules (corresponding to a CN of 3) in its solvation shell along the trajectory. The remaining water molecule formed H-bonds with the other water molecules (see Fig. 14, conf-1-A, conf-1-B, and conf-1-C). Analyzing the evolution of H-bonds with time, we can observe that the remaining water molecule can form up to two H-bonds. These states are transitional states. The bottom of Fig. 14 shows now a trajectory where Li^+ is initially surrounded by four water molecules which do not form H-bonds (conf-1), giving rise to CN = 4; this conformer evolves toward conf-2 where one water molecule moves to the second shell of Li^+, hence decreasing the CN to 3.

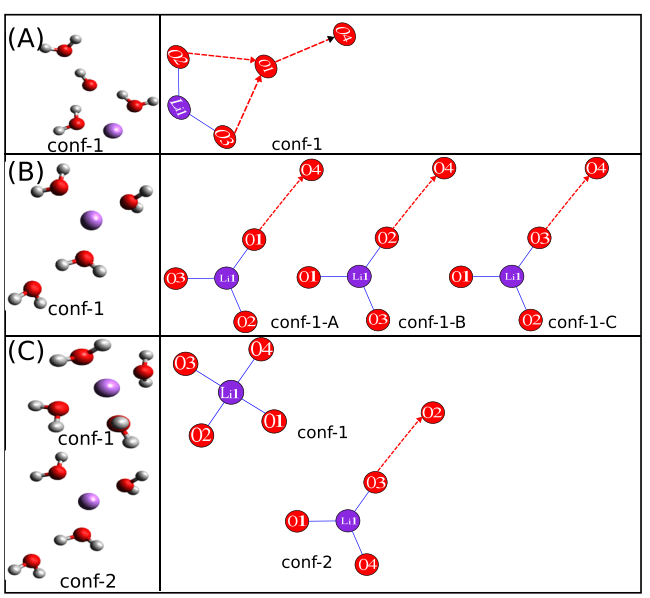

FIG. 14. Conformations identified along the $Li^+(H_2O)_4$ dynamics according to the temperature of the system. The left side shows the 3D representations of conformations. The right side represents 2D simplified graphs of the conformations. (a), (b), and (c) represent conformations identified at 50 K, 300 K, and 400 K, respectively.

This section has especially shown how easily transferable the graph theory algorithms developed in this work at the atomistic granularity level are. The application shown here on a very modest cluster size has already shown how easily one can get information on subtle structural changes that occur over time, not only from the point of view of the "central" metallic ion but also from the point of view of the surrounding water molecules. Once again, the statistical analysis of all events occurring along the trajectories gives a direct structural and chemical overview, ready to use at one glance. The graphs provide the information on the sequence of isomerisation that occur over time, providing direct information on the potential energy surface, that could be further used into a more systematic way to locate minima and transition states. The same algorithms can be readily applied to much larger molecular clusters, containing a much larger number of water molecules, several ions, and clusters made of other H-bonding solvents, without modifications.

D. Direct application to inhomogeneous condensed matter: The air/liquid water interface

The previous illustrations have shown the capabilities of our graph algorithms to analyze the conformational dynamics of gas phase molecules and clusters. The application of

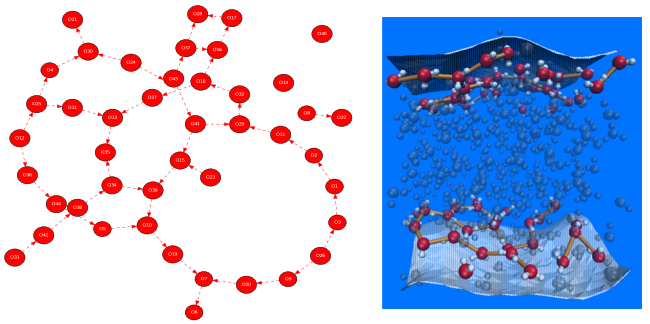

FIG. 15. Left: 2D simplified graph illustrating one conformation explored along the liquid water/air trajectory and showing the extended 2D H-bond network formed by the water molecules that belong to the 3.5 Å interfacial layer. Oxygen atoms are represented in red and hydrogen bonds are represented with the dashed red arcs. Right: side view of the air/liquid water interface highlighting the 2D-H-bond network [Reproduced with permission from Pezzotti *et al.*, J. Phys. Chem. Lett. **8**, 3133–3141 (2017). Copyright 1998 American Chemical Society].

the proposed algorithms is however not restricted to gas phase dynamics and can be applied to more complex and challenging systems such as condensed phase. We straightforwardly apply the whole method to the air/liquid water interface that is of high relevance in the interfacial science community. We have shown in a recent publication[21] that the liquid water at the interface with air is organized into an extended 2D network composed by water-water hydrogen bonds all oriented parallel to the surface. ~90% of the water molecules within the 3.5 Å thickness of the interface are involved into the 2D network. Beyond that thickness, liquid water is recovered. This 2D-H-bond network result has been obtained by a careful and rather painful analysis (and coding) of the H-bond network within the water molecules at the interface with air. We apply here our graph theoric algorithms in order to analyze this interfacial layer. For the illustration presented here, only 200 snapshots are extracted from a total of 20 ps *ab initio* MD trajectory, and only the interfacial region is analyzed (composed of an average of 44 water molecules, 132 atoms). The same 2D-H-bond network as in Ref. 21 has been obtained by the graph theory algorithms developed in the present work; an illustration is shown in Fig. 15 where one can see the extended H-bond network formed in between the water molecules within the 3.5 Å thin interfacial layer. One can immediately note that the water molecules in this layer form H-bonded rings/polygons, the graph theoric methods providing ~58% of the interfacial water molecules being involved in polygons over the whole trajectory analyzed here.

We furthermore analyzed the interfacial layer in terms of the size of the connected components on the mixed graph (i.e., a subgraph in which every vertex is connected to every other node in the sub-graph and is not connected to any other node not being part of the sub-graph) in order to obtain a statistical view on the number of water molecules involved within the extended 2D network. This is shown in Fig. 16, where one can see that the largest extended network is made by ~38-45 water molecules, while smaller components made of less than 10 water molecules can be found disconnected from the extended 2D network.

All graph analyses have been obtained in a few seconds (9.46 s).

As demonstrated using the air-water interface as a test case, the graph theory algorithms developed here are directly transferable to condensed phase systems, without modifications. The application to the air-water interface nicely shows that graph theory approaches are very efficient and provide an automatic identification of similarities/dissimilarities between conformations also when the molecular systems are highly dynamical, i.e., with H-bond networks made of H-bonds that are continuously broken and reformed on time scales of few fs to few ps. Furthermore, specific H-bonds patterns like the 2D-Network in Fig. 15 are easily revealed in the framework of graph theory. The identification and direct visualisation (see Fig. 15) of such highly interconnected H-bonded structures is inaccessible to standard structural methods, while it is automatically obtained with graph theory. This illustration on the air-water interface opens the path to automatically identify complex H-bond patterns and networks in liquids and at interfaces, i.e., the air-water interface, as shown here, solid-water interfaces, liquid-liquid interfaces, and for water at biological interfaces (e.g., around biomolecules or along channels and membranes). The inclusion of solutes/electrolytes, etc, in/at these condensed phase systems is of no consequence for the graph analyses developed here; their presence will indeed be automatically taken

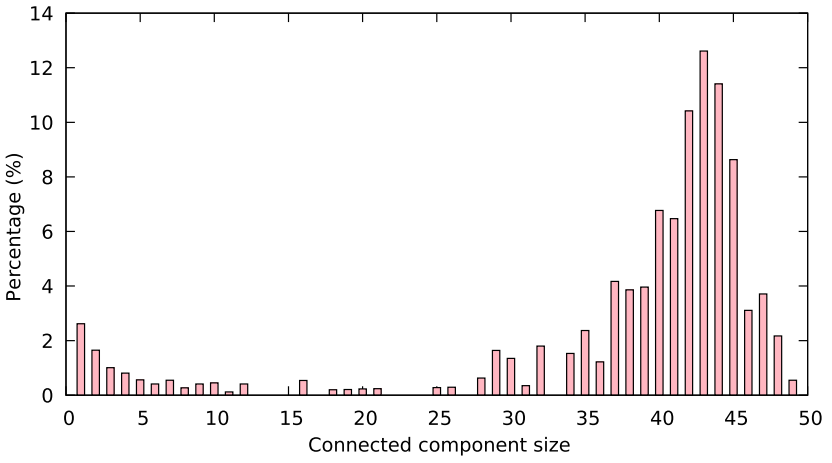

FIG. 16. Distribution of the connected components of the graph made by the water molecules at the air/liquid water interface.

TABLE II. Summary of the conformational analyses with the proposed graph theory and running times (on one processor of a Mac computer).

Chemical formula of the molecule[a]	Number of atoms	Number of snapshots in the trajectory	Number of reference snapshots	Number of conformations identified	Running time (s) (1 proc.)	References for the published trajectories
Trajectories of isolated protonated peptides: H-bonds dynamics						
$C_6H_{13}N_2O_3$ ($NH_3^+ - Ala_2 - COOH$)	24	5 900	17	3	0.52	17
		10 200	7	4	0.88	
		10 200	9	3	0.76	
		27 000	6	3	1.83	
$C_{26}H_{39}N_7O_8$ (Z-Ala_6—COOH)	80	75 000	4	3	20.87	...
Trajectories of collisional peptide fragmentation: H-bond and covalent bond dynamics						
$C_4H_{10}N_3O_2Ar$ (gly2-NH_3^+)	20	51	26	3	0.02	20
Trajectories of $Li^+(H_2O)_n$ clusters: H-bond and electrostatic interaction dynamics						
			2	1	2.29	19
$Li^+(H_2O)_4$	13	50 001	17	1	2.51	
			39	2	2.77	
Trajectories of air/liquid water interface: H-bond dynamics						
$(H_2O)_{38-45}$	114-135	200	200	200	9.46	21

[a]Some trajectories concern the same molecule but they are generated in different thermodynamic conditions.

into account and will be part of the final analyses, as can be typically anticipated from the cluster application reported in Sec. IV C.

E. Performance evaluation of the algorithm

Table II presents a summary of the conformational analyses that have been discussed in the previous examples together with running times for performance assessment. Table II shows, in particular, the very low computational cost of the algorithm. The results are obtained in a few seconds for these rather short DFT-MD/QM-MM (mixed quantum-classical) trajectories on rather small-size molecular systems. More importantly, the table also shows that the number of reference snapshots and the number of conformations explored along the trajectories are negligible when compared to the total number of snapshots generated in the trajectory. For example, along a trajectory of 27 000 snapshots (see the last line of the $C_6H_{13}N_2O_3$ trajectories in Table II), only 6 reference snapshots were identified, which means that orbits have been recomputed 6 times only, the same for the isomorphism tests. These results show that in practice, the theoretical complexity (i.e., non-polynomial complexity because of the isomorphism tests) of the algorithm is never reached. This is at least true for the rather simple conformational dynamics observed in these trajectories.

V. CONCLUSIONS

We have presented graph theoric algorithms for the automatic post-processing of molecular dynamics simulations, including graphs of transitions that give a statistical view of the structures sampled and the interconversion pathways in-between these structures (isomerisation and conformational dynamics) and of the chemical mechanisms and pathways along chemical reactions. We have chosen to maintain the atomistic granularity level of the simulations into the graph theoric representation and have hence devised low computational cost algorithms that allow us to follow in time the changes in the structures. Isomorphism is a key to identify conformations and provide the conformational change(s) by allowing the interchange of atoms of the same chemical nature within the atoms list (i.e., ignore the order of appearance in the snapshot). Despite being a non-polynomial problem and therefore possibly a computationally expensive method, we have shown that "orbits" and "reference snapshots" allow applying this test only a limited number of times along the trajectory (i.e., the theoretical complexity is never reached).

We have demonstrated that the proposed algorithms can be applied to follow in time the isomerisation dynamics of gas phase molecules (here peptides of various length, complexity, and flexibility), the isomerisation being solely due to changes over time in the hydrogen bonds formed and/or proton transfers, and that the same algorithms can be directly applied to analyze chemical dynamics where molecules undergo multiple chemical fragmentations and subsequent chemical reorganisation. As shown, graph theory not only provides a direct and fast methodology to identify and statistically analyze conformers and fragments generated over time but also provides the information about the interconversions in between conformers/isomers, the sequence of events, and their statistical occurrence, via the graphs of transitions. The latter are essential in giving the relevant microscopic statistical view of the events occurring along MD trajectories. They are automatically generated. The atomic level of granularity chosen for graph theory is instrumental in order to follow in time

hydrogen bond(s) and covalent bond(s) changes and reorganisations, where it is essential to keep the information on the chemical nature of the atoms within the molecules. Same algorithms have been used in order to assess the dynamics of clusters made of water molecules solvating a metal ion, both water-water H-bonds and coordination of the metal ion being followed over time. The same algorithms have been straightforwardly applied to condensed matter, here the air/liquid water interface. This particular illustration not only serves the purpose to show that the algorithms are transferable from one molecular system to another, i.e., from the gas phase to condensed matter without any modification, but it also serves the purpose to show that the graph theory here developed at the atomic level of granularity provides an easy means to follow in time rather complex hydrogen bond networks that are otherwise pretty complicated to analyze from the molecular dynamics trajectories since these H-bonded networks are highly dynamical. The graphs easily revealed the extended H-bond network formed by the water molecules at the interface with the air, and also immediately revealed the underlying structural organisation of the water molecules into H-bonded polygons, which sizes could be directly extracted from the graphs.

The graph theoric algorithm presented here can be applied to *ab initio* MD and classical MD simulations These graph theoric methods can furthermore be automatically applied to numerous trajectories analyzed in parallel and hence provide an immediate statistical view of the results.

To conclude, we believe that the present cheminformatics method based on graph theory algorithms for the post-processing of molecular dynamics simulations will have impact for the chemical-physics community as they provide the means to (easily) unravel new chemical physics. They offer unique opportunities to identify and statistically characterize molecular structures, conformational interconversions and chemical reactivity mechanisms, including unique opportunities to unravel complex intertwined extended H-bonding networks of short life-time in the condensed phase. We thus envision that these algorithms will be broadly adopted and will allow bringing forth new chemical physics, for typically, e.g., (heterogeneous) catalysis, atmospheric science, material design at (aqueous) interfaces, biomolecular recognition, etc.

ACKNOWLEDGMENTS

All co-authors acknowledge funding from the LABEX CHARM$_3$AT (LABoratoire d'EXcellence CHARM$_3$AT CHimie des ARchitectures Moléculaires Multifonctionnelles et des MATériaux). S.B. especially benefitted from a Ph.D. funding from LABEX CHARM$_3$AT. Daria Ruth Galimberti (LAMBE UMR8587, UEVE) is acknowledged for very helpful and fruitful discussions during the course of the work.

[1]D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Elsevier, 2001), Vol. 1.
[2]A. R. Leach, *Molecular Modelling: Principles and Applications* (Pearson Education, 2001).
[3]C. Dykstra, G. Frenking, K. Kim, and G. Scuseria, *Theory and Applications of Computational Chemistry: The First Forty Years* (Elsevier, 2011).
[4]W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," J. Mol. Graphics **14**, 33–38 (1996).
[5]N. Trinajstić, *Chemical Graph Theory*, 2nd ed. (CRC Press, 1992).
[6]J.-H. Choi, H. Lee, H. R. Choi, and M. Cho, "Graph theory and ion and molecular aggregation in aqueous solutions," Annu. Rev. Phys. Chem. **69**, 125–149 (2018).
[7]B. L. Mooney, L. R. Corrales, and A. E. Clark, "MoleculaRnetworks: An integrated graph theoretic and data mining tool to explore solvent organization in molecular simulation," J. Comput. Chem. **33**, 853–860 (2012).
[8]A. Ozkanlar and A. E. Clark, "ChemNetworks: A complex network analysis tool for chemical systems," J. Comput. Chem. **35**, 495–505 (2014).
[9]K. Han, R. M. Venable, A.-M. Bryant, C. J. Legacy, R. Shen, H. Li, B. Roux, A. Gericke, and R. W. Pastor, "Graph-theoretic analysis of monomethyl phosphate clustering in ionic solutions," J. Phys. Chem. B **122**, 1484–1494 (2018).
[10]C. M. Tenney and R. T. Cygan, "Analysis of molecular clusters in simulations of lithium-ion battery electrolytes," J. Phys. Chem. C. **117**, 24673–24684 (2013).
[11]F. Pietrucci and W. Andreoni, "Graph theory meets *ab initio* molecular dynamics: Atomic structures and transformations at the nanoscale," Phys. Rev. Lett. **107**, 085504 (2011).
[12]M. Hudelson, B. L. Mooney, and A. E. Clark, "Determining polyhedral arrangements of atoms using PageRank," J. Math. Chem. **50**, 2342–2350 (2012).
[13]B. L. Mooney, L. R. Corrales, and A. E. Clark, "Novel analysis of cation solvation using a graph theoretic approach," J. Phys. Chem. B. **116**, 4263–4275 (2012).
[14]F. Pietrucci and W. Andreoni, "Fate of a graphene flake: A new route toward fullerenes disclosed with *ab initio* simulations," J. Chem. Theory Comput. **10**, 913–917 (2014).
[15]E. Martínez-Núñez, "An automated transition state search using classical trajectories initialized at multiple minima," Phys. Chem. Chem. Phys. **17**, 14912–14921 (2015).
[16]E. Martínez-Núñez, "An automated method to find transition states using chemical dynamics simulations," J. Comput. Chem. **36**, 222–234 (2015).
[17]D. Marinica, G. Gregoire, C. Desfrancois, J. Schermann, D. Borgis, and M. Gaigeot, "*Ab initio* molecular dynamics of protonated dialanine and comparison to infrared multiphoton dissociation experiments," J. Phys. Chem. A. **110**, 8802–8810 (2006).
[18]A. Sediki, L. C. Snoek, and M.-P. Gaigeot, "N–H$^+$ vibrational anharmonicities directly revealed from DFT-based molecular dynamics simulations on the Ala$_7$H$^+$ protonated peptide," Int. J. Mass Spectrom. **308**, 281–288 (2011).
[19]V. Brites, A. Cimas, R. Spezia, N. Sieffert, J. M. Lisy, and M.-P. Gaigeot, "Stalking higher energy conformers on the potential energy surface of charged species," J. Chem. Theory Comput. **11**, 871–883 (2015).
[20]R. Spezia, J. Martens, J. Oomens, and K. Song, "Collision-induced dissociation pathways of protonated Gly$_2$NH$_2$ and Gly$_3$NH$_2$ in the short time-scale limit by chemical dynamics and ion spectroscopy," Int. J. Mass Spectrom. **388**, 40–52 (2015).
[21]S. Pezzotti, D. R. Galimberti, and M.-P. Gaigeot, "2D H-bond network as the topmost skin to the air–water interface," J. Phys. Chem. Lett. **8**, 3133–3141 (2017).
[22]E. M. Luks, "Isomorphism of graphs of bounded valence can be tested in polynomial time," J. Comput. Syst. Sci. **25**, 42–65 (1982).
[23]B. D. McKay, *Practical Graph Isomorphism* (Department of Computer Science, Vanderbilt University, Tennessee, USA, 1981).
[24]B. D. McKay and A. Piperno, "Practical graph isomorphism. II," J. Symbolic Comput. **60**, 94–112 (2014).
[25]S. G. Hartke and A. Radcliffe, "Mckay's canonical graph labeling algorithm," in *Communicating Mathematics* (American Mathematical Society, 2009), Chap. 8, pp. 99–111.
[26]S. Sorlin and C. Solnon, "A new filtering algorithm for the graph isomorphism problem," in *Proceedings of the Third International Workshop on Constraint Propagation and Implementation* (CPAI, 2006), p. 93.
[27]H. L. Bodlaender, "Polynomial algorithms for graph isomorphism and chromatic index on partial k-trees," J. Algorithms **11**, 631–643 (1990).
[28]P. T. Darga, K. A. Sakallah, and I. L. Markov, "Faster symmetry discovery using sparsity of symmetries," in *45th ACM/IEEE, Proceeding in Design Automation Conference, DAC 2008* (IEEE, 2008), pp. 149–154.
[29]T. Junttila and P. Kaski, "Engineering an efficient canonical labeling tool for large and sparse graphs," in *Proceedings of the Ninth Workshop on*

Algorithm Engineering and Experiments (ALENEX) (IEEE, 2007), pp. 135–149.

[30]D. Barth, S. Bougueroua, M.-P. Gaigeot, F. Quessette, R. Spezia, and S. Vial, "A new graph algorithm for the analysis of conformational dynamics of molecules," in *Proceeding in Information Sciences and Systems 2015* (Springer, 2016), pp. 319–326.

[31]J. Kobler, U. Schöning, and J. Torán, *The Graph Isomorphism Problem: Its Structural Complexity* (Springer Science & Business Media, 2012).

[32]L. Babai, A. Dawar, P. Schweitzer, and J. Torán, "The graph isomorphism problem (Dagstuhl seminar 15511)," Dagstuhl Reports **5**, 1–17 (2016).

[33]M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison, "Avogadro: An advanced semantic chemical editor, visualization, and analysis platform," J. Cheminf. **4**, 17 (2012).

[34]A. Cimas, T. Vaden, T. De Boer, L. Snoek, and M.-P. Gaigeot, "Vibrational spectra of small protonated peptides from finite temperature MD simulations and IRMPD spectroscopy," J. Chem. Theory Comput. **5**, 1068–1078 (2009).