Determination of kinetic properties in unimolecular dissociation of complex systems from graph-theory based analysis of an ensemble of reactive trajectories.

Ariel F. Perez-Mellor[1, 2, a)] and Riccardo Spezia[2, b)]
[1)]*LAMBE UMR8587, 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)]*Laboratoire de Chimie Théorique, Sorbonne Université and CNRS, F-75005 Paris,
France*

(Dated: August 10, 2021)

In this paper we report how graph-theory can be used to analyze an ensemble of independent molecular trajectories which can react during the simulation time-length and obtain structural and kinetic information. The method is totally general and here is applied to the prototypical case of gas phase fragmentation of protonated cyclo-di-glycine. The methodology allows to analyze the whole set of trajectories in an automatic computer-based way without the need of visual inspection, but getting all the needed information. In particular, we not only determine the appearance of different products and intermediates, but we can characterize the corresponding kinetics. The use of colored-graph and canonical labeling allows the correct characterization of the chemical species involved. In the present case, the simulations consist of an ensemble of unimolecular fragmentation trajectories at constant energy, such that from the rate constants at different energies the threshold energy can also be obtained for both global and specific pathways. This approach allows the characterization of ion-molecule complexes, likely through a roaming mechanism, by properly taking into account the elusive nature of such species. Finally, it is possible to obtain directly the theoretical mass spectrum of the fragmenting species if the reacting system is an ion, as in the specific example.

Keywords: Unimolecular Fragmentation — Rate Constant Calculations — Graph Theory — Peptides Frangmentation

I. INTRODUCTION

Using dynamical simulations to understand chemical reactivity represents a wide and important field in physical chemistry.[1–5] More recently, thanks to the improved computer power and new developments in electronic structure theory, it becomes possible to follow on-the-fly bond breaking/making processes also for relatively complex molecules, like peptides, sugars or polycyclic aromatic hydrocarbons.[6–8]

While fitted potential energy surfaces are often used to understand in detail reactions of relatively small systems,[9–13] the coupling between dynamics and electronic structure theory made it possible to study the evolution on-the-fly of relatively large systems. For example, uni- and bi-molecular reaction dynamics simulations were largely studied with the aim of understanding several processes, from combustion to atmospheric chemistry, from interfacial reactions to photolysis or astrochemistry.[14–20] Unimolecular reactivity represents an important class of reactions,[21] with a relevant application in the field of mass spectrometry.[22]

Recently, trajectories-based methods were used to understand several gas-phase ion chemistry experiments, like electron ionization mass spectrometry,[23–25] surface-induced dissociation (SID)[6,26,27] or collision-induced dissociation.[28,29] In particular, in the field of tandem mass spectrometry, simulations were able to predict fragmentation products, propose reaction mechanisms and quantify reaction energy thresholds for both statistical and non-statistical processes.[7] Statistical unimolecular fragmentation kinetics of relatively large systems was obtained by activating the initial system with excess vibrational energy. By following the initial product's decay as a function of time, it was possible to obtain unimolecular rate constants and, from Arrhenius-like plots, the corresponding activation energies.[16,30–32]

On-the-fly simulations are based on the time propagation of atomic positions expressed in cartesian coordinates. Thus, from the computational point of view, a molecule is a set of atoms connected by chemical bonds. However, since the system is represented by interacting nuclei and electrons, the bonds are not directly defined. One can state the existence and evolution of such bonds by geometrical and/or electron density criteria. When systems are simple and trajectories relatively few, the outcome of simulations can be analyzed either by *ad hoc* quantifications (distances, angles, etc ...) or by visual inspection of the resulting trajectories. However, thanks to computational improvements over the last years, it is now possible to run several and long trajectories of relatively big systems for which different and complex behaviors are possible. Of course, it is impossible to predict in advance all the possible reaction pathways only based on chemical intuition. The results, however, consist of a huge amount of data from which advanced analysis tools, like graph-theory, are needed to obtain physical properties. For this reason, graph-theory-based methods were used and developed to better understand the behavior of complex molecular systems

[a)] ariel.perezmellor@unige.ch; Present address: Department of Physical Chemistry, University of Geneva, 30 Quai Ernest-Ansermet, 1211 Geneva 4, Switzerland.
[b)] riccardo.spezia@sorbonne-universite.fr

from a relatively important amount of data. Examples can be found in protein folding[33], enzyme kinetics[34], conformational analysis[35], protein structure and flexibility identification[36,37], peptide structure and kinetics studies,[38] or assignment of vibrational normal modes.[39]

Molecular dynamics simulations are widely used in different fields, and they provide a huge amount of data from which important molecular properties should be obtained. To this end, several authors used graph theory to determine the different structures sampled during long simulations. [40–44] Pietrucci and Andreoni developed another useful application to drive biased molecular dynamics simulations from reactants to products without imposing elaborated geometrical constraints. [45,46] Recently, an original way to reduce computational time when analyzing a large set of data issued from molecular simulations using graph theory was proposed by Bougueroua *et al.*[47] Martinez-Nunez and co-workers proposed a specific application to chemical reactivity.[48–50] It is based on molecular dynamics in which the graph theory algorithms automatically determine minima and saddle points along a reaction path. This method was applied to photo-fragmentation,[51] unimolecular fragmentation,[52] and organometallic catalysis.[53]

We recently used graph theory to automatically detect if a molecular system dissociates after internal energy activation and obtain, from an ensemble of quasi-classical trajectories, the distribution of molecular fragments.[54] From this information, an *in silico* mass spectrum can be obtained as well as unimolecular rate constants.[30,55] However, different steps were determined *ad hoc* or not fully automatized.

In the present work, we show how graph-theory can be developed to obtain the structural and kinetic information needed to describe the reactivity of complex molecular systems as obtained from an ensemble of trajectories, here focusing in particular on gas phase unimolecular fragmentation. In particular, we consider the possibility that an initial structure passes through intermediates before dissociating. One important point is to correctly consider the possibility that the system oscillates between intermediate(s) and reactant before it eventually dissociates. In mass spectrometry, it is well known that ion-molecule complexes can be formed before fragmentation.[56,57] A recent study of L-cysteine fragmentation[52] has shown that this can correspond to the dynamical mechanisms called roaming.[58,59] The accurate detection of such ion-molecule complexes' formation can be problematic due to their elusive dynamical nature. The specific analysis of trajectories to identify such states is discussed here.

The methodology is explained in general forms and applied to the prototypical case of fragmentation of protonated cyclo-di-glycine, for which experimental mass spectra are available and fragments known.[60] It will serve as an example to show how we can obtain from graph-theory structural and kinetic information of fragmentation of a large system. In particular, peptides fragmentation represents a typical example in which the correct identification of intermediate structures is crucial to explain reactivity: after activation, the excess proton often moves to different basic sites, weakening the adjacent bonds and inducing fragmentation. This is the so-called "mobile proton model," which was shown to be a general mechanism typical of peptide fragmentation[61–63] but also relevant in other compounds, like nucleic acids.[55,64] In recent works, we have successfully applied some aspects of the methodology proposed here to qualitatively disentangle the complex dynamics in CID experiments of dipeptides that contain the DKP motif, such as cyclo Tyr-ProH$^+$[65] and cyclo Phe-HisH$^+$.[66]

This work illustrates how graph theory can be applied to quantitatively describe such complex unimolecular processes, where the protonated cyclo-di-glycine is used as a test case. In setion II, we first explain how chemical dynamics simulations are performed: they provide this specific example's trajectories, but the methodology is more general and does not depend on how simulations are performed. In section III, we show how graph theory formalism can be adapted to analyze an ensemble of trajectories. We have then applied it to a simple three-state kinetic model (section IV) from which overall properties can be obtained. In section V, we show how the model can be extended to analyze in more details one specific pathway. Finally, we demonstrate how the method can be used to obtain a mass spectrum. Section VII concludes the article.

II. CHEMICAL DYNAMICS SIMULATIONS

Two isomers of cyclo-di-glycine were used as initial structures for the dynamical simulations: the most stable cyclic structure (labeled CYC^{00}) and a linear isomer (LIN^{00}) which is higher in energy and largely populated. They are shown in Figure 1.

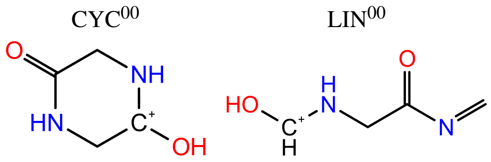

Figure 1. Schematic representation of the two structures used in chemical dynamics simulations. CYC is for the cyclo-di-glycine and LIN for its linear isomer. Both are protonated forms. The label 00 means that they are the most stable structures. The charge is arbitrarily placed on the C atoms, another valence structure can be drawn with the charge on the O atoms and C=O double bonds. The simulations are done in the molecular orbital framework so this will not have any impact.

Trajectories were propagated on-the-fly using the semi-empirical Hamiltonian RM1-D[67,68] to obtain energies and gradients of the two systems. Note that this method was shown to correctly characterize dissociation properties of protonated peptides in the gas phase[31,69,70] and includes dispersion corrections.

To induce unimolecular fragmentations, the systems were activated via excess vibrational energy through microcanonical normal mode sampling.[71] We used several internal energy values in the 131-217 kcal/mol range, listed in Table I (labeled as E_v). Rotational energy was incorporated via a

Boltzmann distribution at 300 K. For each system and total energy, an ensemble of trajectories (about 10 000 per point) was run, integrating Newton's equations of motion numerically via the velocity Verlet algorithm with a time step of 0.1 fs, ensuring energy conservation. Each trajectory was propagated up to 20 or 40 ps depending on the energy. Details on simulation conditions are summarized in Table I, where Λ is the simulation time of each trajectory. In the same table we also report the average potential ($\langle V\rangle$) and kinetic ($\langle K\rangle$) energies (with corresponding standard deviations) of the activated structures obtained from simulations with the energy decomposition method discussed in section IV B where they will also be discussed. The zero of the potential energy was set at the electronic energy of CYC^{00}.

	Λ	N	E_T	E_v	$\langle K\rangle$	$\sigma_{\langle K\rangle}$	$\langle V\rangle$	$\sigma_{\langle V\rangle}$
	ps		kcal/mol					
CYC^{00}	40	10000	158	157	78	13	80	13
	40	9982	168	167	83	14	85	14
	40	9980	178	177	88	14	90	14
	20	11886	188	187	93	15	95	15
	20	9856	198	197	97	16	101	16
	20	11455	208	207	102	17	106	17
	20	9121	218	217	107	18	111	18
LIN^{00}	40	9651	158	131	66	11	92	11
	40	9387	168	141	71	12	97	12
	40	8930	178	151	76	12	102	12
	20	7630	188	161	81	13	107	13
	20	13792	198	171	86	14	112	14
	20	11140	208	181	91	15	117	15
	20	9026	218	191	96	16	122	16

Table I. Simulation details. Λ is the total integration time, N the number of trajectories, E_T, the total energy, E_v the initial excess vibrational energy, $\langle K\rangle$ the average kinetic energy and $\langle V\rangle$ the average potential energy, where $\sigma_{\langle K\rangle}$ and $\sigma_{\langle V\rangle}$ are the corresponding standard deviations.

All simulations were performed with the general chemical dynamics software Venus[72] which is interfaced with MOPAC 5.022mn[73] for electronic structure calculations.

III. GRAPH THEORY ANALYSIS OF AN ENSEMBLE OF REACTIVE TRAJECTORIES

Chemical dynamics trajectories provide a huge amount of data in terms of positions (and momenta) of the different atoms as a function of time. We now define the key quantities we can obtain by applying graph theory to this set of data. These quantities will be used to obtain the different physical quantities. In the present application we will focus in particular on what is relevant to obtain rate constants and (time dependent) theoretical mass spectrum.

A. Data storage

In all trajectories, the information is saved every τ steps. In the present case we have chosen $\tau = 50$ fs, and this will depend on the characteristic time-scale of the process of interest and on the storage capabilities. As a result of this time discretization, several 2D-arrays are constructed to store and analyze the information appropriately.

In the following, $\mathbf{X}[i,j]$ denotes a generic 2D-array where i indicates the trajectory number ($i \in [1,N]$) and j the snapshot ($j \in [1,M]$ with $M = \Lambda/\tau + 1$). We have, for example, the atomic positions $\mathbf{XYZ}[i,j]$, the potential $\mathbf{POT}[i,j]$, the kinetic $\mathbf{KIN}[i,j]$, and total energy of the system $\mathbf{TOE}[i,j]$.

In general, two kinds of arrays $\mathbf{X}[i,j]$ can be distinguished according to the type of variable it might contain: continuous or discrete. $\mathbf{X}[i,j]$ is said to be a continuous (discrete) 2D-array if and only if (iff) the elements X_{ij} are continuous (discrete). A continuous array can be transformed into a discrete array through a discretization process, which will consist of introducing threshold values. A typical example used in this work is transforming the cartesian $\mathbf{XYZ}[i,j]$ array into the adjacency $\mathbf{ADJ}[i,j]$ array, as explained in the next subsection.

B. Building undirected and unweighted κ-Coloured Graph

A critical aspect of the unimolecular fragmentation simulation is to identify whether an isomerization or fragmentation event occurs from a computer analysis of time evolution of cartesian coordinates without any visual inspection. At this end we used different tools of graph theory.[74]

The analysis of a set of N-trajectories begins with the treatment of the aforementioned $\mathbf{XYZ}[i,j]$-array. It is a 2D-array where each XYZ_{ij} element contains the position of the atoms for each trajectory i at the discretized time $(j-1)\,\tau$.

The geometry contained in XYZ_{ij} can be converted into a simple undirected and unweighted κ-colored graph $G = (V,E)$ by following a discretization criterion that establishes whether two atoms are linked or not. The finite set of vertices $V = \{v_1, \ldots, v_\lambda\}$ denotes all the λ-atoms while the finite set $E \subseteq V \times V$ of edges (v_n, v_m) with $m,n \in [1,\lambda]$ indicates all the possible connections between them. Each vertex v_n is colored with a suitable color $C = \{1,2,\ldots\kappa\}$ according to the type of atom it represents. Therefore, the maximum number of different types of chemical atoms in the molecule defines the color number κ of G. In addition, the coloring of the vertices $\pi \in V \times C$ is known as a partition of the nodes. These concepts are directly taken from standard graph theory.[74]

A practical way of representing a graph G is through its adjacency matrix $A \in R^{\lambda\times\lambda}$ where $\lambda = |V|$, the number of vertices corresponding to the number of atoms. It is a symmetric $(0,1)$-matrix in which the diagonal elements are all zeros. The rest of the elements A_{nm} are one iff exists an edge between vertex v_n and v_m, otherwise they are zero. To build the A matrix for a given XYZ_{ij} molecular geometry we define the threshold bond distances as the sum of the atomic radii of the two atoms $R_n + R_m$ involved in the bond, multiplied by a factor α:

$$A_{nm} = \begin{cases} 1 \rightarrow |\vec{r}_n - \vec{r}_m| \leq \alpha \cdot (R_n + R_m) \\ 0 \rightarrow |\vec{r}_n - \vec{r}_m| > \alpha \cdot (R_n + R_m) \end{cases}$$

The α parameter allows for bond elongation during the molecular dynamics, i.e., vibrations around stables geometries. It means that all the geometries around stable isomers have the same $G(V,E)$. The optimum α value of 1.264 was determined in previous studies[30,54] since it is able to provide reasonable chemical structures avoiding unphysical atomic valences. For the atomic radii, we used previously reported values[54], and notably:

	H	C	N	O
R (Å)	0.372	0.781	0.807	0.715

The α value remains constant in all the systems studied in this work. To overcome the dependency of α with vibrational energy, some corrections were introduced during the definition of the states, as will be detailed in the section IV A.

Thus, for each XYZ_{ij}, there exists a simple undirected κ-coloured graph that can be represented by its adjacency matrix ADJ_{ij}. The partition is built with the set of atom types respecting the order of appearance in the XYZ file; in our case a possible representation is thus $\pi = \{HHHHHHHCCCCNNOO\}$. This representation was chosen to address events such as isomerization and reactivity specifically. Simultaneously, structures that differ, for example, by the rotation of one or more atoms or groups, will not be considered different. Shape descriptors such as inertia moments or radius of gyration could be used to discriminate between rotamers, for example, but it is out of the scope in this work which is devoted on describing chemical reactivity and thus not considered presently.

In Figure 2, we show a snapshot extracted from the simulations to illustrate all the steps of the methodology developed here. Table II displays the adjacency matrix of the structure sketched in Figure 2 constructed following the steps mentioned above.

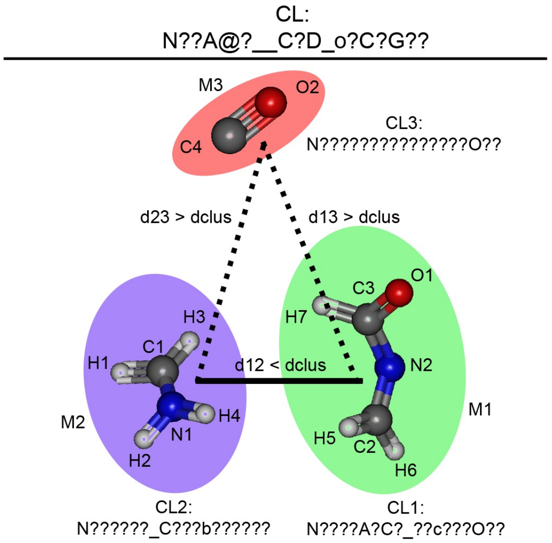

Figure 2. A snapshot taken from the simulations used to illustrate the graph decomposition method used in the text. CL stands for the canonical label, while dclus represent the cut-off distance used to identify the cluster formation; see text for details.

	H1	H2	H3	H4	H5	H6	H7	C1	C2	C3	C4	N1	N2	O1	O2
H1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
H2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
H3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
H4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
H5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
H6	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
H7	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
C1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
C2	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0
C3	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0
C4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
N1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0
N2	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
O1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
O2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

Table II. An adjacency matrix of the snapshot depicted in Figure 2. The bonds were colored accordingly to the fragment to which they belong.

C. Derived Information From ADJ [i,j]-array

The analysis of the 2D-array **ADJ** [i,j] through graph-theory tools is a key point of this study. In the following we detail some useful 2D-arrays that can be obtained from **ADJ** [i,j].

- **ADJ** [i,j] + π → **CAN** [i,j]. Having defined the partition π of the graph ADJ_{ij} and by using a graph isomorphism algorithm, one can get the canonical label, CAN_{ij}, of ADJ_{ij}. Therefore, each CAN_{ij}-element is the canonical label of its corresponding ADJ_{ij}. The labelling process was carried out by the use of the *amtog* and *labelg* tools included in the NAUTY package[75]. The introduction of this graph invariant allows to overcome the isomorphism problem which otherwise will occur in the automatic analysis of the trajectories. Therefore, one can count precisely the amount of different graphs that are formed during the dynamics. This precise counting is impossible using only the adjacency matrix formalism. The permutation of two rows or two columns of the same atom type inherently implies a change in the adjacency matrix. At the same time, the resulting structure is chemically identical to that of the precursor. A typical example is when two hydrogen atoms exchange themselves. In other terms, canonical labeling is an efficient way to identify if two graphs are identical or not. Two graphs are said to be identical iff they share the same canonical label; otherwise, they are different. On the top of Figure 2 is represented the canonical label (CL) of the adjacency matrix of the entire snapshot (Table II) by using the partition introduced above. The Graph6 format was used for the canonical representation.[75]

- $\mathbf{ADJ}\,[\mathrm{i,j}] \rightarrow \mathbf{CLU}\,[\mathrm{i,j}]$. Fruitful information can be extracted by taking into account the connectivity of the graph. The graph $\mathrm{ADJ_{ij}}$ can be classified as connected or disconnected. A graph is said to be connected when every pair of vertices in the graph is connected, it means that there is a path between $\mathrm{v_n}$ and $\mathrm{v_m}$. Chemically speaking, this corresponds to an undissociated molecule. Otherwise, they are called disconnected, which corresponds to two or more molecules. Moreover, every disconnected graph is made by a set of connected subgraphs. By identifying and counting all the connected subgraphs in $\mathrm{ADJ_{ij}}$, using the depth-first search (DFS) algorithm[76], it is possible to decompose $\mathrm{ADJ_{ij}}$ into all the connected entities, such that,

$$\mathrm{ADJ_{ij}} = \sum_{\mathrm{k=1}}^{\mathrm{D}} \mathrm{ADJ_{ij}^{k}} \tag{1}$$

where, D is the number of disconnected components of the associated graph $\mathrm{ADJ_{ij}}$. An example of this decomposition is given in Table II, where the full matrix can be seen as the sum of the matrices of the three resulting fragments, each labeled with a different color (red, blue and green). Then, the canonical label of every $\mathrm{ADJ_{ij}^{k}}$ is calculated by using the same partition π previously defined (labeled CL1, CL2 and CL3 in Figure 2). The latter allows an adequate count of the different fragments.

Having determined the atoms that belong to a given fragment, we calculate (in the case of a disconnected graph) the centers of mass (COM) of every connected subgraph and the distances between them (labeled d12, d23, and d13 in Figure 2). This is an important information for determining the formation of ion-molecule complexes as explained further below.

All this information is stored in the $\mathbf{CLU}\,[\mathrm{i,j}]$ array.

- $\mathbf{ADJ}\,[\mathrm{i,j}] \rightarrow \mathbf{CLU}\,[\mathrm{i,j}] + \mathrm{M} \rightarrow \mathbf{MAS}\,[\mathrm{i,j}]$. Once all connected subgraphs are identified in $\mathrm{ADJ_{ij}}$, their corresponding vertices are sum-weighted by the atom mass. Then, each $\mathrm{MAS_{ij}}$-element corresponds to the set of mass recognized in $\mathrm{ADJ_{ij}}$. For example, $\mathrm{MAS_{ij}} = \{\mathrm{M1,M2,M3}\}$ where $\mathrm{M1} > \mathrm{M2} > \mathrm{M3}$ if three different connected subgraphs were detected in $\mathrm{ADJ_{ij}}$, as in Figure 2. In the present work we used the following masses:

	H	C	N	O
M (uma)	1.00782	12.00000	14.00307	15.99490

which correspond to what used in the dynamics.

- $\mathbf{ADJ}\,[\mathrm{i,j}] \rightarrow \mathbf{CLU}\,[\mathrm{i,j}] + \mathrm{M} \rightarrow \mathbf{MAS}\,[\mathrm{i,j}] \rightarrow \mathbf{CMAS}\,[\mathrm{i,j}]$. Similar to the α parameter, we introduce a cut-off distance (dclus) between the COM to determine whether two connected subgraphs form a complex or not. Then, by inspecting the corresponding adjacency matrix $\mathrm{B} \in \mathrm{R^{D\times D}}$, we can detect such events. The B matrix is defined as,

$$\mathrm{B_{pq}} = \begin{cases} 1 \rightarrow \mid \vec{\mathrm{R}}_{\mathrm{p}}^{\mathrm{CM}} - \vec{\mathrm{R}}_{\mathrm{q}}^{\mathrm{CM}} \mid \leq \mathrm{dclus} \\ \\ 0 \rightarrow \mid \vec{\mathrm{R}}_{\mathrm{p}}^{\mathrm{CM}} - \vec{\mathrm{R}}_{\mathrm{q}}^{\mathrm{CM}} \mid > \mathrm{dclus} \end{cases}$$

where p and q run from one up to the number of disconnected graphs identified in the snapshot. Once the formation of an ion-complex is detected, their masses are calculated and stored in **CMAS**. As a consequence, each $\mathrm{CMAS_{ij}}$-element corresponds to the set of cluster mass recognized in $\mathrm{ADJ_{ij}}$. For example, $\mathrm{CMAS_{ij}} = \{\mathrm{M1+M2,M3}\}$ if a complex is formed between M1 and M2 as in the example of Figure 2.

This approach of using only the distance dclus as a criterion to determine whether two fragments form a complex or not could be not suitable for larger systems assuming extended configurations (as can occur in peptides and more in general in large polymers). This is due to the fact that the distance between the outer atom in the fragment and the COM can be considerably large. A possible solution can be to divide the whole molecule in sub-units and then perform the same analysi with respect to the COM of each sub-unit. In the case of (bio)-polymers, one can use as sub-units the different monomers which are normally relatively small and compact.

The dclus parameter was obtained from the analysis of trajectories which form disconnected graphs, and in particular the $\mathrm{CYC^{00}}$ simulations at 177 kcal/mol and $\mathrm{LIN^{00}}$ at 151 kcal/mol. For the two sets we calculate the distances between the COM of the different disconnected graphs obtained from which we get a pair distribution function of fragment distances. The normalized distribution functions (PDF) are shown in Figure 3. The two PDFs show a maximum at around 3.75 Å which corresponds mainly to the transient separation of hydrogen atoms from the main structure as detected by the algorithm (this does not correspond to a physical phenomenon of dissociation, but to a hot vibrational state), and sketched in Figure 3. Thus, this 3.57 Å distance can be seen as an effective radius of all the connected graphs, and it can serve as a reference to establish the cut-off for ion-molecule complex detection, which must be larger than it. A cut-off of 8 Å is more than twice this value and, as shown in Figure 3 can represent a good estimation of a distance to automatically identify ion-molecule complexes from the analysis of an ensemble of trajectories.

- $\mathbf{ADJ}\,[\mathrm{i,j}] \rightarrow \mathbf{CLU}\,[\mathrm{i,j}] \rightarrow \mathbf{CYC}\,[\mathrm{i,j}]$. The cyclomatic number of the graphs can give the number of rings present in the structure, $\mathrm{c} = |\mathrm{E}| - |\mathrm{V}| + \mathrm{D}$ where $|\mathrm{E}|$, $|\mathrm{V}|$ and D are the number of edges, the number of vertices and the number of disconnected components

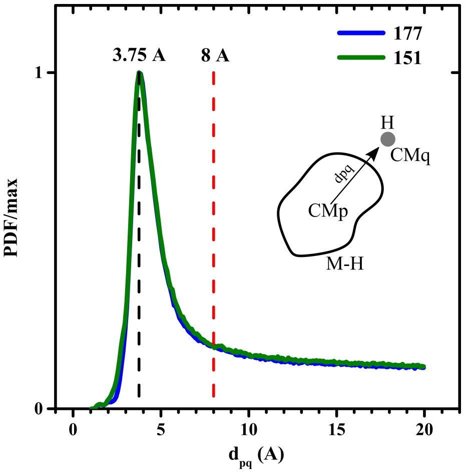

Figure 3. The PDF of the distance between disconnected graphs normalized at its maximum value as obtained from two simulations: CYC^{00} with an internal energy of 177 kcal/mol (blue line) and LIN^{00} at 151 kcal/mol (green line). The two vertical lines highlight the maximum of the PDF and the dclus value used. The sketch summarizes the origin of the major contributors to the peak of the PDF.

of the associated graph, respectively. The number of edges can be computed directly from the adjacency matrix, such that:

$$|E| = \frac{1}{2}\sum_{n=1}^{|V|}\sum_{m=1}^{|V|} A_{nm} \qquad (2)$$

while the value of D was already obtained in $\mathbf{CLU}[i,j]$. Therefore, the CYC_{ij}-element account for the number of cycles contained in the adjacency matrix ADJ_{ij}. For example, if $c = 1$ the graph (and so the corresponding molecule) has one ring.

Information	Matrix
The position of the atoms	XYZ_{ij}
The potential energy of the graph	POT_{ij}
The kinetic energy of the graph	KIN_{ij}
The adjacency matrix of the graph	ADJ_{ij}
The total energy of the graph	TOE_{ij}
The partition of the graph	π
The canonical label of the graph	CAN_{ij}
The number of connected subgraphs in the graph The adjacency matrices of the subgraphs The canonical labels of the subgraphs The distances between connected subgraphs	CLU_{ij}
The masses of the connected subgraphs	MAS_{ij}
The masses of the clusters	$CMAS_{ij}$
The number of cycles contained in the graph	CYC_{ij}

Table III. Summary of all the information derived and stored from the present analysis conducted on the ensemble of trajectories.

Table III summarizes all the information described and extracted from the simulations. With all of them in our hands, it is possible to identify several events occurring in the simulations, such as isomerization, fragmentation, or ion-molecule complex formation.

Before entering into the detailed analysis of the trajectories, we will present some properties of the 2D arrays that will be essential for the analysis and further discussion.

D. Properties of a discrete $\mathbf{X}[i,j]$

All the discrete $\mathbf{X}[i,j]$ 2D-arrays described previously fulfill the following properties:

1. **Maximum number of different items.** $\mathbf{X}[i,j]$ has a maximum number of different items $\rho_{\mathbf{X}} \in [1, N \cdot M]$. The minimum value of $\rho_{\mathbf{X}}$ corresponds to one if all the elements of $\mathbf{X}[i,j]$ are identical while its maximum value corresponds to $N \cdot M$ if all the items are different. Therefore, the sample space of $\mathbf{X}$ can be defined as $\Omega_{\mathbf{X}} \in \{X^1, X^2, ..., X^{\rho}\}$. This property of **CAN** reflects the number of different graphs created during the simulation. This establishes all the graphs populated at least once.

2. **Appearance.** For any of the $\Omega_{\mathbf{X}}$-items, there is a (0,1)-matrix, i.e. $X^{\omega} \mid \mathbf{X}[i,j]$ where $\omega \in [1, \rho_{\mathbf{X}}]$, that accounts for the appearance of X^{ω} in $\mathbf{X}[i,j]$, such that:

$$(X^{\omega} \mid \mathbf{X})_{ij} = \begin{cases} 1 \rightarrow X_{ij} = X^{\omega} \\ 0 \rightarrow X_{ij} \neq X^{\omega} \end{cases}$$

From this property of **CMAS** it is possible to identify when a given cluster is created.

3. **Residence time distribution.** The residence time of a given X^{ω} is the amount of time that it appears continuously in a given trajectory; therefore, the residence time distribution (RTD) is the frequency distribution of the residence time for the whole ensemble of trajectories. It should be noted that the resolution of RTD is related to the value of τ employed. From this property of **CAN** it is possible to determine a lifetime of the desired graph if and only if τ is short enough.

4. **Maximum-residence time distribution.** The maximum-residence time distribution (MRTD) is the frequency distribution of the maximum residence time value per trajectory. The schematic representation of the algorithm used to obtain RTD and MRTD is reported in Supporting Information, Table S1. Note that, only those values for which $RT > 0$ and $MRT > 0$ are useful for the analysis. From this property of **CAN** it is possible to determine the maximum time that a graph can be formed continuously.

5. **Time-dependent probability and standard deviation.** All the items in $\Omega_{\mathbf{X}}$ have associated 1D-matrices which account for the probability $\mathbf{P}^{\omega}[j]$ with the associated standard deviation $\sigma^{\omega}[j]$ of the appearance of

X^{ω} in $\mathbf{X}[\mathrm{i,j}]$ along the discrete time $(\mathrm{j}-1)\,\tau$ for an ensemble of N-trajectories, such that:

$$\mathrm{P}^{\omega}_{\mathrm{j}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{X}^{\omega} \mid \mathbf{X})_{\mathrm{ij}} \tag{3}$$

It corresponds to a reduction of dimensionality of $\mathrm{X}^{\omega} \mid \mathbf{X}[\mathrm{i,j}]$ by calculating the average over the trajectories. From this property of **CMAS** the population of a given cluster is obtained as a function of time .

6. **Mean value or hierarchy.** The average value of $\mathbf{P}^{\omega}[\mathrm{j}]$ is expressed directly by:

$$\langle \mathbf{P}^{\omega} \rangle = \frac{1}{\mathrm{N}\cdot\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{M}} (\mathrm{X}^{\omega} \mid \mathbf{X})_{\mathrm{ij}} \tag{4}$$

This value can be used as a criterion to identify whether an item is more important than others during the simulations. The item X^{μ} is more relevant than X^{ν} iff the $\langle \mathbf{P}^{\mu} \rangle$ is greater than $\langle \mathbf{P}^{\nu} \rangle$. From this property of **CAN** it is possible to establish a hierarchy relative to a global population of a graph in the simulation.

7. **Shannon's entropy.** The disorder of $\mathbf{X}[\mathrm{i,j}]$ can be measured by using Shannon's entropy formula, such that:

$$\mathrm{S}_{\mathbf{X}} = -\sum_{\omega=1}^{\rho_{\mathbf{X}}} \langle \mathbf{P}^{\omega} \rangle \cdot \ln \langle \mathbf{P}^{\omega} \rangle \tag{5}$$

From this property of **CAN** it is possible to obtain a value that reflects the disorder of the system in the simulation.

8. **Closure.** As a matter of consequence from the property 6, the closure condition can be written as,

$$\sum_{\omega=1}^{\rho_{\mathbf{X}}} \langle \mathbf{P}^{\omega} \rangle = \frac{1}{\mathrm{N}\cdot\mathrm{M}}\sum_{\omega=1}^{\rho_{\mathbf{X}}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{M}} (\mathrm{X}^{\omega} \mid \mathbf{X})_{\mathrm{ij}} = 1 \tag{6}$$

IV. THREE-STATE KINETIC MODEL

We now consider the evolution of an ensemble of trajectories whose initial conditions were set in order to reproduce a specific physical condition. In the present case is the vibrational activation under microcanonical conditions. The following refers to the application of the general graph-theory approach described previously to the reactivity of an ensemble of microcanonical trajectories (all trajectories have the same internal energy). The method is general and can be applied without loosing generality to canonical simulations, for example. Since the system has at t = 0 enough energy to react, we now describe the different steps necessary to follow the graph modification in time and then obtain the kinetic properties.

The sample space $\Omega_{\mathbf{CAN}}$ covers all the physically possible graphs $\rho_{\mathbf{CAN}}$ that can be formed considering the λ-atoms. This number depends on the number of bonds of the λ-atoms and increases exponentially with their number. Nevertheless, during the dynamics only a reduced number of graphs are reached as a function of the total energy of the system, the starting activated graph (SAG), the total simulation time (Λ), and the number of propagated trajectories (N).

The selection of $\Omega_{\mathbf{CAN}}$ as a work-space avoids counting as different graphs those snapshots in which the system exchange identical functional groups (H, H^{+}, CH_3, etc). The features of $\Omega_{\mathbf{CAN}}$ can be extracted directly from the properties of the $\mathbf{CAN}[\mathrm{i,j}]$ described previously iff the number of propagated trajectories is large enough. For example, the number of visited graphs $\Omega_{\mathbf{CAN}}(\mathrm{E}_{\nu},\mathrm{N},\Lambda)$ during the simulations for a given E_{ν}, N and Λ is $\rho_{\mathbf{CAN}}$.

We should note that if we use the simple adjacency matrix space, $\Omega_{\mathbf{ADJ}}$, the description of the dynamics could be problematic from a chemical point of view. In fact, two graphs representing the same chemical structure could have different adjacency matrices, and thus, some additional information should be included. The use of canonical label avoids this further complication.

A. Partitioning of the Work-Space and Definition of the States

The first element of an $\Omega_{\mathbf{CAN}}$ is the canonical label (CL) of the starting activated graph (SAG). It corresponds to the CL CAN[1] of the equilibrium starting geometry $\mathbf{Q}_0$. The visited graphs during the simulations can be divided into two regions. The first one covers all the connected graphs. Chemically speaking, it accounts for all the possible isomers or intermediate states before the system breaks eventually into two or more parts. The second zone encompasses all the disconnected graphs. In this region, the primary fragmentations happen. All the related information to this classification is stored in the $\mathbf{CLU}[\mathrm{i,j}]$ array. For a given trajectory i at time $(\mathrm{j}-1)\,\tau$, the $\left(\mathrm{CLU}^{=1} \mid \mathbf{CLU}\right)_{\mathrm{ij}}$ element is one when we have a connected graph and analogously $\left(\mathrm{CLU}^{>1} \mid \mathbf{CLU}\right)_{\mathrm{ij}}$ element is one when a disconnected occurs, otherwise they are zero. The label "= 1" is used when one connected subgraph is detected while "> 1" when they are two or more. Figure 4 shows the resulting Venn diagram of $\Omega_{\mathbf{CAN}}$ accordingly to the partitioning we employ.

The primary fragmentation process of the SAG can be analyzed from the time-evolution of three different global states:

- Primary fragmentation state $|\mathrm{PF}\rangle$. A trajectory is said to provide a primary fragmentation event if the system stays a given amount of time, Δ, in the disconnected region continuously. The Δ value must be large enough to account for only those trajectories that once they reach the disconnected region never recross to the connected one. Of course, the time discretization, $\Delta/\tau = \beta \in \mathbb{N}$, due to the storing frequency time holds (and in any case it will be valid since computer trajectories are always time discretized). The $|\mathrm{PF}\rangle$ assignment starts by building the primary fragmentation

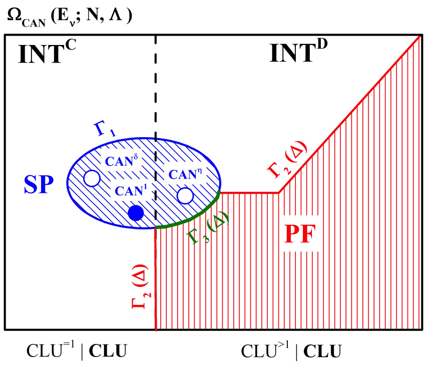

Figure 4. Venn diagram of $\Omega_{\mathbf{CAN}}$ displaying the definition of the states. Γ establishes the boundaries between different states.

guess matrix (PFG), such as:

$$\mathrm{PFG}\,[\mathrm{i,j}] = \mathrm{CLU}^{>1} \mid \mathbf{CLU}\,[\mathrm{i,j}] \tag{7}$$

Subsequently, the (0,1)-matrix that represents $|\mathrm{PF}\rangle$ is obtained. A schematic representation of the adopted algorithm is reported in the Supporting Information, Table S2. Note that here we are dealing with gas phase fragmentation for which once the $|\mathrm{PF}\rangle$ is obtained it typically does not come back to the connected region. The approach can be in principle extended with physical conditions in which a chemical equilibrium is obtained between reactants and products with some (minor) modification. However, this will not be discussed in the present work.

Note that a trajectory to be classified as being in the primary fragmentation region has to be a disconnected graph for a given amount of time. So the PF state is not a set of disconnected graphs, but a condition that they must meet along with the time.

The use of Δ implies a reduction in the dimensionality of the PF-matrix, $\mathrm{N} \times (\mathrm{M}-\beta)$, compared to PFG-matrix, $\mathrm{N} \times \mathrm{M}$. This means that the trajectories can be analyzed only during $\Lambda - \Delta$ time. Then, a key point is the proper estimation of Δ. Using a too small value will increase the number of incorrect assignment of PFs, while a too large value will decrease the statistics in the analysis at the same computational cost. In other words, a larger value of Δ will provide fewer ion-molecule complexes wrongly assigned as primary fragmentation.

To evaluate in the present case a reasonable value of Δ and provide an approach which can be used for other problems, we consider the maximum residence time distribution of PFG $[\mathrm{i,j}]$, previously defined. This gives the information about the maximum time the system stays in the disconnected region before re-entering into the connected one. The MRTD of the PFG matrix for two simulations of CYC^{00} and LIN^{00} at 177 and 151 kcal/mol, respectively, is reported in the Supporting Information, Figure S1. Using $\Delta = 5$ ps the error is less than 1%, which can be considered as a good compromise and used in the following. Note that this allows to remove only 5 ps from the analysis, thus providing 15 and 35 ps that can be analyzed from trajectories long 20 and 40 ps, respectively. Using a larger Δ will imply to run longer simulations or to have a smaller statistical sampling for the analysis (so a larger uncertainty).

- Starting point state, $|\mathrm{SP}\rangle$. It corresponds to the set of CLs which have the same chemical structure of the geometry considered as initial state: $\mathrm{SP} = \{\mathrm{CAN}^1, \mathrm{CAN}^2, ..., \mathrm{CAN}^\eta\}$. This problem can arise because the initial structure is activated randomly and differently for each trajectory. The automatic algorithm can find a different connectivity and thus a different CAN, while they represent the same physical state and one should analyze them together.

Three types of subsets can be distinguished. The first one is simply $\mathrm{SP}^{\mathrm{SAG}} = \{\mathrm{CAN}^1\} \subseteq \mathrm{SP}$, corresponding to the CL of SAG, with the corresponding appearance matrix:

$$\mathrm{SP}^{\mathrm{SAG}}\,[\mathrm{i,j}] = \mathrm{CAN}^1 \mid \mathbf{CAN}\,[\mathrm{i,j}] \tag{8}$$

This is the major contributor to the starting points state, nevertheless, during the microcanonical normal mode sampling of initial conditions, $\mathbf{Q}_0 + \Delta\mathbf{Q}$, some other canonical labels are inherently populated and recognized as different in the analysis. Of course, they depend on the α parameter used in the construction of the connectivities. However, it will be simpler and more practical to keep α constant and elaborate a strategy to identify the other CANs. They can be of two kinds which we list in the following.

- $\mathrm{SP}^{\mathrm{C}} = \{\mathrm{CAN}^2, \mathrm{CAN}^3, ..., \mathrm{CAN}^\delta\} \subseteq \mathrm{SP}$: ensemble of CL of connected (C) graphs that have a similar structure to the SAG. It includes those graphs barely populated directly from the SAG that contain unphysical atomic valencies and, in some cases, those graphs directly populated from the SAG by a fast dynamical process such as proton transfer – this is likely to occur when treating protonated peptides, for example, in which compact structures are common in the gas phase. The corresponding appearance matrix is:

$$\mathrm{SP}^{\mathrm{C}}\,[\mathrm{i,j}] = \sum_{q=2}^{\delta} \mathrm{CAN}^q \mid \mathbf{CAN}\,[\mathrm{i,j}] \tag{9}$$

- $\mathrm{SP}^{\mathrm{D}} = \{\mathrm{CAN}^{\delta+1}, \mathrm{CAN}^{\delta+2}, ..., \mathrm{CAN}^\eta\} \subseteq \mathrm{SP}$: set of canonical labels of disconnected (D) graphs formed as a result of bond breaking directly from SAG when $\mathrm{E_v}$ increases. Typically, these disconnected graphs result from the $\mathrm{X-H}$ bond rupture due to the fixed value of α. A crucial requirement is that any of these graphs should not be involved in the primary fragmentation event, e.g., dehydrogenation or deprotonation. Therefore, the appearance matrix becomes:

$$\mathrm{SP^D}\left[\mathrm{i,k}\right] = \sum_{\mathrm{q}=\delta+1}^{\eta} \mathrm{CAN^q} \mid \mathbf{CAN}\left[\mathrm{i,k}\right] \circ \left\{\mathrm{J}\left[\mathrm{i,k}\right] - \mathrm{PF}\left[\mathrm{i,k}\right]\right\} \tag{10}$$

where $\mathrm{k} \in [1, \mathrm{M}-\beta]$ and $\mathrm{J}[\mathrm{i,k}]$ is a all-ones matrix. The circle represents the Hadamard product between matrices. As can be seen directly, the $\mathrm{SP^D}$ matrix shows a reduction in the dimensionality due to its dependence on PF.

The (0,1)-matrix of the starting point state can be directly obtained by summing all of its components and considering the reduction of the dimensionality:

$$\mathrm{SP}\left[\mathrm{i,k}\right] = \mathrm{SP^{SAG}}\left[\mathrm{i,k}\right] + \mathrm{SP^C}\left[\mathrm{i,j}\right] + \mathrm{SP^D}\left[\mathrm{i,k}\right] \tag{11}$$

The set of graphs chosen as the starting point state for $\mathrm{CYC^{00}}$ and $\mathrm{LIN^{00}}$ system are reported in the Supporting Information, Figures S2 and S3. As can be seen, a proper selection of $\mathrm{SP^C}$ and $\mathrm{SP^D}$ makes $|\mathrm{SP}\rangle$ mostly independent on the α value. Only the trajectories that after the vibrational normal-mode sampling used to generate initial conditions, provide (so for t = 0) a canonical label equals to SP were propagated, being $\mathrm{SP}[\mathrm{i},1] = 1\,\forall \mathrm{i}$. This procedure allows to control the activation process, avoiding that the automatic analysis identifies some unphysical isomerization or fragmentation at t = 0. Finally, we can obtain the time-evolution probability and the associated standard deviation of $|\mathrm{SP}\rangle$.

An example of the time evolution of the different starting points defined previously ($\mathrm{SP^{SAG}}$, $\mathrm{SP^C}$ and $\mathrm{SP^D}$) for two simulations, $\mathrm{CYC^{00}}$ at 217 kcal/mol and $\mathrm{LIN^{00}}$ at 191 kcal/mol, is reported in Supporting Information, Figure S4. As expected, the $\mathrm{SP^{SAG}}$ represents the major contribution to the starting-point state while $\mathrm{SP^C}$ and $\mathrm{SP^D}$ represent a limited fraction.

- Intermediate state $|\mathrm{INT}\rangle$: all snapshots that are not classified neither $|\mathrm{SP}\rangle$ nor $|\mathrm{PF}\rangle$. The (0,1)-matrix is defined as:

$$\mathrm{INT}\left[\mathrm{i,k}\right] = \mathrm{J}\left[\mathrm{i,k}\right] - \mathrm{SP}\left[\mathrm{i,k}\right] - \mathrm{PF}\left[\mathrm{i,k}\right] \tag{12}$$

where $\mathrm{k} \in [1, \mathrm{M}-\beta]$ and $\mathrm{J}[\mathrm{i,k}]$ is an all-ones matrix. As a matter of fact, the INT matrix also shows a reduction in the dimensionality due to its dependence on the PF-matrix.

There are two types of intermediate states, and notably:

i. Connected intermediates $\mathrm{INT^C}$: they are the set of connected graphs (isomers) reached by the system before it reacts, for which

$$\mathrm{INT^C}\left[\mathrm{i,j}\right] = \mathrm{CLU}^{=1} \mid \mathbf{CLU}\left[\mathrm{i,j}\right] - \mathrm{SP^{SAG}}\left[\mathrm{i,j}\right] - \mathrm{SP^C}\left[\mathrm{i,j}\right] \tag{13}$$

or

$$\mathrm{INT^C}\left[\mathrm{i,j}\right] = \sum_{\mathrm{q}=\eta+1}^{\xi} \mathrm{CAN^q} \mid \mathbf{CAN}\left[\mathrm{i,j}\right] \tag{14}$$

where $\{\mathrm{CAN}^{\eta+1}, \mathrm{CAN}^{\eta+2}, ..., \mathrm{CAN}^{\xi}\}$ is the canonical labelling of the connected graphs reachable in the simulation.

ii. Disconnected intermediates $\mathrm{INT^D}$: it is a disconnected area where the primary fragmentation requirement is not fully accomplished. These structures in fact will come back to a connected graph after a given time. In the case of ion fragmentation, this region is where ion-molecule complexes take place. More in general, this region identifies the formation of (transient) complexes, which can occur also in the case of neutral fragmentation and can be related to a roaming mechanism.[58,59] We thus have:

$$\mathrm{INT^D}\left[\mathrm{i,k}\right] = \mathrm{CLU}^{>1} \mid \mathbf{CLU}\left[\mathrm{i,k}\right] - \mathrm{PF}\left[\mathrm{i,k}\right] - \mathrm{SP^D}\left[\mathrm{i,k}\right] \tag{15}$$

We should note that in the present approach the same graph can be classified as INT or PF depending on its fate. For example, if inside the Δ time a disconnected graph comes back to the connected region it is classified as INT, otherwise as PF. Given the definition of these three states it is possible to follow their evolution as a function of time from which rate constants can be extracted, as we will show in Section IV C. Before, we will show in the next section that also energetic information can be obtained.

B. Energy Decomposition

The energy information stored in the $\mathbf{POT}[\mathrm{i,j}]$-array can be decomposed into three different arrays according to the defined states, such as:

$$\begin{aligned}
\mathbf{POT}^{\mathrm{SP}}\left[\mathrm{i,j}\right] &= \mathrm{SP}\left[\mathrm{i,j}\right] \circ \mathbf{POT}\left[\mathrm{i,j}\right] \\
\mathbf{POT}^{\mathrm{INT}}\left[\mathrm{i,k}\right] &= \mathrm{INT}\left[\mathrm{i,k}\right] \circ \mathbf{POT}\left[\mathrm{i,k}\right] \\
\mathbf{POT}^{\mathrm{PF}}\left[\mathrm{i,k}\right] &= \mathrm{PF}\left[\mathrm{i,k}\right] \circ \mathbf{POT}\left[\mathrm{i,k}\right]
\end{aligned} \tag{16}$$

The $\circ$ denotes the element-wise product (Hadamard multiplication) as introduced before. Each term corresponds to the potential energy of the three states: starting point, intermediate and primary fragmentation, respectively. The same decomposition can be applied to $\mathbf{KIN}[\mathrm{i,j}]$-array. From these newly defined continuous arrays the energy gap between the states can be obtained.

The $\langle \mathrm{V} \rangle$-value can be used to provide an estimation of the relative stability of the two structures during the dynamics. The potential energy difference between $\mathrm{CYC^{00}}$ and $\mathrm{LIN^{00}}$ at 0 K is about 26 kcal/mol at RM1-D level of theory which is used in the trajectories, while from the simulations it is between 11 and 12 kcal/mole depending on the activation energy, as shown in Table I. This will correspond to an effective

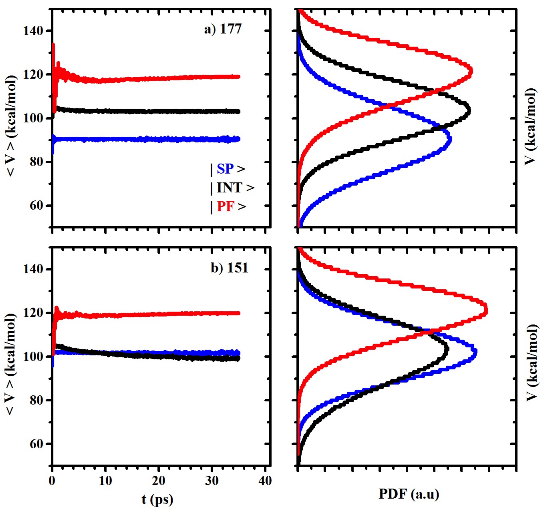

Figure 5. Average potential energy as the function of time (left panel) and corresponding probability density function (right panel) as obtained from chemical dynamics simulations at two activation energies : a) CYC^{00} at 177 kcal/mol and b) LIN^{00} at 151 kcal/mol.

energy difference which will be discussed later when comparing the threshold energies obtained from the fragmentation simulations.

In Figure 5 we show for each state the average (over the different trajectories per each snapshot) potential energy of the three states as a function of time as well as the corresponding probability density function (PDF). As an example, we display values for simulations of CYC^{00} and LIN^{00} at two activation energies (177 and 151 kcal/mol, respectively). Note that by considering the energy difference between the minimum energy geometries of the two structures (26 kcal/mol), the two sets have the same excess energy..

In the CYC^{00} system the three states are very-well resolved in energy. The maximum values of the PDF are located at 90.5, 103.5 and 121.5 kcal/mol for $|\mathrm{SP}\rangle$, $|\mathrm{INT}\rangle$ and $|\mathrm{PF}\rangle$, respectively. The LIN^{00} has the $|\mathrm{SP}\rangle$ located 11 kcal/mole above the $|\mathrm{SP}\rangle$ of CYC^{00}, while the $|\mathrm{PF}\rangle$ has the same energy as in CYC^{00}. This reflects that the fragmentation products are in average very similar for the two systems. This is because LIN^{00} is the most populated isomer during the fragmentation of CYC^{00} (as we will detail later). The average potential energy of $|\mathrm{INT}\rangle$ in LIN^{00} varies along the time because the system will isomerize going to the most stable graph, CYC^{00}. That explains why the distribution goes below the $|\mathrm{SP}\rangle$ energy distribution. Similar behaviors were obtained for other internal energies.

The same energy decomposition can be applied to every canonical label. Thus, the mean values of the potential energy of a given graph CAN^{q} can be determined by:

$$\mathrm{POT}^{q}\,[\mathrm{i},\mathrm{j}] = \mathrm{CAN}^{q} \mid \mathbf{CAN}\,[\mathrm{i},\mathrm{j}] \circ \mathbf{POT}\,[\mathrm{i},\mathrm{k}] \qquad (17)$$

The same holds for kinetic energy.

This approach will provide mean values of kinetic and potential energies of the states defined by the state partitioning.

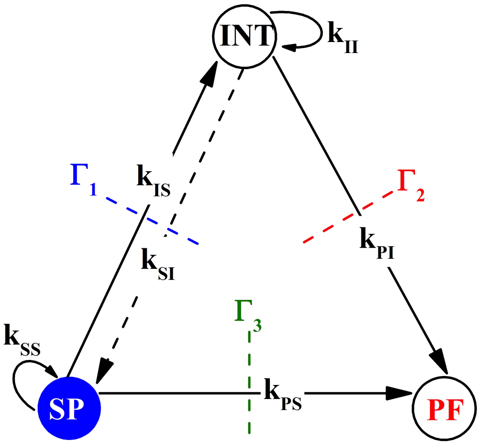

Figure 6. Markov diagram for the three-state system that models the unimolecular dissociation in the gas phase.

They can be used to determine the stability of the different states and/or canonical labels from an average and dynamical perspective.

C. Kinetic model for a three-state system

By using the three states defined and determined as detailed previously, we can now set up a kinetic scheme from which "effective" microcanonical rate constants can be obtained. We call them effective because they reflect the simplification of a complex kinetic network to a simple three-state model. In other words, the intermediate and product states do not correspond to specific molecules or geometries, but to a global definition. The kinetic scheme can be constructed directly from the Venn diagram previously introduced. From it a three-state Markov model (shown in Figure 6) describes the statistical kinetic process of primary fragmentation in the gas phase.

The $|\mathrm{SP}\rangle$ state is represented by a filled blue circle to indicate that at $t = 0$, it is the only populated state. Previously we have highlighted that the choice of the Δ-parameter guarantees that once the system reacts, either from the initial state or from the intermediate state, it never returns back to reactant or intermediate state. This is a typical situation of gas-phase reaction, in which the fragmentations are not reversible since the probability that the two fragments encounter each other in vacuum is almost zero. This means that a given trajectory will cross the Γ_3 or Γ_2 dividing surfaces only once and there are no reverse rate constants from PF to SP or INT. As we have discussed in the definition of INT state, if two (or more) molecules are formed but they form back one molecule the state is labeled as INT. In other words PF state cannot by definition comes back to reactants, and the choice of Δ is crucial. Of course, further fragmentations are possible, but they are not discussed here. They can be treated with the same approach using a specific fragmentation product as initial state.

The kinetic equations of this general scheme can be written as:

$$\frac{dN_{SP}}{dt} = k_{SI} \cdot N_{INT} - (k_{IS} + k_{PS}) \cdot N_{SP}$$

$$\frac{dN_{INT}}{dt} = k_{IS} \cdot N_{SP} - (k_{SI} + k_{PI}) \cdot N_{INT} \qquad (18)$$

$$\frac{dN_{PF}}{dt} = k_{PS} \cdot N_{SP} + k_{PI} \cdot N_{INT}$$

where N_{SP}, N_{INT} and N_{PF} denotes the abundance of the starting point, intermediate and primary fragmentation states, respectively while k_{SI}, k_{IS}, k_{PS} and k_{PI} are the corresponding "effective" rate constants which for the present simulations are microcanonical. In general, for each k_{ij}, "i" indicates the destination state and "j" the source state. The diagonal elements can be expressed as: $k_{SS} = -(k_{IS} + k_{PS})$ and $k_{II} = -(k_{SI} + k_{PI})$. From now on, the word "effective" is suppressed from the discussion for simplicity.

The solution of the system of linear first-order differential equations assuming that at $t = 0$ the only populated state is $N_{SP}(0) = N_0$, can be obtained using the Laplace-Carson transform method.[22,77] The analytical solution is reported in the Supporting Information.

Since the energy is conserved and enough to pass the barrier, the limit for $t \rightarrow \infty$ is $N_{SP}/N_0 = 0$, $N_{INT}/N_0 = 0$ and $N_{PF}/N_0 = 1$.

The microcanonical rate constants are obtained by making a simultaneous fit using the solutions of the differential equations system from the population of the states obtained in simulations. The simultaneous fit was made using a home-made program developed in C++ through the modular scientific software ROOT. The MINUIT library was employed during the minimization procedure. The ROOT scripts are provided in a public repository (https://github.com/afperezmellor?tab=repositories).

To obtain all the kinetic parameters we employed two approaches: (A) all the rate constants are obtained from the simultaneous fit; (B) the primary fragmentation rate constant is obtained from the reaction flux, and the others from the fit. We now show how we obtained the rate constants using the two approaches and how they compare each other.

1. *Full simultaneous fit (method A)*

The microcanonical rate constants are obtained from each ensemble of trajectories as a function of the activation energy by fitting the populations of SP, INT and PF using the kinetic equations described above by a simultaneous fit. In Figure 7 we report the resulting fits together with the population evolution obtained from CYC^{00} simulations at the different activation energies (the same from LIN^{00} are reported in Supporting Information, Figure S5). As can be clearly seen, the fit is very good and thus the rate constants can be extracted safely. This shows that the simple three-state kinetic model can globally catch the fragmentation dynamics of this system.

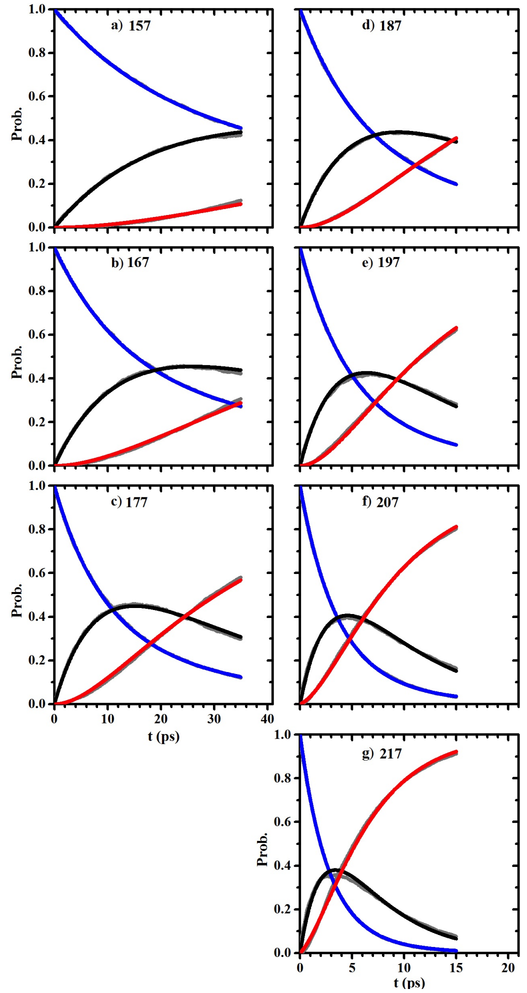

Figure 7. Populations of the three states (gray lines) at different vibrational energies (reported in kcal/mol) of CYC^{00} system fitted by the simultaneous fit, method A. Blue: starting point state; black: intermediate state; red: primary fragmentation state.

2. *Reaction flux and partial fit (method B)*

In the second approach used (method B) we first estimate the most relevant reactive rate constant from the simulations by a direct measure using the flux.[78]

In fact, the primary fragmentation can be described as a reaction flux in sample space across the dividing surfaces, $\Gamma_2(\Delta)$ and $\Gamma_3(\Delta)$. The role of the Δ parameter is crucial to minimize the flux through these surfaces. We should notice that $\Gamma_2(\Delta)$ and $\Gamma_3(\Delta)$ do not correspond to any saddle point geometry, but they represent abstract dividing surfaces. This has the advantage of getting rate constants which represent the overall kinetics, independently on geometrical details which can, in general, make impossible to conceive and analyze the full set of reaction pathways characteristic of fragmentation of complex systems. We will discuss how a specific pathway of particular interest can be studied out from the global pathway description in section V.

In this framework, the reactive rate constants can be obtained as a function of time from the flux of reactive graphs passing across the critical region divided by the remaining graphs in the departure state, such that:

$$k_{PI}(t) = \frac{1}{\tau}\frac{|INT \rightarrow PF\rangle}{|INT \rightarrow INT\rangle} \qquad k_{PS}(t) = \frac{1}{\tau}\frac{|SP \rightarrow PF\rangle}{|SP \rightarrow SP\rangle} \quad (19)$$

In the systems studied here, there are only a few trajectories that react directly from the starting point, so the sampling will not be enough to get a converged value. Therefore, we have calculated only k_{PI} directly from the flux.

The rate constants obtained via the flux method for CYC^{00} simulations are reported in the Supporting Information (Figures S6 and S7). The rate constants are here time dependent and we observe a not negligible spread. At low energies this spread is larger, probably because the fragmentation is not fully converged. LIN^{00} simulations show a similar behavior.

We have then used the position of the maximum value of the probability distribution functions as k_{PI} and fitted the other rate constants. As previously, the fit results describe very well the simulation data (see Figure S8 in the Supporting Information). We should notice that this approach has the disadvantage that a huge number of trajectories is necessary to have a convergence in the flux calculations, while less are needed when using only the solution of the kinetic scheme from population evolution.

D. Rate Constants

The rate constants obtained at different energies and with the two methods are reported in Table S3. As we can notice, the two methods are in very good agreement providing very similar results.

For both systems, the rate constant values increase monotonically with internal energy. Only the k_{SI} of CYC^{00} are almost energy independent. In the CYC^{00} system, the k_{IS} is always larger than k_{PI}. A different trend is observed for LIN^{00} system, where at low activation energy k_{IS} is higher than k_{PI} and vice-versa for larger internal energy.

The k_{PI} values for LIN^{00} system are always higher than its counterpart CYC^{00} at the same total energy. This observation is easily explained by the fact that the primary fragmentation state is found on average at the same potential energy level for both systems for the same total energy (see Figure 5). In addition, the starting point state CYC^{00} is lower in energy by 11-12 kcal/mol (as obtained by the simulations at different activation energies) with respect to LIN^{00}. Therefore, the energy gap between the two states, $|PF\rangle - |SP\rangle$, is smaller in LIN^{00} simulations than in CYC^{00}. We should remind that they share the same reactive basin and, then, the system LIN^{00} will have more primary fragmentation events than its counterpart CYC^{00} under the same total energy condition.

E. Threshold Energies

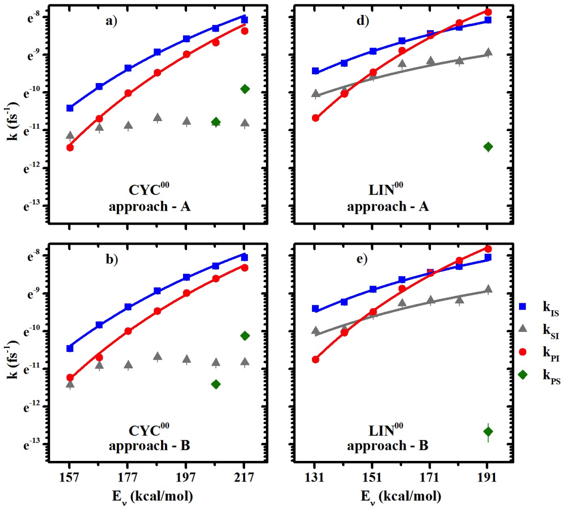

Figure 8. Microcanonical rate constant values (symbols) of the different processes as a function of the internal energies. The colored line represents the RRK fit from which the energy threshold and the effective frequency factors are obtained.

Using the two approaches described previously, we obtained the rate constants as a function of the excess energy for the two systems. The rate constants are obtained from microcanonical simulations and thus their energy dependence should follow the classical RRKM theory (also called RRK theory):[21]

$$k(E_v) = \nu\left(1 - \frac{E_a}{E_v}\right)^{s-1} \quad (20)$$

where ν and E_a are the effective frequency and threshold energy, respectively, while s is the number of vibrational degrees of freedom of the reactant. ν and E_a are two adjustable parameters that can be obtained by fitting the rate constants at different energies via Equation 20. Figure 8 shows how rate constants behave as a function of energy, E_v, with the corresponding RRK fits. As previously discussed, a clear energy dependence was observed for k_{IS} and k_{PT} in both CYC^{00} and LIN^{00} simulations, and k_{SI} in LIN^{00} simulations. Note that for k_{PS} we have too few points to obtain a reasonable fit. As already noticed, the k_{SI} in CYC^{00} simulations do not show any energy dependence. This is probably due to the fact that the threshold energy is too low and that there is no time for energy randomization of the INT state before reaction.

From the RRK fits, we obtained frequency factors, ν, and threshold energies, E_a, which are listed in Table IV.

As can be seen, the RRK model explains properly the evolution of the microcanonical rate constants, obtained from the simulations, as a function of the excess energy. As expected, the LIN^{00} system needs less energy to form the primary fragmentation products than CYC^{00} system. The difference between the threshold energies, 11 kcal/mol, corresponds to the energy difference between the starting point states obtained from the potential energy average (between 11 and 12 kcal/mol as a function of the activation energy). This result strengthens the picture for which the stability of the states is better described by the expectation value of the

	CYC^{00} Method		LIN^{00} Method	
	A	B	A	B
ν_{IS}	0.110 ± 0.014	0.111 ± 0.016	0.005 ± 0.0006	0.005 ± 0.0009
$\mathrm{E}^{\mathrm{a}}_{\mathrm{IS}}$	30.480 ± 0.517	30.518 ± 0.561	13.845 ± 0.561	13.725 ± 0.675
ν_{SI}			0.001 ± 0.0004	0.001 ± 0.0007
$\mathrm{E}^{\mathrm{a}}_{\mathrm{SI}}$			11.147 ± 1.116	11.850 ± 1.619
ν_{PI}	0.400 ± 0.096	0.263 ± 0.037	0.113 ± 0.009	0.133 ± 0.013
$\mathrm{E}^{\mathrm{a}}_{\mathrm{PI}}$	37.859 ± 0.919	36.169 ± 0.919	26.375 ± 0.919	26.915 ± 0.362

Table IV. Effective frequencies factors (in fs^{-1}) and energy thresholds (in kcal/mol) obtained following the methods to obtain the rate constants and then fitted using the RRK model.

potential energy from the simulations rather than from the difference between potential energy values at minima.

Concluding, the proposed three-state model gives a general overview of the description of the primary fragmentation events providing quantitative information on rate constants and threshold energies.

F. Structures of the Intermediate States

The intermediate state is defined as a state which is different from the reactant but not (yet) primary fragmentation state. We now show that our approach can provide more molecular information on this state from graph-theory based analysis. First we can consider three subsets: linear connected intermediates, cyclic connected intermediates, and ion-molecule complexes (IMC). The first and the second subsets account for the linear and cyclic structures reached after the starting structure is activated. The latest are those disconnected structures that do not accomplished the primary fragmentation requirement.

The branching ratio of the intermediate state as a function of time as obtained at a given energy (the same total energy) from CYC^{00} and LIN^{00} simulations is shown in Figure 9. For the CYC^{00} system, the linear structures represent almost 90% of the intermediate state while the cyclic structure is around 7% and IMC 3%. A completely different picture is observed for LIN^{00} system. The linear structure drops down up to 47% while the cyclic formation grows up to 47%. The IMC represents around 7% of the intermediate state.

The explanation of such difference lies in the nature of the isomerization process. The CYC^{00} system, which is the global minimum, has a cyclic structure. Once it is activated, the ring opens and thus the mostly formed intermediates are linear, such as LIN^{00}. Note that in Figure 9 we show the ratio between three classes of intermediate states (INT) and thus the initial state (SP) is not included. After the ring opening, several proton transfers take place populating many linear structures. On the other hand, when the linear, LIN^{00}, is activated, the main isomerization event is the formation of a cyclic structure, and notably CYC^{00}, which is the most stable isomer.

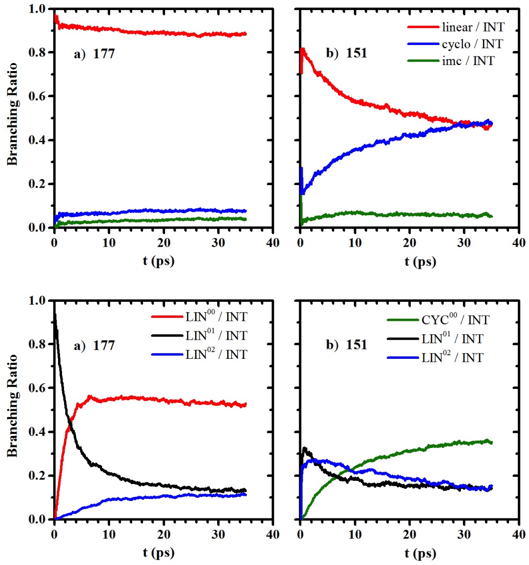

Figure 9. Branching ratios : Upper panels, decomposition of the intermediate state between linear, cyclic, and IMC structures at the same total energy. a) CYC^{00} simulations at 177 kcal/mol. b) LIN^{00} simulations at 151 kcal/mol; Lower panels, the three most relevant structures of the intermediate state as obtained from simulations: a) CYC^{00} at 177 kcal/mol and b) LIN^{00} at 151 kcal/mol.

It is possible to follow the time-abundance of the different structures of the intermediates. We report in Figure 9 the evolution of the most relevant ones (the corresponding structures are reported in the Supporting Information, Figure S9, together with the other intermediates). We can notice that the linear structures mostly populated correspond to different tautomers: the proton transfer is a key reaction occurring to these intermediate structures, in particular for linear structures, following the mobile-proton model.[61–63] On the other hand, once a cyclic structure is formed this is much more stable than the linear counterpart. These processes are typical of peptide fragmentation and we have noticed some of them in previous simulations:[6,7,30,31,52,69,70,79–83] here our approach is able to automatically characterize them and provide information on the different structures before their visual inspection.

The structures of the ion-molecule complexes (IMC) obtained during the simulations are reported in Figure S10 of the Supporting Information. They appear as pairwise fragments; one entity is charged while the other is neutral.

During the dynamic, they may interchange some functional groups, as observed here, resulting in different IMC structures: IMC^{01}_{01}, IMC^{01}_{02}, IMC^{01}_{03} and IMC^{01}_{04}. This is an indication that a roaming mechanism took place.

V. PROPERTIES OF A SPECIFIC PATHWAY

Up to now, the kinetic properties were obtained for the primary fragmentation state as a whole without any differentiation on their structure. We now show how the kinetic scheme can be adapted to obtain the desired properties for one specific path of interest. The approach is totally general, and of course if more specific products are interesting the procedure should be repeated for each pathway. Of course, one needs to have enough trajectories leading to a specific pathway to have statistically significative information.

When a primary fragmentation is identified, we can distinguish three classes of events:

1. Direct primary fragmentation : $M^+ \rightarrow M_1^+ + M_2$. It occurs when the molecule reacts directly forming two entities of masses M_1 and M_2. The information related to this events can be extracted from the **MAS** array. It is then useful to focus on a particular event, say $M^\alpha = \{M_1, M_2\}$ which can be monitored during simulation using the properties described previously. The condition that a specific fragmentation occurs formally corresponds to :

$$\mathrm{MASD}\,[i,k] = \mathrm{MAS}^\alpha | \mathbf{MAS}\,[i,k] \circ \mathrm{PF}\,[i,k] \qquad (21)$$

Then, MASD (0,1)-array displays whether a given snapshot contains an event of direct primary fragmentation or not, and the time evolution can also be obtained.

2. Ion-Molecule formation : $M^+ \rightarrow M_3^+ \cdots M_4 + M_2$ where $M_3^+ + M_4 = M_1^+$. This process occurs when a complex is obtained during the fragmentation that gives M_1^+. This event can be followed through the **CMAS** array which, as previously seen, takes into account the formation of clusters. Similar to what done for direct events, and considering the same $M^\alpha = \{M_1, M_2\}$ event, it can be stated that:

$$\mathrm{MASC}\,[i,k] = \mathrm{MAS}^\alpha | \mathbf{CMAS}\,[i,k] \circ \mathrm{PF}\,[i,k] \qquad (22)$$

Note that this is not an intermediate ion-molecule complex, but a case in which one (at least) product is itself a complex.

3. Higher-order of fragmentation event: they are structures which are neither direct fragmentation nor ion-molecule formation events. These events are predominant when the activation energy is too high and the system breaks suddenly in many pieces.

The formation of a given mass channel $M^\alpha = \{M_1, M_2\}$ via direct primary fragmentation can be obtained by:

$$\mathrm{MASG}\,[i,k] = \mathrm{MASD}\,[i,k] + \mathrm{MASC}\,[i,k] \qquad (23)$$

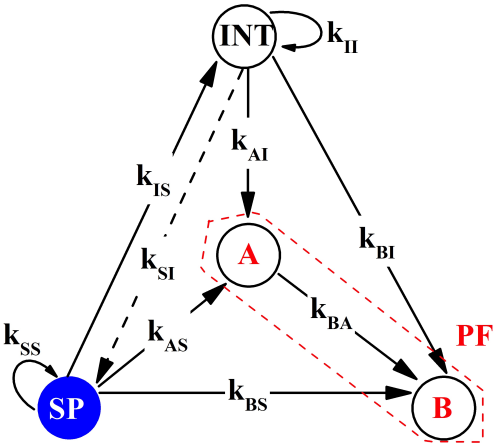

Figure 10. Markov diagram for a particular four-state system that explains the formation of a given fragmentation product by primary fragmentation in the gas phase.

It should be noted that M_2 can fragment subsequently into different pieces depending on its nature and excess energy. In that case, one can follow simply M_1^+ and $M_3^+ \cdots M_4$ as primary fragmentation products, disregarding the fate of the neutrals. At this end one should use the information on the fragmentation products. These events were not found here due to the low mass of the neutrals and relatively low fragmentation energy.

This approach will provide the different fragmentation channels, in terms of clusters and fragments (an example is given in Table S4 of the Supporting Information). This information will be useful to build the theoretical mass spectrum, as we will show in section VI. Before, we will show how the kinetic parameters of a specific fragmentation product can be obtained.

Once we monitor the appearance of specific fragmentation channels, it is then possible to determine their unimolecular kinetics. Similar to how it was done to explain the global process of primary fragmentation, to extract the rate constants we first built a kinetic model with four states, shown in Figure 10. The primary fragmentation state is now divided into two new states, $|A\rangle$ and $|B\rangle$. The first takes into account the process of interest, for example, the formation $M^\alpha = \{M_1, M_2\}$ while the second all the other processes, so that $|B\rangle = |PF\rangle - |A\rangle$.

After establishing all possible interconnections between the states and considering that once the system reacts it does not return to the precursor state, the system of ordinary differential equations will be :

$$
\begin{aligned}
\frac{\mathrm{dN_{SP}}}{\mathrm{dt}} &= \mathrm{k_{SI}} \cdot \mathrm{N_{INT}} - (\mathrm{k_{IS}} + \mathrm{k_{AS}} + \mathrm{k_{BS}}) \cdot \mathrm{N_{SP}} \\
\frac{\mathrm{dN_{INT}}}{\mathrm{dt}} &= \mathrm{k_{IS}} \cdot \mathrm{N_{SP}} - (\mathrm{k_{SI}} + \mathrm{k_{AI}} + \mathrm{k_{BI}}) \cdot \mathrm{N_{INT}} \\
\frac{\mathrm{dN_{A}}}{\mathrm{dt}} &= \mathrm{k_{AS}} \cdot \mathrm{N_{SP}} + \mathrm{k_{AI}} \cdot \mathrm{N_{INT}} - \mathrm{k_{BA}} \cdot \mathrm{N_{A}} \\
\frac{\mathrm{dN_{B}}}{\mathrm{dt}} &= \mathrm{k_{BS}} \cdot \mathrm{N_{SP}} + \mathrm{k_{BA}} \cdot \mathrm{N_{A}} + \mathrm{k_{BI}} \cdot \mathrm{N_{INT}}
\end{aligned}
\tag{24}
$$

The analytical solution (reported in the Supporting Information) of this system was obtained in a similar way to that of three-state model using the Laplace-Carson transform method. It is important to note that there are four totally independent populations $\mathrm{N_{SP}}$, $\mathrm{N_{INT}}$, $\mathrm{N_A}$, and $\mathrm{N_B}$ and seven rate constants to be determinated: $\mathrm{k_{IS}}$, $\mathrm{k_{SI}}$, $\mathrm{k_{AS}}$, $\mathrm{k_{BS}}$, $\mathrm{k_{AI}}$, $\mathrm{k_{BI}}$ and $\mathrm{k_{BA}}$. However, two have already been obtained in the three-state system $\mathrm{k_{IS}}$ and $\mathrm{k_{SI}}$ together with the sum of two of them as, $\mathrm{k_{PS}} = \mathrm{k_{AS}} + \mathrm{k_{BS}}$ and $\mathrm{k_{PI}} = \mathrm{k_{AI}} + \mathrm{k_{BI}}$. Therefore the problem is reduced to four populations and three rate constants, $\mathrm{k_{AS}}$, $\mathrm{k_{AI}}$ and $\mathrm{k_{BA}}$. The rate constant $\mathrm{k_{AI}}$ accounts for the formation of the mass channel through a global intermediate while $\mathrm{k_{AS}}$ does the same but from the starting point. Finally, the rate constant $\mathrm{k_{BA}}$ corresponds to the secondary fragmentation of the channel of interest if there is any.

A simultaneous fit of the populations, following the method A, was performed to obtain the rate constants involving the $|\mathrm{A}\rangle$ state, $\mathrm{k_{AS}}$, $\mathrm{k_{AI}}$ and $\mathrm{k_{BA}}$. The remaining rates constants, $\mathrm{k_{IS}}$, $\mathrm{k_{SI}}$, $\mathrm{k_{PI}}$ and $\mathrm{k_{PS}}$, are constrained to the values obtained during the simultaneous fit, by using the approaches A or B, of the three-state scheme.

We applied this approach to two primary fragmentation products, which were the most abundant observed: (i) the loss of neutral CO and (ii) the formation of methaniminium cation ($\mathrm{H_2NCH_2^+}$). As previously, we obtained the rate constants from the time evolution of the states and then we fit their energy dependence using the RRK model. The RRK fit results are reported in Table S5. As for the global fit, the values obtained with method A and B are very similar. For simplicity we will comment in the following the values obtained with method A.

Interestingly, in the case of CO-loss we can follow two reactions: (i) the formation of the CO-loss product from an intermediate state (characterized by ν_{AI} and $\mathrm{E^a_{AI}}$) and (ii) the degradation of it forming other products (characterized by ν_{BA} and $\mathrm{E^a_{BA}}$). For the $|\mathrm{INT}\rangle \to |\mathrm{A}\rangle$ reaction, the threshold energy ($\mathrm{E^a_{AI}}$) is slightly lower than the corresponding global threshold energy ($\mathrm{E^a_{PI}}$), which is what expected since this channel is the most abundant one. The same picture is obtained for $\mathrm{CYC^{00}}$ and $\mathrm{LIN^{00}}$ simulations, just the barrier of these lasts is lower. Notably, the differences in threshold energy reflect also in this case the difference of 11-12 kcal/mol obtained from average potential energy analysis.

For the $\mathrm{H_2NCH_2^+}$ formation we do not observe any secondary fragmentation and the energy threshold from $\mathrm{CYC^{00}}$ simulations are about 14.5 kcal/mol higher than the one obtained from $\mathrm{LIN^{00}}$ simulations. This is slightly higher than the average energy difference between the two states, but not large enough to suggest that other processes are relevant.

VI. THE THEORETICAL MASS SPECTRUM

We now show how this approach can finally provide a theoretical mass spectrum. More detailed description of the reactivity will be left to specific studies as done previously for other systems.[65,66] Experimentally this is obtained by counting the number of structures as a function of their mass-to-charge (m/z) ratio. Theoretically we can obtain it in the same way using the rigorous definition of a primary fragmentation product given before. Also an estimation can be made by counting the elements in the $\Omega_{\mathbf{CMAS}}$ work-space. Every item in $\Omega_{\mathbf{CMAS}}$ includes the set of cluster mass formed during the simulations.

All the different items $\rho_{\mathbf{CMAS}}$ are directly related to the possible reactions. For any of these different items $\omega \in [1, \rho_{\mathbf{CMAS}}]$, the charge distribution is calculated a posteriori from the expectation values of the electrostatic potential (ESP) of the molecule on a uniform distribution of points as implemented in Mopac.[73]

Experimentally, the most abundant m/z products observed in fragmentation of protonated cyclo-di-glycine are:[60] 90.0, 87.0, 70.1, 59.0, 58.1, 41.9 and 30.1. We obtained all of them from the different simulations (results for the two initial species and the different excess energies are summarized in Tables S6 and S7 of the Supporting Information) as well as other products. The description of the observed ions is reported in Table S8. Note that from simulations we obtain a high-resolution spectrum with no additional cost. Furthermore, it would be eventually possible to consider isotopic abundances, assuming that the fragmentation process is not largely affected by difference in atomic masses (which is what generally is done when doing isotopic labeling in mass spectrometry).

Similarly to experiments, it is possible to report the relative abundance of the species of interest as a function of activation energy (an example is reported in the Supporting Information, Figure S11).

VII. CONCLUSIONS AND OUTLOOKS

In this work, we have reported a new application of graph theory to analyze an ensemble of reactive trajectories of a complex system. We focus our attention on the gas-phase unimolecular dissociation related to mass spectrometry, but the approach is totally general. One key aspect is the use of the canonical labeling which allows to automatically classify as the same molecular structures two graphs with different adjacency matrices. A flowchart summarizing the key steps to obtain the main information to characterize the reactivity using graph-theory and kinetic analysis is reported in Figure 11.

We should note that recently the spectral graph-theory was used to analyze molecular reactivity,[50,84] which could be less time consuming for much larger systems. Presently the

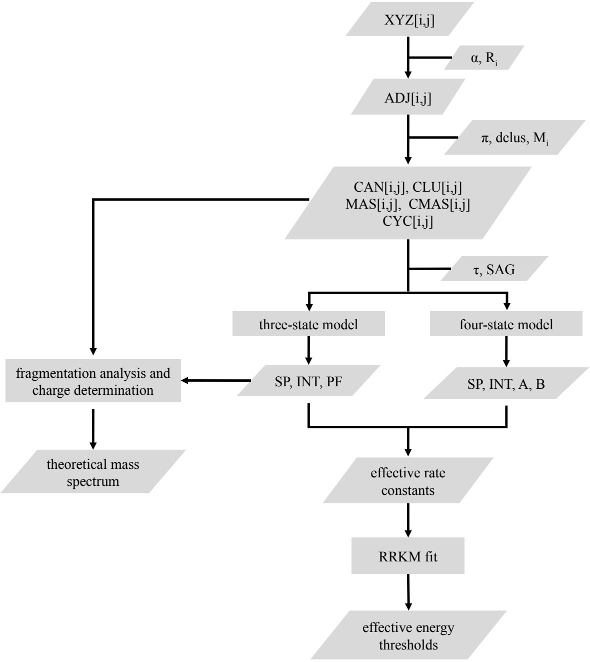

Figure 11. Schematic flowchart summarizing the present approach. We report the different matrices obtained in their sequence and the different parameters, states and properties as defined in the text.

graph-theory analysis using the canonical labeling is much faster than the production of trajectories, and thus the computational time for the trajectory analysis was not a problem. However, when using spectral-graph theory, as in the SPRINT coordinate approach[45] which considers the permutation of identical atoms, a matrix diagonalization is needed which would also be time consuming for very large systems. Future comparisons of the different approaches will surely be useful given the increasing interest of applying graph theory to study chemical reactivity.

Based on the present approach, it is possible to obtain kinetic properties, in terms of rate constants and, after a typical RRK fit, threshold energies. Here we assumed that the products cannot form back the reactants, as it is pertinent in gas phase fragmentation, but a particular care is given in identify transient species which can be obtained, as the ion-molecule complexes, this being relevant to point out possible roaming mechanisms. The method, however, could be extended to physical conditions (and typically reactions in solution) in which a dynamical equilibrium between reactants and products is reached. The main limitations will be in having enough direct dynamics trajectories to have a good statistical sampling. Also the kinetic scheme should be adapted in this case. Future applications and developments of the present method in that direction are surely welcome.

Both general three-state and specific four-state models can be used, this last allowing a more specific analysis of a particular fragmentation product. We should also note that we obtained rate constants also from the flux: results are similar to the simple fit of the population while this last needs less trajectories to have a good convergence. We thus suggest in future to use it when possible.

A particularity of the approach is the general definition of an intermediate state, which is rather abstract (this is useful to obtain rate constants of fragmentation products) but molecular details can also be extracted. Specifically to ion fragmentation is the possibility to identify ion-molecule complexes which are often suggested when explaining products observed in mass spectrometry.

Finally, we have coupled the graph theory analysis to a charge localization method leading to a full theoretical mass spectrum which can be obtained automatically. We have applied it to a prototypical system representative of peptide fragmentation, we now plan to use it in more complex and new systems in order to describe not only qualitatively but also quantitatively (thanks to the access to threshold energies) the fragmentation processes.

SUPPLEMENTARY MATERIAL

Supplementary material is available, where we report: (I) eleven additional figures, (II) eight additional tables and (III-IV) the analytical solutions of 3-state and 4-state kinetic systems. The scripts used in the kinetic fits are reported in a public GitHub repository (https://github.com/afperezmellor?tab=repositories).

AUTHOR'S CONTRIBUTIONS

All authors contributed equally to this work.

ACKNOWLEDGMENTS

This work was supported by ANR DynBioReact (Grant No. ANR-14-CE06-0029-01). We thank Dr. Anne Zehnacker, Dr. Alejandro Gutierrez Quintanilla, Dr. Latevi Max Lawson Daku, Dr. Arnulf Rosspeintner, and Prof. Thomas Burgi for useful discussions and suggestions on the manuscript. We also would like to thank the late Prof. William L. (Bill) Hase for several discussions we will clearly miss in the future.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] N. C. Blais and D. L. Bunker, J. Chem. Phys. **37**, 2713 (1962).

[2] M. Karplus, R. N. Porter, and R. D. Sharma, J. Chem. Phys. **43**, 3259 (1965).

[3] D. Troya and G. C. Schatz, J. Chem. Phys. **120**, 7696 (2004).

[4] U. Lourderaj, K. Park, and W. L. Hase, Int. Rev. Phys. Chem. **27**, 361 (2008).

[5] S. Pratihar, X. Ma, Z. Homayoon, G. L. Barnes, and W. L. Hase, J. Am. Chem. Soc. **139**, 3570 (2017).

[6] S. Pratihar, G. Barnes, and W. Hase, Chem. Soc. Rev. **45**, 3595 (2016).

[7] A. Martin-Somer, V. Macaluso, G. L. Barnes, L. Yang, S. Pratihar, K. Song, W. L. Hase, and R. Spezia, J. Am. Soc. Mass Spectrom. **31**, 2 (2020).

[8] A. Simon, M. Rapacioli, G. Rouaut, G. Trinquier, and F. X. Gadéa, Phil. Trans. R. Soc. A **375**, 20160195 (2017).

[9] P. G. Jasien and R. Shepard, Int. J. Quantum Chem. **34**, 183 (1988).

[10] B. J. Braams and J. M. Bowman, Int. Rev. Phys. Chem. **28**, 577 (2009).

[11] J. M. Bowman, G. Czako, and B. Fu, Phys. Chem. Chem. Phys. **13**, 8094 (2011).

[12] C. Qu, Q. Yu, and J. M. Bowman, Ann. Rev. Phys. Chem. **69**, 151 (2018).

[13] R. Conte, C. Qui, P. L. Houston, and J. M. Bowman, J. Chem. Theory. Comp. **16**, 3264 (2020).

[14] J. A. Nummela and B. K. Carpenter, J. Am. Chem. Soc. **124**, 8512 (2002).

[15] G. Vayner, S. V. Addepalli, K. Song, and W. L. Hase, J. Chem. Phys. **125**, 014317 (2006).

[16] L. Yang, R. Sun, and W. L. Hase, J. Chem. Theory Comput. **7**, 3478 (2012).

[17] Y.-C. Han, P.-Y. Tsai, J. M. Bowman, and K.-C. Lin, Phys. Chem. Chem. Phys. , 18628 (2017).

[18] R. Perez-Soto, S. A. Vazquez, and E. Martinez-Nunez, Phys. Chem. Chem. Phys. **18**, 5019 (2016).

[19] J. Li, B. Zhao, D. Xie, and H. Guo, J. Phys. Chem. Lett. **11**, 8844 (2020).

[20] G. Trinquier, A. Simon, M. Rapacioli, and F. X. Gadéa, Mol. Astrophys. **7**, 37 (2017).

[21] T. Baer and W. L. Hase, *Unimolecular reaction dynamics: theory and experiments*, (Oxford University Press, New York, 1996).

[22] K. Song and R. Spezia, *Theoretical Mass Spectrometry*, (De Gruyter, Berlin, 2018).

[23] S. Grimme, Angew. Chem. Int. Ed. **52**, 6302 (2013).

[24] C. A. Bauer and S. Grimme, J. Phys. Chem. A **120**, 3755 (2016).

[25] C. A. Bauer and S. Grimme, J. Phys. Chem. A **118**, 11479 (2014).

[26] S. Pratihar, D. G. Bhakta, S. C. Kohale, J. Laskin, and W. L. Hase, Phys. Chem. Chem. Phys. **16**, 23769 (2014).

[27] S. Pratihar, G. L. Barnes, J. Laskin, and W. L. Hase, J. Phys. Chem. Lett. **7**, 3142 (2016).

[28] O. Meroueh, Y. Wang, and W. L. Hase, J. Phys. Chem. A **106**, 9983 (2002).

[29] R. Spezia, J.-Y. Salpin, M.-P. Gaigeot, W. Hase, and K. Song, J. Phys. Chem. A **113**, 13853 (2009).

[30] R. Spezia, A. Martin-Somer, V. Macaluso, Z. Homayoon, S. Pratihar, and W. L. Hase, Faraday Discuss. **195**, 599 (2016).

[31] Z. Homayoon, S. Pratihar, E. Dratz, R. Snider, R. Spezia, G. Barnes, V. Macaluso, A. Martin-Somer, and W. L. Hase, J. Phys. Chem. A **120**, 8211 (2016).

[32] A. Malik, Y.-F. Lin, S. Pratihar, L. Angel, and W. L. Hase, J. Phys. Chem. A **123**, 6868 (2019).

[33] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich, Proc. Natl. Acad. USA **103**, 17747 (2006).

[34] K.-C. Chou, Biophys. Chem. **35**, 1 (1990).

[35] J. Koca, S. Perez, and A. Imberty, J. Comput. Chem. **16**, 296 (1995).

[36] D. J. Jabobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, Proteins: Struct. Funct. Gen. **44**, 150 (2001).

[37] Y. Yan, S. Zhang, and F.-X. We, Proteome Science **9**, S17 (2011).

[38] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).

[39] D. R. Galimberti, S. Bougueroua, J. Mahé, M. Tommasini, A. M. Rijs, and M.-P. Gaigeot, Faraday Discuss. **217**, 67 (2019).

[40] B. L. Mooney, L. R. Corrales, and A. E. Clark, J. Comput. Chem. **33**, 853 (2012).

[41] B. L. Mooney, L. R. Corrales, and A. E. Clark, J. Phys. Chem. B **116**, 4263 (2012).

[42] M. Tenney and R. T. Cygan, J. Phys. Chem. C **117**, 24673 (2013).

[43] A. Ozkanlar and A. E. Clark, J. Comput. Chem. **35**, 495 (2014).

[44] K. Han, R. M. Venable, A.-M. Bryant, C. J. Legacy, R. Shen, H. Li, B. Roux, A. Gericke, and R. W. Pastor, J. Phys. Chem. B **122**, 1484 (2018).

[45] F. Pietrucci and W. Andreoni, Phys. Rev. Lett. **107**, 085504 (2011).

[46] F. Pietrucci and W. Andreoni, J. Chem. Theory Comput. **10**, 913 (2014).

[47] S. Bougueroua, R. Spezia, S. Pezzotti, S. Vial, F. Quessette, D. Barth, and M.-P. Gaigeot, J. Chem. Phys. **149**, 184102 (2018).

[48] E. Martinez-Nunez, Phys. Chem. Chem. Phys. **17**, 14912 (2015).

[49] E. Martinez-Nunez, J. Comput. Chem. **36**, 222 (2015).

[50] A. R. an R. Rodriguez-Fernandez, S. A. Vazquez, G. L. Barnes, J. J. P. Stewart, and E. Martinez-Nunez, J. Comput. Chem. **39**, 1922 (2018).

[51] R. Perez-Soto, S. A. Vazquez, and E. Martinez-Nunez, Phys. Chem. Chem. Phys. **18**, 5091 (2016).

[52] V. Macaluso, D. Scuderi, M. E. Crestoni, S. Fornarini, D. Corinti, E. Dalloz, E. Martinez-Nunez, W. L. Hase, and R. Spezia, J. Phys. Chem. A **123**, 3685 (2019).

[53] J. A. Varela, S. A. Vazquez, and E. Martinez-Nunez, Chem. Sci. **8**, 3843 (2017).

[54] Y. Jeanvoine, A. Largo, W. L. Hase, and R. Spezia, J. Phys. Chem. A **122**, 869 (2018).

[55] A. Carrà, V. Macaluso, P. W. Villalta, R. Spezia, and S. Balbo, J. Am. Soc. Mass Spectrom. **30**, 2771 (2019).

[56] D. J. McAdoo, Mass Spectrom. Rev. **7**, 363 (1988).

[57] P. Longevialle, Mass Spectrom. Rev. **11**, 157 (1992).

[58] F. A. L. Mauguiere, P. Collins, Z. C. Kramer, B. K. Carpenter, G. S. Ezra, S. C. Farantos, and S. Wiggins, Annu. Rev. Phys. Chem. **68**, 499 (2017).

[59] J. M. Bowman and P. L. Houston, Chem. Soc. Rev. **46**, 7615 (2017).

[60] P. Y. I. Shek, J. K.-C. Lau, J. Zhao, J. Grzetic, U. H. Verkerk, J. Oomens, A. C. Hopkinson, and K. W. M. Siu, Int. J. Mass Spectrom. **316-318**, 199 (2012).

[61] A. R. Dongré, J. Jones, A. Somogyi, and V. H. Wysocki, J. Am. Chem. Soc. **118**, 8365 (1996).

[62] V. H. Wysocki, G. Tsaprailis, L. Smith, and L. Breci, J. Mass. Spectrom. **35**, 1399 (2000).

[63] R. Boyd and A. Somogyi, J. Am. Soc. Mass Spectrom. **21**, 1275 (2010).

[64] E. Rossich Molina, D. Ortiz, J.-Y. Salpin, and R. Spezia, J. Mass Spectrom. **50**, 1340 (2015).

[65] A. Pérez-Mellor, I. Alata, V. Lepere, R. Spezia, and A. Zehnacker-Rentien, Int. J. Mass Spectrom. **465**, 116590 (2021).

[66] A. Pérez-Mellor, K. L. Barbu-Debus, V. Lepere, I. Alata, R. Spezia, and A. Zehnacker-Rentien, Europ. Phys. J. D **75**, 165 (2021).

[67] G. B. Rocha, R. O. Freire, A. M. Simas, and J. J. P. Stewart, J. Comput. Chem. **27**, 1101 (2006).

[68] S. Grimme, J. Comput. Chem. **25**, 1463 (2009).

[69] G. L. Barnes and W. L. Hase, J. Am. Chem. Soc. **131**, 17185 (2009).

[70] G. L. Barnes, K. Young, L. Yang, and W. L. Hase, J. Chem. Phys. **134**, 094106 (2011).

[71] W. L. Hase and D. G. Buckowski, Chem. Phys. Lett. **74**, 284 (1980).

[72] W. L. Hase, R. J. Duchovic, X. Hu, A. Komornicki, K. F. Lim, D.-H. Lu, G. H. Peslherbe, K. N. Swamy, S. R. V. Linde, A. Varandas, H. Wang, and R. J. Wolf, QCPE Bull. **16**, 671 (1996).

[73] J. J. P. Stewart, L. K. Fiedler, J. Zheng, I. Rossi, W.-P. Hu, G. C. Lynch, Y.-P. Liu, P. Zhang, Y.-Y. Chuang, and J. P. et al., “MOPAC, version 5.022mn based on MOPAC 5.0,” (2015).

[74] N. Trinajstic, *Chemical Graph Theory* (CRC Press, Boca Raton, FL, 2000).

[75] B. D. McKay and A. Piperno, J. Symbolic Comput. **60**, 94 (2014).

[76] S. Even, *Graph Algorithms*, 2nd ed., edited by G. Even (Cambridge University Press, 2011).

[77] N. M. Rodiguin and E. H. Rodiguina, *Consecutive Chemical Reactions* (D. Van Nostrand Co., New York, 1964).

[78] R. F. Grote and J. T. Hynes, J. Chem. Phys. **73**, 2715 (1980).

[79] Z. Gregg, W. Ijaz, S. Jannetti, and G. L. Barnes, J. Phys. Chem. C **118**, 22149 (2014).

[80] D. Ortiz, P. Martin-Gago, A. Riera, K. Song, J.-Y. Salpin, and R. Spezia, Int. J. Mass Spectrom. **335**, 33 (2013).

[81] R. Spezia, J. Martens, J. Oomens, and K. Song, Int. J. Mass Spectrom. **388**, 40 (2015).

[82] R. Spezia, S. B. Lee, A. Cho, and K. Song, Int. J. Mass Spectrom. **392**, 125 (2015).

[83] A. Martin-Somer, J. Martens, J. Grzetic, W. L. Hase, J. Oomens, and R. Spezia, J. Phys. Chem. A **122**, 2612 (2018).

[84] S. Vazquez, X. L. Otero, and E. Martinez-Nunez, Molecules **23**, 3156 (2018).