ReacNetGenerator: an automatic reaction network generator for reactive molecular dynamics simulations.

Zeng, Jinzhe; Cao, Liqun; Chin, Chih-Hao; et.al.

Zeng, J., Cao, L., Chin, C.-H., Ren, H., Zhang, J. Z. H., & Zhu, T. (2020). ReacNetGenerator: an automatic reaction network generator for reactive molecular dynamics simulations. In Physical chemistry chemical physics (Vol. 22, Issue 2, pp. 683–691). Royal Society of Chemistry . https://doi.org/10.7282/00000188
Document Version: Accepted Manuscript (AM)

ReacNetGenerator: An Automatic Reaction Network Generator for Reactive Molecular Dynamic Simulations

Jinzhe Zeng[1†], Liqun Cao[1†], Chih-Hao Chin[1,2*], Haisheng Ren[3*], John Z.H. Zhang[1,2,4] and Tong Zhu[1,2*]

[1]*Shanghai Engineering Research Center of Molecular Therapeutics & New Drug Development, School of Chemistry and Molecular Engineering, East China Normal University, Shanghai, 200062, China*

[2]*NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai, 200062, China*

[3]*School of Chemical Engineering, Sichuan University, Chengdu 610065, China*

[4]*Department of Chemistry, New York University, New York 10003, United States*

[†]These authors contribute equally.

ABSTRACT

Reactive molecular dynamics (MD) simulation makes it possible to study the reaction mechanism of complex reaction systems at the atomic level. However, the analysis of the MD trajectories which contain thousands of species and reaction pathways has become a major obstacle to the application of reactive MD simulation in large-scale systems. Here, we report the development and application of the Reaction Network Generator (ReacNetGenerator) method. It can automatically extract the reaction network from the reaction trajectory without any predefined reaction coordinates and elementary reaction steps. Molecular species can be automatically identified from the cartesian coordinates of atoms and the hidden Markov model is used to filter the trajectory noises which makes the analysis process easier and more accurate. The ReacNetGenerator has been successfully used to analyze the reactive MD trajectories of the combustion of methane and 4-component surrogate fuel for rocket propellant 3 (RP-3), and it has great advantages in efficiency and accuracy compared to traditional manual analysis.

View Article Online
DOI: 10.1039/C9CP05091D

1 Introduction

In the past decades, reactive molecular dynamics (MD) simulation has been widely used to study the reaction mechanism of many complex molecular systems, such as combustion, explosion, and heterogeneous catalysis. Among all reactive MD method, *ab initio* molecular dynamics simulation is undoubtedly the most rigorous and accurate one[1]. Unfortunately, the computational cost inherent in the QM calculation severely limits the simulation scale of the AIMD method. With the rapid development of computer hardware and algorithms, some AIMD methods have recently begun to handle larger scale systems. Wang *et al.*[2, 3] had proposed an *ab initio* nanoreactor and used it to explore the formation pathways of glycine in the atmosphere of early Earth. The nanoreactor adopts graphics processing units (GPUs) to accelerate the electronic structure calculation, thus has higher efficiency than conventional QM calculations based on CPUs.

On the other hand, QM data can be used to train semi-empirical QM methods and empirical force fields, such as the density-functional based tight-binding (DFTB[4]) method and the reactive force field (ReaxFF[5-9]). These methods trade accuracy for the lower computational expense, making it possible to reach simulation scales that are orders of magnitude beyond what is tractable for conventional QM methods. By using the DFTB method, Zhang and co-workers studied the early decay mechanism of the shocked ε-2,4,6,8,10,12-hexanitro-2,4,6,8,10,12-hexaazaisowurtzitane (CL-20)[10], Lei et al. simulated the reaction process of graphene synthesis via detonation at different oxygen/acetylene mole ratios[11]. Compared with the DFTB method, the reactive force field (ReaxFF[12]) has a more obvious advantage in efficiency. By using reactive MD with ReaxFF, Li and co-workers have studied the pyrolysis of Liulin coal. The modeled system contains 28351 atoms, and the simulation was performed for 250ps[13]. In addition, Han et al. performed long-time ReaxFF MD simulations of fuel-rich combustion for up to 10 ns to explore the initial formation mechanism of soot nanoparticle[14].

Large molecular system and long simulation time can produce complicated MD trajectories which contain a great number of reaction events and molecular species. Manual analysis of such trajectories is unrealistic, which motivates the development of computational algorithms that can analyze these trajectories automatically. Liu et al. has proposed the VARxMD (Visualization and Analysis of Reactive Molecular Dynamics) software which can analyze and visualize reactions from the atomic coordinates and bond order information from the MD trajectory with ReaxFF[15]. Dontgen et al. also developed the ChemTraYzer (Chemical Trajectory Analyzer) software for postprocessing and analyzing of the ReaxFF trajectories[16]. In the preparation of this article, a new method is presented based on ChemTraYzer[17], which can give more accurate rate constants and construct a skeletal reaction

View Article Online
DOI: 10.1039/C9CP05091D

network. Recently, Martínez and co-workers proposed a method called Nebterpolation which can discover and refine reaction pathways from the AIMD trajectory[3].

Despite the success of these methods, they are either not easily accessible due to license issue, or are not user-friendly, and still require a large amount of human resources to participate in the analysis. In this study, a new tool called ReacNetGenerator (Reaction Network Generator) was developed. It takes atomic coordinates as the only necessary input, and can automatically extract rich information from the trajectory, such as molecular species and reaction events. Reaction networks can be constructed based on this information and be interactively explored by the user. The paper is organized as follows: the algorithms of ReacNetGenerator is described in detail in the next section. In Section 3, the ReacNetGenerator was successfully used to uncover the reaction mechanism of the combustion of methane and the 4-component RP-3 jet fuel. And finally, a conclusion is given in Section 4.

2 Computational Methods

2.1 Simulation details

In this work, two fuel oxidation systems were constructed by using the Amorphous Cell module in the Materials Studio software package[18]. One is the methane oxidation system and the other is a four-component surrogate model of the RP-3 jet fuel[19] (42% *n*-dodecane, 40% *n*-decane, 13% ethylcyclohexane, and 5% *p*-xylene, as shown in Fig. S1). A specified amount of oxygen molecules that can completely oxidize the fuel molecules was added to the system. For methane, the simulated system is a 3-dimensional periodic box containing 50 CH_4 and 100 O_2 molecules with a bulk density of 0.25g/cm^3. For RP-3, the simulated system contains 1606 molecules and 6490 atoms, and the density is 0.5g/cm^3 (Fig. S2). Specific densities were chosen to be consistent with the previous studies[20, 21] to facilitate the comparison.

Both systems are firstly equilibrated at 298K for 250ps using the Forcite module in Materials Studio to get a reasonable initial configuration. Then a 2.5ns MD simulation with ReaxFF was performed for each system by using the LAMMPS package[22]. The parameters of Chenoweth et al. (the CHO-2008 parameter set) were employed[23]. The temperature was set to 3000 K. The time step was set to 0.1fs, and the coordinates of atoms were recorded every 10fs. The NVT ensemble was used for both the equilibration and production run. The Berendsen thermostat was used and the damping parameter was set to 100fs. Then the trajectories of entire processes were used to analyze in subsequent steps.

2.2 The ReacNetGenerator method

View Article Online
DOI: 10.1039/C9CP05091D

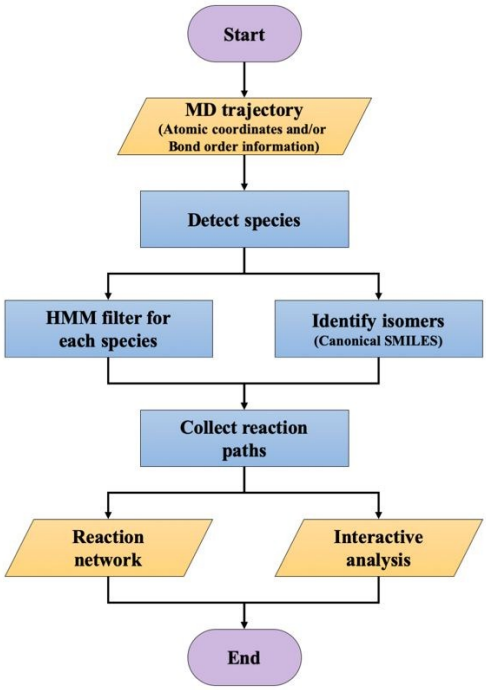


Figure. 1 The flowchart of the ReacNetGenerator.

The flowchart of the ReacNetGenerator method is shown in Fig. 1, which consists of several modules and algorithms. The key input of ReacNetGenerator is the MD trajectory from either ReaxFF MD or AIMD simulation, and the bond order information from ReaxFF MD simulation can also be an additional input. After reading these input files, the connectivity of the atoms in each snapshot is determined firstly from the coordinates and/or bond orders. Then the species (including molecules and radicals) are detected according to the atomic connectivity. However, reactive MD simulation normally contains large-amplitude molecular vibrations and collisions, thus it is very rough to use the distance between atoms to judge the existence of chemical bonds. Many "noise" molecules that are unstable or even unreasonable in energy or structure will be detected from the first step. In their previous works[2, 3], Wang et al. filtered these noises by using a two-state hidden Markov model (HMM). We also adopted this algorithm in the ReacNetGenerator. To facilitate the analysis, all detected species are indexed by canonical simplified molecular input line entry specification (SMILES) to guarantee its uniqueness. Isomers are also identified among species according to SMILES. Then the path and quantity of reactions in the whole trajectory are calculated. With the reaction matrix generated, the force-directed algorithm will be used to construct a reaction network. All the results including the network, species, and reactions are put in an interactive web page, and the user can analyze the local reaction network for any given species by mouse-clicking. Below are details of several key procedures:

View Article Online
DOI: 10.1039/C9CP05091D

2.2.1 Acquisition of species information

An MD trajectory that contains atomic coordinates is the only necessary input of ReacNetGenerator. Then the Open Babel software[25] will be used to convert the atomic coordinates to bond information. In Open Babel, a chemical bond will be assigned to two atoms if their distance d satisfies the following criteria:

$$0.4\text{Å} \leqslant d \leqslant r_i + r_j + 0.45\text{Å} \quad (1).$$

Where r_i and r_j are atomic covalent radius of atom *i* and *j*. The bond type, bond order, hybridized center, and aromaticity are then determined by the local bond geometry. The periodic boundary condition is considered by slightly modify the source code of Open Babel. As a result, ReacNetGenerator can process trajectory from any kind of reactive MD simulation, such as AIMD and ReaxFF MD. The using of Open Babel can be skipped if additional bond order information is provided by the ReaxFF MD simulation. After the bond information is obtained, each snapshot in the trajectory can be treated as an undirected graph. Each atom in the snapshot is a node in the graph and each chemical bond can be treated as an edge. Then a molecular fragments (species) can be considered as a subgraph. Identifying all species can be considered as a graph traversal problem in the graph theory. A general graph traversal algorithm, called Depth-first search[26], is used in this study. By using this way, all species in a given trajectory can be obtained.

2.2.2 Filtering "noise" by HMM

As mentioned above, a lot of species detected from the last step are useless for analyzing. To accelerate the analysis, the existence of species is firstly converted into 0-1 signals, and a two-state HMM[3, 27] is adopted in the ReacNetGenerator to smooth the existence signal.

The HMM model can be described as a transition matrix **A**, an emission matrix **B**, and an initial state vector π. The raw (existence) signal of a given species can be considered as a visible output sequence O^s:

$$O^s = \{o_1^s, o_2^s, \cdots o_t^s \cdots o_T^s\} \quad (2),$$

where the superscript *s* represents a given species, and the subscript *t* represents time (or the t^{th} frame in the trajectory), there are a total of *T* frames in the trajectory. o_t^s can be any value in set *V*:

$$V = \{v_1 = 1, v_2 = 0\} \quad (3),$$

where 1 means the species exists, and 0 means no.

Similarly, the noise-filtered signal we expect can be considered as the hidden sequence H^s in the HMM:

$$H^s = \{h_1^s, h_2^s, \cdots h_t^s \cdots h_T^s\} \quad (4),$$

and h_t^s can be any value in set Q:

$$Q = \{q_1 = 1, q_2 = 0\} \quad (5).$$

The mathematical description of **A**, **B**, and π is given by

$$\begin{cases} \boldsymbol{A} = [a_{ij}]_{2\times 2} \\ \boldsymbol{B} = [b_j(k)]_{2\times 2} \\ \boldsymbol{\pi} = (\pi_1, \pi_2) \end{cases} \quad (6).$$

where $a_{ij} = P(h_{t+1}^s = q_j \mid h_t^s = q_i), i = 1,2, j = 1,2$ is the probability of the transition from state q_i at t to q_j at $t+1$. $b_j(k) = P(o_t^s = v_k \mid h_t^s = q_j), k = 1,2, j = 1,2$ is the probability of observing v_k when the hidden state is q_j. $\pi_i = P(h_1^s = q_i), i = 1,2$ is the start probability for state q_i.

The next task is to find the most likely hidden sequence H^s (called Viterbi path) that can produce the observation sequence O^s:

$$Viterbi\ path = \max_{H^s} P(H^s, O^s) \quad (7),$$

where

$$P(H^s, O^s) = P(h_1^s = q_i) \prod_{t=1}^{T} P(h_t^s = q_j \mid h_{t-1}^s = q_i) P(o_t^s = v_k \mid h_t^s = q_j) \quad (8).$$

In order to obtain the Viterbi path, a dynamic programming algorithm called the Viterbi algorithm[28], which has been widely used in telecom decoding, deep-space communications, speech recognition, and bioinformatics is employed.

Note that $P(h_t^s = q_j \mid h_{t-1}^s = q_i)$, $P(o_t^s = v_k \mid h_t^s = q_j)$, and $P(h_1^s = q_i)$ are respectively values of **A**, **B**, and π, so the Viterbi path is determined by the parameters of the HMM model. Choosing smaller values in the off-diagonal of **A** will force the sequence H^s to have fewer transitions. Choosing larger values in the off-diagonal of **B** will allow the H^s to deviate more often from the O^s. Both changes can increase the filtering ability of the HMM model.

In ReacNetGenerator, the initial state π will not influence the filtering ability of the HMM model. For **A** and **B**, one can randomly select parts of the trajectory of several species for manual analysis and labeling the noise signals. Then the maximum likelihood estimation method can be used to calculate their matrix elements:

$$a_{ij} = \frac{\alpha_{ij}}{\sum_j \alpha_{ij}}, i = 1, 2; j = 1, 2 \quad (9).$$

View Article Online
DOI: 10.1039/C9CP05091D

Where α_{ij} is the frequency at which the signal transition (q_i to q_j) from time t to $t+\Delta t$ (Δt is the time interval in the trajectory.)

Similarly,

$$b_j(k) = \frac{\beta_{jk}}{\sum_k \beta_{jk}}, j = 1, 2; k = 1, 2 \quad (10).$$

Where β_{ij} is the frequency of v_k is observed form the state q_j.

In addition, users may use some off-the-shelf theory to estimate HMM parameters from the original trajectory, such as the Baum-Welch algorithm[29]. However, it must be pointed out that none of these methods give the perfect parameters for the HMM model. This is because we cannot get the species information in the trajectory in advance, and it is even less likely to mark every noise signal. A better strategy is to perform multiple rounds of analysis and gradually reduce the filtering power of the HMM model. The skeletal reaction mechanism in which the major species participate is obtained first and then the more detailed local reaction network can be analyzed for the species of interest.

The default HMM parameters in ReacNetGenerator are consistent with the previous study[3], where

$$\begin{cases} \mathbf{A} = \begin{pmatrix} 0.999 & 0.001 \\ 0.001 & 0.999 \end{pmatrix} \\ \mathbf{B} = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix} \\ \boldsymbol{\pi} = (0.5, 0.5) \end{cases} \quad (11).$$

The parameters are further verified by combining manual analysis and the maximum likelihood estimation. Usually, only signals with a lifetime shorter than 100 fs are filtered out. In our simulation, the atomic coordinates are saved every 10fs. This setting is widely adopted by MD simulations with ReaxFF, while AIMD usually uses smaller time intervals to store coordinates. Therefore, we think the default HMM parameters are a good starting point for analysis.

A concern with the HMM approach is that it might filter out actual reaction events which are rare. However, when faced with the trajectory that contains thousands of species and nanosecond lengths, users are more concerned with the main reaction mechanisms. While the HMM method can greatly reduce the complexity of the trajectory analysis. In the future study, we will introduce energy criteria to validate whether there is a reaction and combine the machine learning methods to make the analysis results more accurate.

2.2.3 Construction of the reaction network

To construct a reaction network, each kind of species should be treated as a node in the network. Therefore, all detected species are indexed by canonical SMILES[30] to guarantee its uniqueness. Isomers are also identified according to SMILES codes. In order to improve the accuracy and efficiency of isomer recognition, the VF2 algorithm[31] is also provided. VF2 is a depth-first backtracking algorithm that is widely used to solve the subgraph isomorphism problem. After filtering out noises, the reaction paths and the number of intermolecular reactions can be calculated. A reaction matrix can be generated as

$$\mathbf{R} = [r_{ij}], i = 1,2,...,N; j = 1,2,...,N \quad (12),$$

where r_{ij} is the number of reactions from species s_i to s_j. Finally, with the reaction matrix, a reaction network can be constructed. Here the NetworkX package[32] is used to make a graph which indicates the reactions between species. Fruchterman-Reingold force-directed algorithm[33] is used to make the layout of nodes relate to the number of reactions, which can ensure that the graph is more legible and easy to read. The distance of two species in the network, the color and thickness of the line connected them are determined by the number of reactions between them, making the reaction network more intuitive.

Here three options are offered: a) showing the reaction network formed by top N (an integer given by the user, 20 by default) species which have the most reactions; b) showing the reaction network starting with reactants and contains N species which have the most reactions with them; c) showing the reaction network formed by top N species which contain specific elements and have the most reactions.

It must be noted that the reaction network shown in this way is only one of the basic outputs of ReacNetGenerator, and users can get more detailed information. As mentions above, the reaction network along with all species and reactions will be shown on an interactive web page. A filter is provided to the user to study reactions that involve specific species. In addition, when the user clicks on any of the species, ReacNetGenerator will show the top 5 species which have the most reaction with it. If the user double clicks on a species, it will be temporarily hidden. Therefore, with simple mouse-clicking, users can explore the reactions of any species they are interested in and build a corresponding reaction network. The user can save the network as a picture for further analysis or presentation at any time during the analysis.

3 Results and Discussion

3.1 Identify reaction pathways and noise filtering

The input of the ReacNetGenerator is the trajectory from a reactive MD simulation. The trajectory contains many individual reaction events, but in the beginning, we don't know when the reaction occurred and which atoms participated in the reaction. Thus, the first task is to

identify and extract these reaction events from the MD trajectory. However, as mentioned above, it is very rough to use the distance between atoms to judge the existence of chemical bonds as reactive MD trajectories usually contain large-amplitude molecular vibrations and collisions. Therefore, the hidden Markov model is employed to filter out the "noise" from the trajectory and make it easier to analyze. Fig. 2 shows two examples of the HMM filtering process. As can be seen, the propyl radical appears very frequently during certain periods of the trajectory, but only occasionally appear in other time periods, the HMM signal accurately reflect its existence. On the other hand, the HOHOOH radical is more like a water molecule and a hydrogen peroxide closely contact with each other by collision, its signal is very sparse and the lifetime is very short, so the HMM successfully filtered it out.

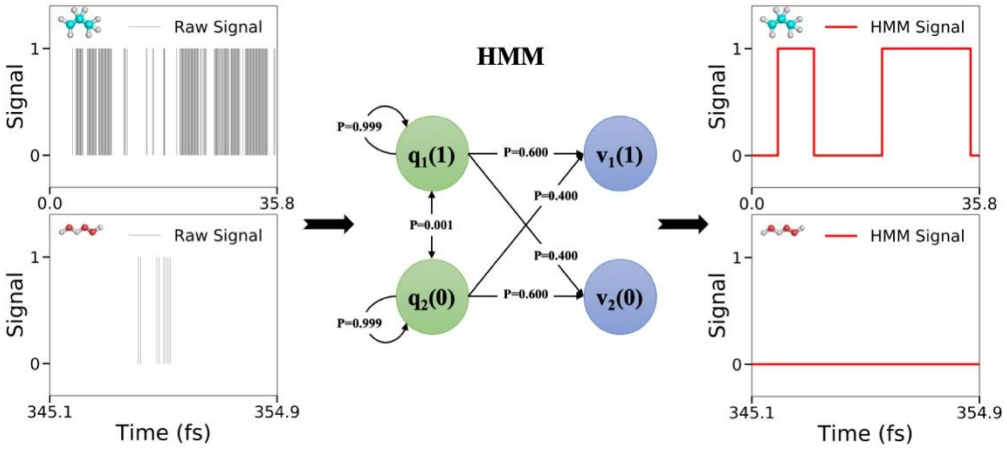


Figure 2. Existence signals of two species processed by the hidden Markov model.

After the HMM process, the connectivity of all atoms is tracked to detect all reaction events in the trajectory.

3.2 Analysis of the MD trajectory of methane combustion.

As the first example, we analyzed a ReaxFF MD trajectory of methane combustion reactions. Fig. 3 and Fig. S3 show the time evolution of the number of important reactants, intermediates and products observed during the simulation. After about 500ps, several species such as CH_2O, H_2O, CH_3OH, CO, CO_2, and CH_3OOH were formed. The main reactions completed after about 1.5ns. The number of reactants and main products almost maintained constant. Then the trajectory was processed by the ReacNetGenerator. Depending on the needs of the user, species with specific elements can be selected for analysis, such as analysis of reactions between all carbon-containing species. The output of ReacNetGenerator consists of three parts: a reaction network, the list of all species, and the list of reactions between species.

All of the information is contained in an interactive web page, making the results clearer and more intuitive.

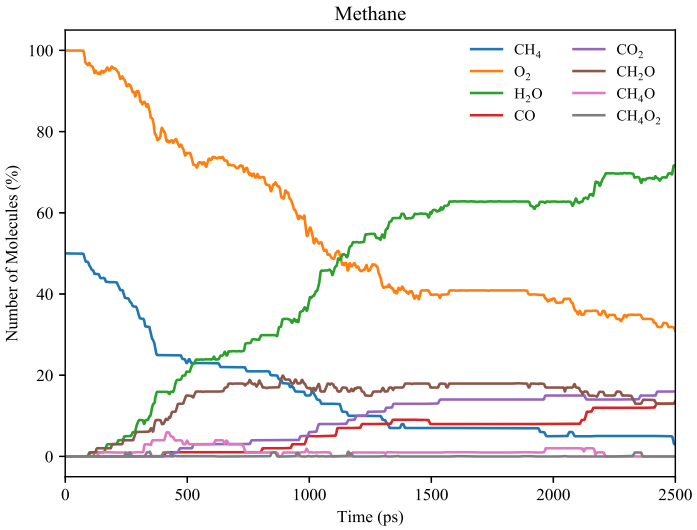


Figure 3. The number of important reactants, products, and intermediates evolved with simulation time.

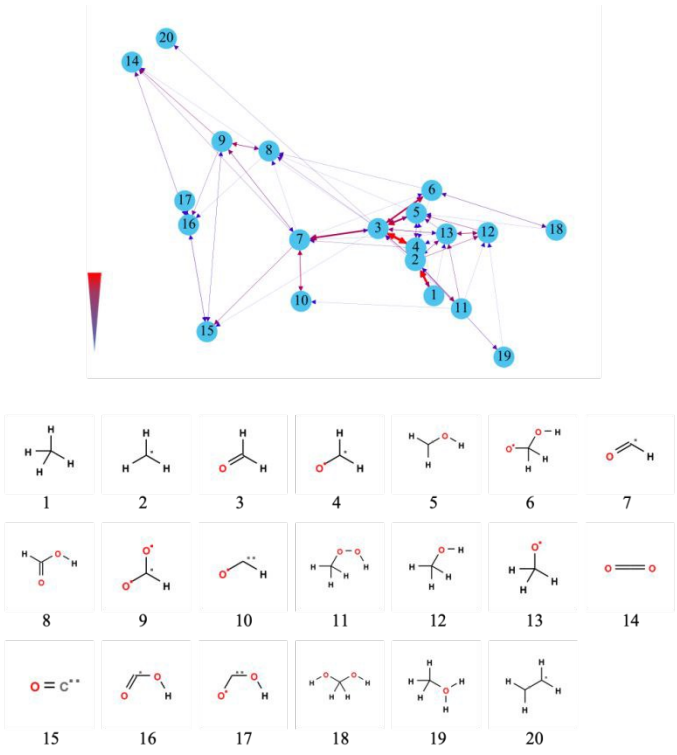


Figure 4. The reaction network formed by the top 20 species which contain carbon and have the most reactions in the trajectory of the methane oxidation. Species are shown in blue spheres, and chemical reactions are indicated using colored arrows. The closer of the spheres, the thicker of the line, the redder of the color indicates the more reactions between species.

View Article Online
DOI: 10.1039/C9CP05091D

Fig. 4 shows one such representation of a reaction network derived from the ReacNetGenerator based on the trajectory of methane oxidation. In the analysis, only the species which contain carbon were considered. In the reaction network, each species is represented by a blue sphere. The colored arrows linked the species represent the reactions between them, the closer of the sphere, the thicker of the line, the redder of the color indicates the more reactions between species. One can find a number of reaction paths from methane to carbon dioxide from Fig. 4, which are highly consistent with the previous work[20], in which the reactive MD simulation was performed under similar conditions and the manual analysis was used.

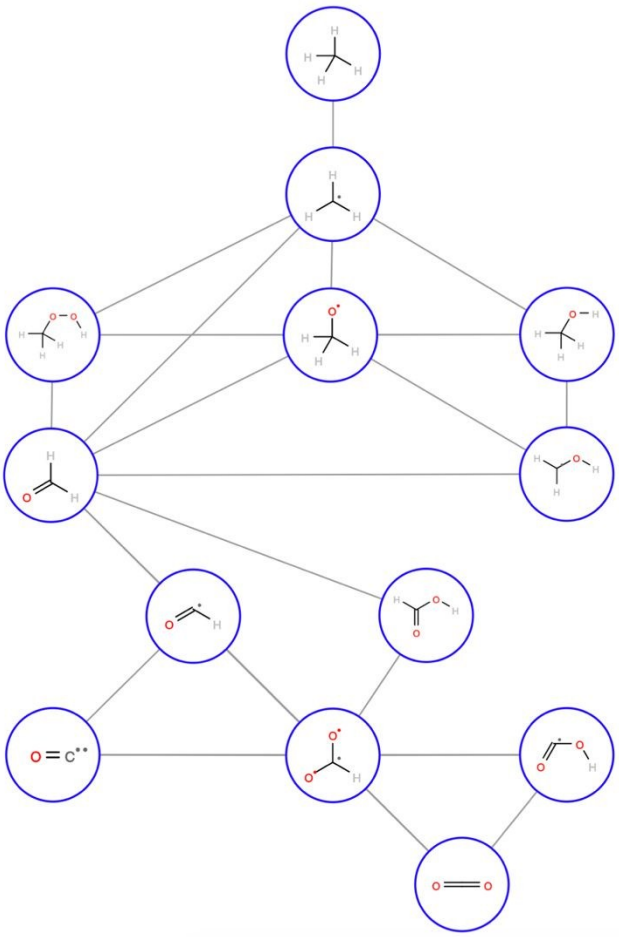


Figure 5. Reaction paths from methane to carbon dioxide derived from ReacNetGenerator.

In addition, if we are interested mainly in a particular species, we can map out the local reaction network of closely related compounds (namely, the species that appear on either side of chemical equations that lead to the species of interest) by the mouse clicking (Video S1). Focusing on specific species also allows us to trace the reaction pathways that lead from the starting molecules. Formaldehyde is a key intermediate that participates in the oxidation of

methane. Fig. 5 shows several reaction paths from CH_4 to CO and CO_2 obtained by interactive mouse clicking. In one path, $CH_4 \rightarrow \bullet CH_3 \rightarrow CH_3O\bullet \rightarrow CH_2OH \rightarrow H_2CO \rightarrow HCO\bullet \rightarrow CO$, the methyl radical ($\bullet CH_3$) is oxidized to the formaldehyde, which then loses a hydrogen atom to form the formyl ($HCO\bullet$) radical. The formyl radical can subsequently lose a hydrogen atom via collisional dissociation or reaction with molecular oxygen, thereby forming the carbon monoxide (CO). As can be seen, formaldehyde and $CH_3O\bullet$ are among the most highly connected compounds in the reaction network, participating in more than 10 reactions with other species. These highly connected compounds have in common the ability to take part in several different reaction types. For example, formaldehyde is found to participate in proton transfer, nucleophilic addition and dihydrogen reactions. These findings are consistent with the previous study[20] and experiments[34], but ReacNetGenerator is more intuitive and saves a lot of time than manual analysis.

3.3 Analysis of the MD trajectory of RP-3 combustion.

As a second example, the ReacNetGenerator was used to analyze the trajectory of the combustion of a four-component surrogate model of the RP-3 jet fuel.

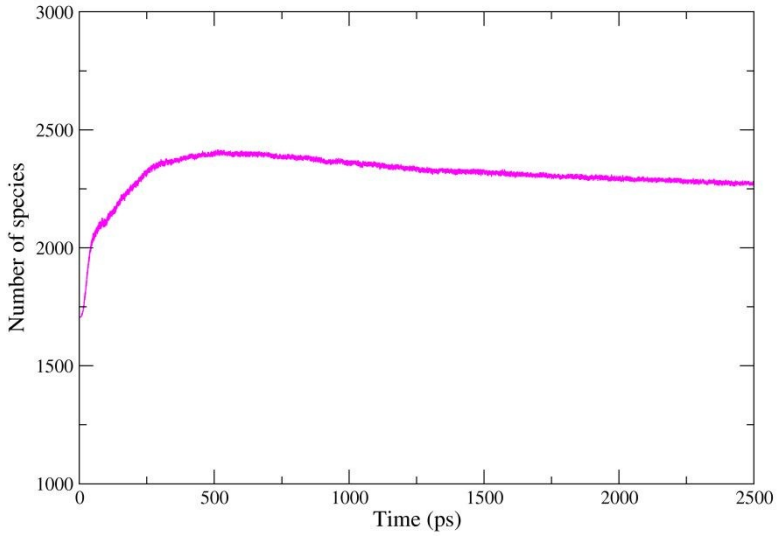


Figure 6. The change in the number of species during the MD simulation.

Fig. 6 shows the change in the number of species during the MD simulation. For such a medium-sized system, the number of species in some snapshots has exceeded 2,400. Manual extraction of reaction events from tens of thousands of snapshots has been an almost impossible task, thus automated methods must be developed. At the same time, the introduction of "noise" filtering is also very necessary.

Fig. 7 depicts reaction networks of RP-3 with and without the HMM filter, respectively. For each situation, 20 species participated in the most reactions are taken to construct the network. As can be seen, the network without HMM is relatively simple, mainly composed of "reversed motions" between two species. For example, the reaction path between species 5 and 6, 9 and 10 only reflect the forming and breaking of a double bond. In addition, species 11 is more like a close connection formed by the temporary collision of a water molecule and a CO_2. In contrast, the HMM filtered most of the "noise" species, which makes the network more reasonable, and contains more useful information.

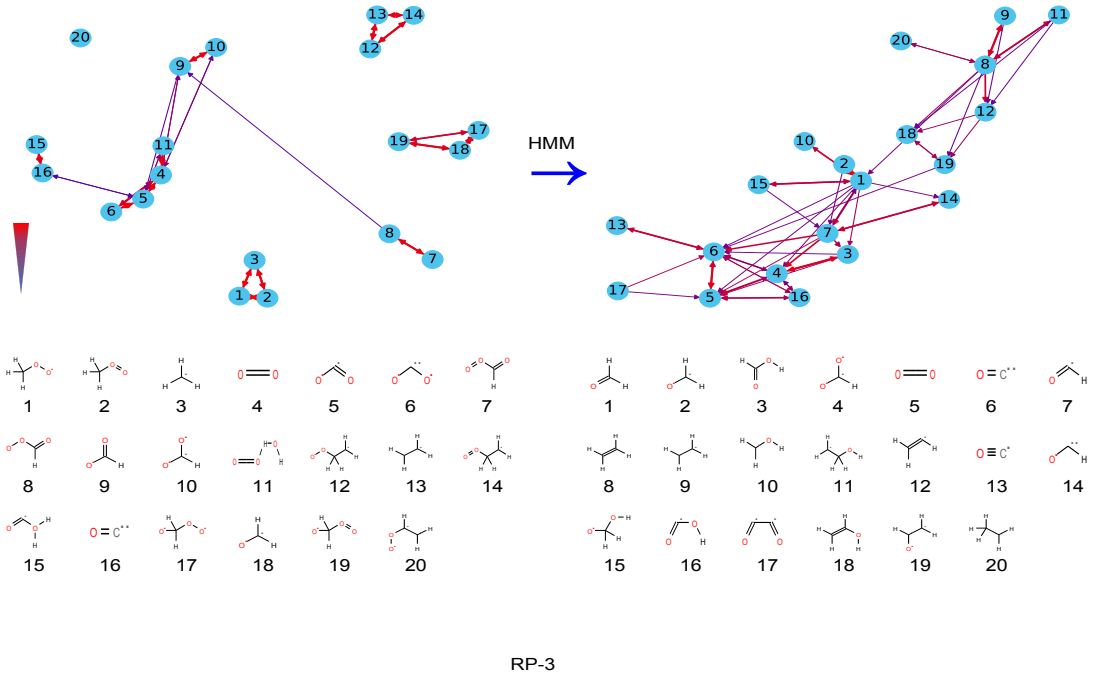


Figure 7. The reaction network for a 2.5ns reactive MD simulation of 4-compont RP-3 oxidation with ReaxFF. The left panel is the network without HMM filtering and the right one is the network with HMM.

View Article Online
DOI: 10.1039/C9CP05091D

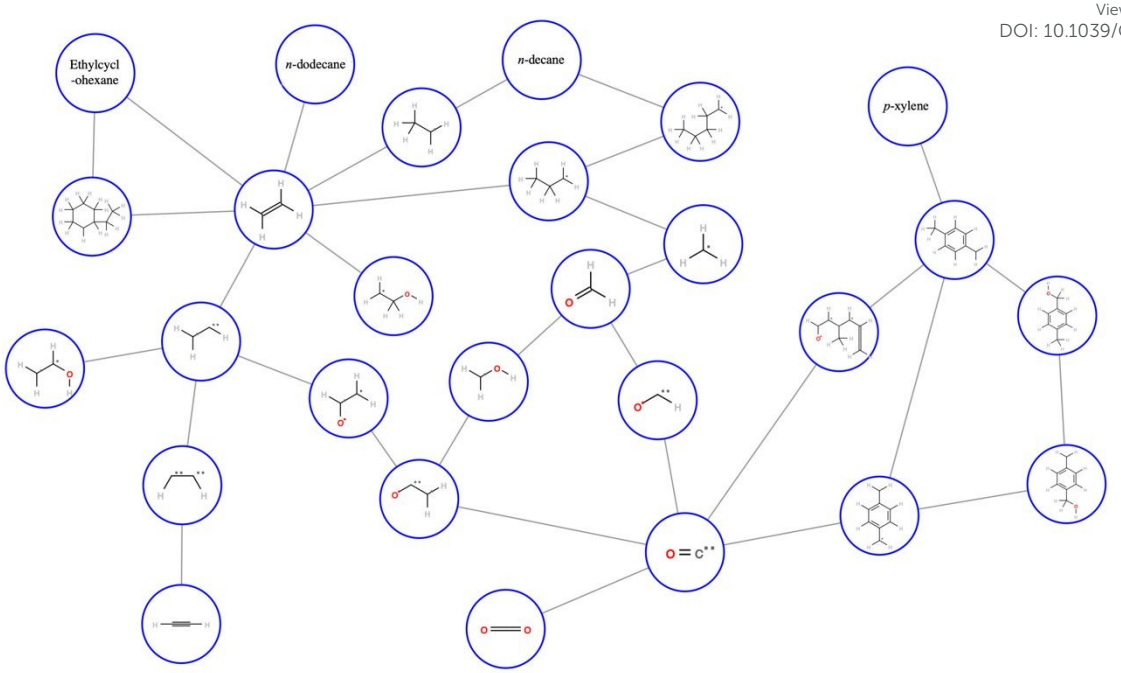


Figure 8. Reaction routes of RP-3 derived from ReacNetGenerator, which is modified slightly. For the sake of clarity, the name of the reactants instead of their 2D diagrams are used. When using ReacNetGenerator, users can easily zoom the web page which contains the analysis results, all pictures on the web page are vector graphics to ensure clarity.

Several reaction paths in the trajectory are detected on the result-containing web page by mouse-clicking and are shown in Fig. 8. The distribution of reactants and products during the MD simulation is shown in Fig. S4. Referring to these two figures, one can find that the first hydrocarbon to decompose was the *n*-dodecane, which underwent homolytic bond cleavage to form the ethene and ethenyl radical species. The second hydrocarbon to decompose was the *n*-decane, which underwent homolytic bond cleavage to form the allyl and pentyl radical species. The third one was the ethylcyclohexane, which underwent homolytic bond cleavage to form the ethylene and ethenyl radical species. Several previous studies[35,36,37,38] indicate that under high-temperature conditions the decomposition pathway may be more important than hydrogen abstraction, which is consistent with the current results.

Hydrogen dissociation is also observed from the methyl group of *p*-xylene to form hydroperoxyl and xylyl radicals. The methyl group of toluene was found to react with O_2, which is consistent with the previous study[39]. *p*-Xylene was the last hydrocarbon to react in this simulation, the activation was facilitated by a hydroperoxy radical, which results in the formation of 4-methylbenzyl and a $CH_3Ph(OH)CH_2$ radical. Since this system was relatively rich in hydrocarbons, the initiation reactions observed involve both decomposition and

reactions with radicals in the system instead of direct oxidation of the hydrocarbons by molecular oxygen as observed in the lean fuel cases.

Fig. S5 shows that the amount of ethylene increases rapidly at the beginning of the trajectory and then slowly decreases, corresponding to the rapid consumption of long-chain alkanes (Fig. S4). At the same time, the number of formaldehydes also keep increasing until about 700ps, and then slowly decreased. These trends are consistent with experimental observations.[40] According to Fig. 8, ethylene is the highest connected compound in the reaction network, participating in more than 10 reactions with other species. Ethylene is then oxidized to formaldehyde, which then loses a hydrogen atom to form formyl radical and it can subsequently lose a hydrogen atom via collisional dissociation or reaction with molecular oxygen, thereby forming carbon monoxide or carbon dioxide.

4. Conclusions

In this study, the ReacNetGenerator method was developed to automatically process the trajectory of the reactive MD simulations, uncovering the detailed reaction events. This method has a huge advantage in efficiency compared to manual analysis and can be particularly useful for large molecular systems. Compared with existing methods, ReacNetGenerator has the following characteristics and advantages:

Firstly, ReacNetGenerator is open-source and can be obtained and modified freely (*https://github.com/tongzhugroup/reacnetgenerator*). Secondly, an MD trajectory that contains atomic coordinates is the only necessary input, which makes the ReacNetGenerator not only suitable for MD simulations based on reactive force fields but also for AIMD simulations. Thirdly, ReacNetGenerator employed the HMM algorithm to filter out unreal species caused by "reversed motions" and instantaneous collisions. In using this way, real reactions in the simulation are highlighted and the analysis time is obviously saved. Next, all the analysis results are put on an interactive web page that is easy to use. Users can use the mouse clicking to carefully analyze the local reaction network of any given species. Finally, ReacNetGenerator supports parallel computing. A lot of optimizations in its speed and memory usage were performed, including data compressing and C bindings, so it is very efficient and can be used to analyze large trajectories.

Combined with relatively macroscopic data, such as the evolution of the number of species over time, the method was successfully used in the analysis of the MD trajectories of the oxidation of methane and a 4-component RP-3 fuel, and the results were highly consistent with previous theoretical studies and experimental measurements.

Further improvements to the ReacNetGenerator method will focus on two aspects. Firstly, the function of calculating the reaction rate statistically will be included. Secondly, although the HMM can filter out the "noise" species in the trajectory, there are still some false positive reactions, and the HMM may also filter out some important intermediates. We will introduce energy criteria to validate whether there is a reaction and combine the machine learning methods to make the analysis results more accurate. Related research is ongoing in our laboratory.

Although no perfect, we believe that the current version of ReacNetGenerator can already assist the analysis of complex reactive MD trajectories and contribute to our understanding of reaction mechanisms in the area of combustion, catalysis, and astrochemistry.

View Article Online
DOI: 10.1039/C9CP05091D

Conflicts of interest

There are no conflicts to declare.

Corresponding Author

*E-mail: zhjin@chem.ecnu.edu.cn

*E-mail: renhs@scu.edu.cn

*E-mail: tzhu@lps.ecnu.edu.cn

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. 91641116). J. Zeng was partially supported by the National Innovation and Entrepreneurship Training Program for Undergraduate (201910269080). We also thank the ECNU Multifunctional Platform for Innovation (No. 001) for providing supercomputer time.

References

1. M. E. Tuckerman, *J. Phys.: Condens. Matter*, 2002, **14**, 1297-1355.
2. L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande and T. J. Martínez, *Nat. Chem.*, 2014, **6**, 1044.
3. L.-P. Wang, R. T. McGibbon, V. S. Pande and T. J. Martinez, *J. Chem. Theory Comput.*, 2016, **12**, 638-649.
4. D. Porezag, T. Frauenheim, T. Kohler, G. Seifert and R. Kaschner, *Phys. Rev. B*, 1995, **51**, 12947-12957.
5. A. C. Van Duin, S. Dasgupta, F. Lorant and W. A. Goddard, *J. Phys. Chem. A*, 2001, **105**, 9396-9409.
6. A. C. van Duin, Y. Zeiri, F. Dubnikova, R. Kosloff and W. A. Goddard, *J. Am. Chem. Soc.*, 2005, **127**, 11053-11062.
7. A. Strachan and E. Kober, *J. Chem. Phys.*, 2005, **122**, 054502.
8. A. Strachan, A. C. van Duin, D. Chakraborty, S. Dasgupta and W. A. Goddard III, *Phys. Rev. Lett.*, 2003, **91**, 098301.
9. G. Shchygol, A. Yakovlev, T. Trnka, A. C. van Duin and T. Verstraelen, *J. Chem. Theory Comput.*, 2019.
10. X. Xue, Y. Wen and C. Zhang, *J. Phys. Chem. C*, 2016, **120**, 21169-21177.
11. T. Lei, W. Guo, Q. Liu, H. Jiao, D. B. Cao, B. Teng, Y. W. Li, X. Liu and X. D. Wen, *J. Chem. Theory Comput.*, 2019, **15**, 3654-3665.
12. K. Chenoweth, S. Cheung, A. C. Van Duin, W. A. Goddard and E. M. Kober, *J. Am. Chem.*

View Article Online
DOI: 10.1039/C9CP05091D

Soc., 2005, **127**, 7192-7202.

13. M. Zheng, X. X. Li, J. Liu, Z. Wang, X. M. Gong, L. Guo and W. L. Song, *Energy & Fuels*, 2014, **28**, 522-534.
14. S. Han, X. Li, F. Nie, M. Zheng, X. Liu and L. Guo, *Energy & Fuels*, 2017, **31**, 8434-8444.
15. J. Liu, X. Li, L. Guo, M. Zheng, J. Han, X. Yuan, F. Nie and X. Liu, *J. Mol. Graph. Model.*, 2014, **53**, 13-22.
16. M. Döntgen, M.-D. Przybylski-Freund, L. C. Kröger, W. A. Kopp, A. E. Ismail and K. Leonhard, *J. Chem. Theory Comput.*, 2015, **11**, 2517-2524.
17. Y. Z. Wu, H. Sun, L. Wu and J. D. Deetz, *J. Comput. Chem.*, 2019, **40**, 1586-1592.
18. Dassault Systèmes BIOVIA, Materials Studio, San Diego: Dassault Systèmes, 2017.
19. D. Zheng, W.-M. Yu and B.-J. Zhong, *Acta Physico-Chimica Sinica*, 2015, **31**, 636-642.
20. Z. He, X.-B. Li, L.-M. Liu and W. Zhu, *Fuel*, 2014, **124**, 85-90.
21. X.-L. Liu, X.-X. Li, S. Han, X.-J. Qiao, B.-J. Zhong and L. Guo, *Acta Physico-Chimica Sinica*, 2016, **32**, 1424-1433.
22. S. Plimpton, *J. Comput. Phys.*, 1995, **117**, 1-19.
23. K. Chenoweth, A. C. Van Duin and W. A. Goddard, *J. Phys. Chem. A*, 2008, **112**, 1040-1053.
24. S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach*, Pearson Education Asia, Hong Kong, 2011.
25. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminform.*, 2011, **3**, 33.
26. R. Tarjan, *SIAM J. Comput*, 1972, **1**, 146-160.
27. L. R. Rabiner, *Proc. IEEE*, 1989, **77**, 257-286.
28. G. D. Forney, *Proc. IEEE*, 1973, **61**, 268-278.
29. L. R. Rabiner, *Proc. IEEE*, 1989, **77**, 257-286.
30. RDKit: Open-source cheminformatics*; http://www.rdkit.org*
31. L. P. Cordella, P. Foggia, C. Sansone and M. Vento, *ITPAM*, 2004, **26**, 1367-1372.
32. A. Hagberg, P. Swart and D. S Chult, *Exploring network structure, dynamics, and function using NetworkX*, Los Alamos National Lab.(LANL), Los Alamos, 2008.
33. T. M. Fruchterman and E. M. Reingold, *Software: Practice and Experience*, 1991, **21**, 1129-1164.
34. G.-P. Smith, D.-M. Golden, M. Frenklach, N.-W. Moriarty, B. Eiteneer, M. Goldenberg, C.-T. Bowman, R.-K. Hanson, S. Song, W.-C. Gardiner Jr., Vi.-V. Lissianski, and Z.-W. Qin, *http://www.me.berkeley.edu/gri_mech/*.
35. X. You, F. N. Egolfopoulos and H. Wang, *Proc. Combust. Inst.*, 2009, **32**, 403-410.
36. Z. Zhao, J. Li, A. Kazakov, F. L. Dryer and S. P. Zeppieri, *Combust. Sci. Technol.*, 2004, **177**,

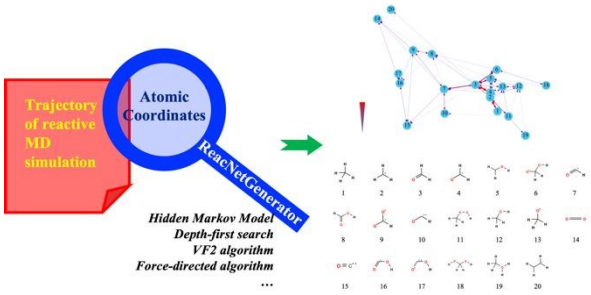
89-106.

37. Q.-D. Wang, X.-X. Hua, X.-M. Cheng, J.-Q. Li and X.-Y. Li, *J. Phys. Chem. A*, 2012, **116**, 3794-3801.

38. Z. Wang, L. Zhao, Y. Wang, H. Bian, L. Zhang, F. Zhang, Y. Li, S. M. Sarathy and F. Qi, *Combust. Flame*, 2015, **162**, 2873-2892.

39. X.-M. Cheng, Q.-D. Wang, J.-Q. Li, J.-B. Wang and X.-Y. Li, *J. Phys. Chem. A*, 2012, **116**, 9811-9818.

40. G. WC Jr, *Gas-phase combustion chemistry*, Springer Science & Business Media, New York, 1999.



TOC: The ReacNetGenerator program can automatically extract reaction information from reactive MD trajectory and construct reaction networks.