ReacNetGen: an Automatic Reaction Network Generator for Reactive Molecular Dynamic Simulations

Jinzhe Zeng[1†], Liqun Cao[1†], John ZH. Zhang[1,2,3], Chih-Hao Chin[1]* and Tong Zhu[1,2]*

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

Email: zhjin@chem.ecnu.edu.cn, tzhu@lps.ecnu.edu.cn

†These authors contribute equally.

Abstract: The reactive molecular dynamics is widely used in the field of computational chemistry to study the reaction mechanisms in molecular systems. However, complex trajectories that are difficult to analyze have become a major obstacle to its application in large-scale systems. In this work, a new approach named ReacNetGen is developed to obtain reaction networks based on reactive MD simulations. Molecular species can be automatically generated from the 3D coordinates of atoms in the trajectory. The hidden Markov model is used to filter the noises in the trajectory, which makes the analysis process easier and more accurate. Compared with manual analysis, the advantage of this method in terms of efficiency is very obvious for large-scale simulation trajectories. It has been successfully used in the analysis of the simulated oxidation of 4-component RP-3 and methane.

1. Introduction

In the past decades, reactive molecular dynamic (MD) simulation has been widely used to study complex molecular systems with chemical reactions, such as combustion, explosion, and heterogeneous catalysis[1]. Compared with experimental methods, reactive simulation can give detailed reaction mechanisms on the atomic level. One of the most frequently used reactive MD method is the combination of classical MD engine and the Reactive Force Field (ReaxFF) proposed by van Duin et al[2]. The ReaxFF based on general bond-order potentials that describe the changing process of atoms and their connections in the molecular system continuously, laying the foundation for the simulation of the reaction in the system. Empirical parameters used in ReaxFF are calibrated from quantum chemistry calculations and experimental data for large numbers of reactions. Therefore, its accuracy is close to DFT, but can handle systems with a larger number of atoms and longer simulation times. By using reactive MD with ReaxFF, Li and co-workers have studied the pyrolysis of Liulin coal model with 28351 atoms[3] and oxidation of RP-1 jet fuel for 10ns[4]. Reaction events can also be simulated with *ab initio* molecular dynamics (AIMD). In their previous work, Wang et al. have proposed an *ab initio* nanoreactor[5] to explore new pathways for glycine synthesis from primitive compounds proposed to exist on the early Earth at the HF/3-21G level. The nanoreactor adopts graphics processing units (GPUs) to accelerate the electronic structure calculation, thus has higher efficiency than conventional QM calculations based on CPUs.

Large molecular models and long simulation time can produce complicit MD trajectories which contain a great number of reaction events and molecular species. This motivates the development of computational algorithms that can analyze these trajectories automatically. Liu et al. have proposed the VARxMD software which can analyze and visualize reactions from the coordinates and bond orders of atoms in the ReaxFF trajectory[6]. Martinez also developed a method called *nebterpolation* to discovery and refine reaction pathways from the AIMD trajectory[7]. In this work, a new tool called ReacNetGenerator (Reaction Network Generator) was developed for automatically extract molecular

species from ReaxFF and AIMD trajectories and generate reaction networks. The methods, algorithms, and applications of this method will be presented in the following sections.

2. Methods

The ReacNetGen consists of several modules and algorithms to process the information from the given trajectory. The flowchart of ReacNetGen is shown in Fig. 1. The key input of ReacNetGen is the MD trajectory from either ReaxFF or AIMD simulation, 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 bonding orders. Then the molecular fragments 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. A lot of "noise" molecules which are unstable in energy or structure will be detected from the first step. In their previous work, Wang et al. filtered these noises by using a two-state hidden Markov model (HMM)[7]. We also adopted this algorithm in the ReacNetGen approach. To facilitate the analysis, all detected molecules are indexed by canonical SMILES to guarantee its uniqueness. Isomers are identified among molecules according to SMILES, and the path and quantity of reactions in the whole trajectory are calculated. With the reaction matrix generated, the force-directed algorithm was used to draws a reaction network.

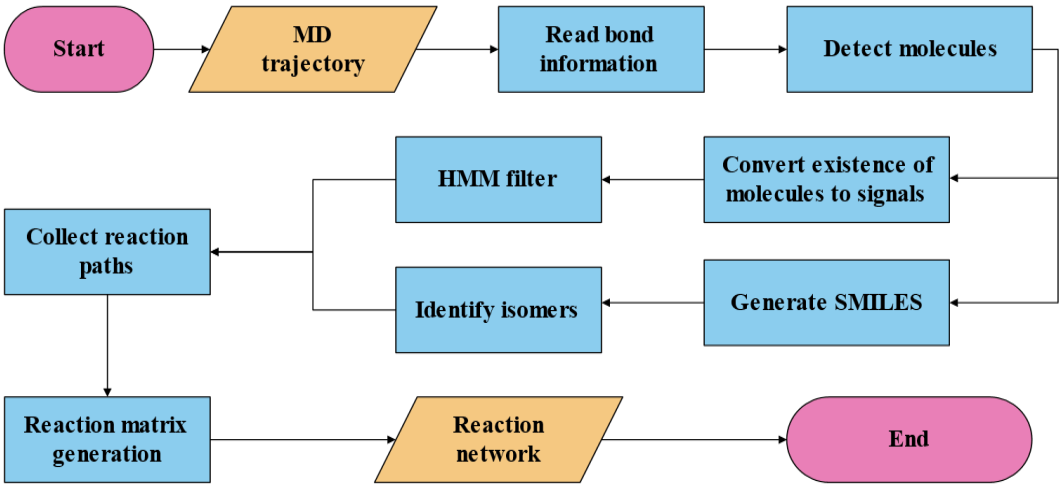

Figure 1. The flowchart of the ReacNetGen.

Details of these algorithms are discussed below.

2.1 Acquisition of molecular information.

There are two types of input files that can be read by ReacNetGen, the first and the necessary one is the trajectory from reactive MD, the second one can be the bond information normally given by simulation using ReaxFF. In fact, atomic coordinates can be converted to the bond information with the Open Babel software[8]. As a result, ReacNetGen can both process ReaxFF trajectories, AIMD trajectories, and other kinds of reactive trajectories.

With the bond information, molecules can be detected from atoms by Depth-first search[9] at every timestep. By using this way, all molecules in a given trajectory can be acquired. Molecules consisting of same atoms and bonds can be considered as the same molecule.

2.2 Noise filtering by hidden Markov model.

As mentioned above, a lot of molecules detected from the last step are useless for analyzing, such as a concussion between [H]O[H]O([H])[H] and water. If we keep these molecules during the analysis, a large amount of time and memory will be wasted. In this work, these useless molecules are considered as noise and will be filtered.

In order to filter noise, a two-state hidden Markov model (HMM)[7, 10] was adopted, which can be described as a transition matrix A, an emission matrix B, and an initial state vector π.

Here, the existence of molecules can be converted into 0-1 signals. Timesteps of molecules are converted to a visible output sequence $\mathrm{O}^m = (\mathrm{o}_t^{\mathrm{m}})$ given by

$$o_t^m = \begin{cases} 1, \ if \ m \ exists, \\ 0, \ \text{otherwise.} \end{cases} \quad (1)$$

Similarly, a hidden state sequence $\mathrm{I}^m = (i_t^{\mathrm{m}})$ is given by

$$i_t^m = \begin{cases} 1, \ if \ m \ exists, \\ 0, \ \text{otherwise.} \end{cases} \quad (2)$$

The set of possible states Q and the set of possible outputs V are given by

$$\mathrm{Q} = \{\mathrm{q}_1, \mathrm{q}_2\}, \mathrm{V} = \{\mathrm{v}_1, \mathrm{v}_2\} \quad (3)$$

where $\mathrm{q}_1 = \mathrm{v}_1 = 1$ and $\mathrm{q}_2 = \mathrm{v}_2 = 0$. The mathematical description of A, B and π is given by

$$A = [a_{ij}]_{N \times N} \quad (4)$$

$$B = [b_j(k)]_{N \times M} \quad (5)$$

$$\pi = (\pi_i). \quad (6)$$

Note that $a_{ij} = P(i_{t+1} = q_j | i_t = q_i), \mathrm{i} = 1,2, \dots, \mathrm{N}, \mathrm{j} = 1,2, \dots, \mathrm{N}$ is the probability of transition from state q_i at t to q_j at $\mathrm{t}+1$. $b_j(k) = P(o_t = v_k | i_t = q_j), \mathrm{k} = 1,2, \dots, \mathrm{M}, \mathrm{j} = 1,2, \dots, \mathrm{N}$ is the probability for state q_j of observing v_k. $\pi_i = P(i_1 = q_i), \mathrm{i} = 1,2, \dots, \mathrm{N}$ is the start probability for state q_i.

In order to predict the state sequence according to the output sequence, Viterbi algorithm[11] is used to acquire the path with the most likely hidden state sequence V called Viterbi path given by

$$P(I, O) = P(i_1 = q_i) \prod_{i=1}^{t} P(i_t = q_j | i_{t-1} = q_i) P(o_t = v_k | i_t = q_j) \quad (7)$$

$$V = \max_I P(I, O). \quad (8)$$

Note that $P(i_t = q_j | i_{t-1} = q_i)$, $P(o_t = v_k | i_t = q_j)$, and $P(i_1 = q_i)$ are respectively values of A, B and π in HMM, so Viterbi path is based on HMM parameters. Choosing smaller values in the off-diagonal of A will force the state sequence I to have fewer transitions. Choosing larger values in the off-diagonal of B will allow the Markov process to deviate more often from the output sequence O. Both two changes can increase the strength of the noise filter[7].

In summary, the HMM signals of all molecules are converted from the origin signals by the Viterbi algorithm, so that large amounts of molecules are filtered. In this work, $\pi = (0.5, 0.5)$, $A = \begin{bmatrix} 0.999 & 0.001 \\ 0.001 & 0.999 \end{bmatrix}$, and $B = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$ are taken from Ref. 7.

2.3 Generate of the reaction network.

To produce a reaction network, every molecule (species) should be treated as a node in the network. Therefore, all detected species are indexed by canonical SMILES[12] to guarantee its uniqueness. Isomers are also identified according to SMILES codes. The VF2 algorithm[13] can be also used to identify isomers, which is an option in ReacNetGen. After filtering out noise, the reaction path of atoms and the number of intermolecular reactions can be calculated.

A reaction network cannot accommodate too many species, so only the first species which have the most reactions are taken. A reaction matrix can be generated as

$$R = [a_{ij}],\ i = 1,2,\dots,100;\ j = 1,2,\dots,100 \qquad (9)$$

where a_{ij} is the number of reactions from species s_i to s_j.

Finally, with the reaction matrix, a reaction network can be drawn. Here the *NetworkX*[14] package is used to make a graph which indicates reactions between species. Fruchterman-Reingold force-directed algorithm[15] is used to make the layout of nodes relate to reaction quantity and different colors and widths of lines are drawn depending on reaction quantity. The distance of two species in the network, the color and thickness of the line between them are determined by the number of their reactions, making the reaction network more intuitive.

3. Results and discussions

Signals of three selected species before and after noise filtering by HMM are shown as Fig. 2. As can be seen, the first two molecules appear very frequently during certain periods of the trajectory, but only occasionally appear in other time periods. The HMM signal accurately reflect their existence. The third species is more like a water and a hydrogen peroxide molecule which are in close contact due to the collision. Its signal is very sparse, and its lifetime is very short, so the HMM successfully filtered it out.

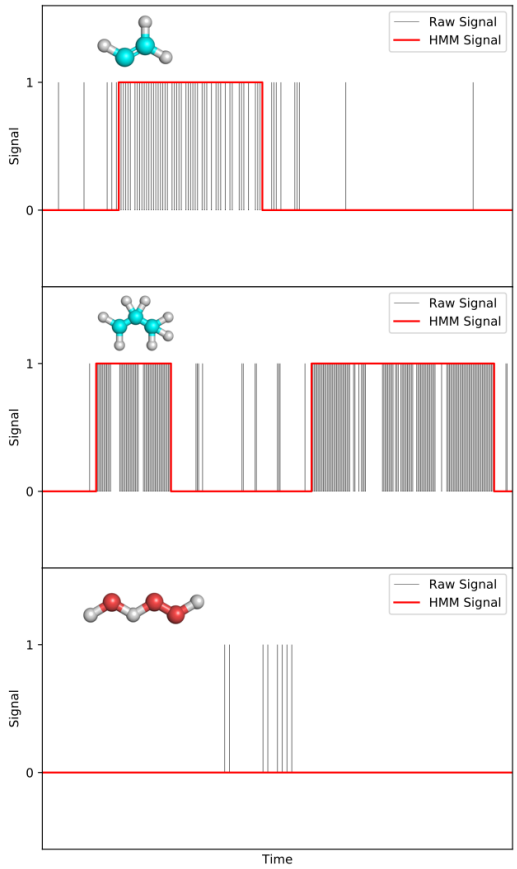

Figure 2. The performance of HMM for three selected species.

Fig. 3 depicts reaction networks with HMM and without HMM for the reactive MD trajectories of the oxidation of 4-component RP-3. In addition to the trajectories, the bond information given by ReaxFF is also used as input to the ReacNetGen. For each situation, 20 species with the most reactions are taken

to draw the network. As can be seen, the network without HMM is relatively sample, mainly composed of reactions between two species. For example, the reaction path between species 5 and 6, 9 and 10 only reflects the forming and breaking of a double bond. In contrast, the HMM filtered most of the unstable species, which makes the network more reasonable, and contains more useful information.

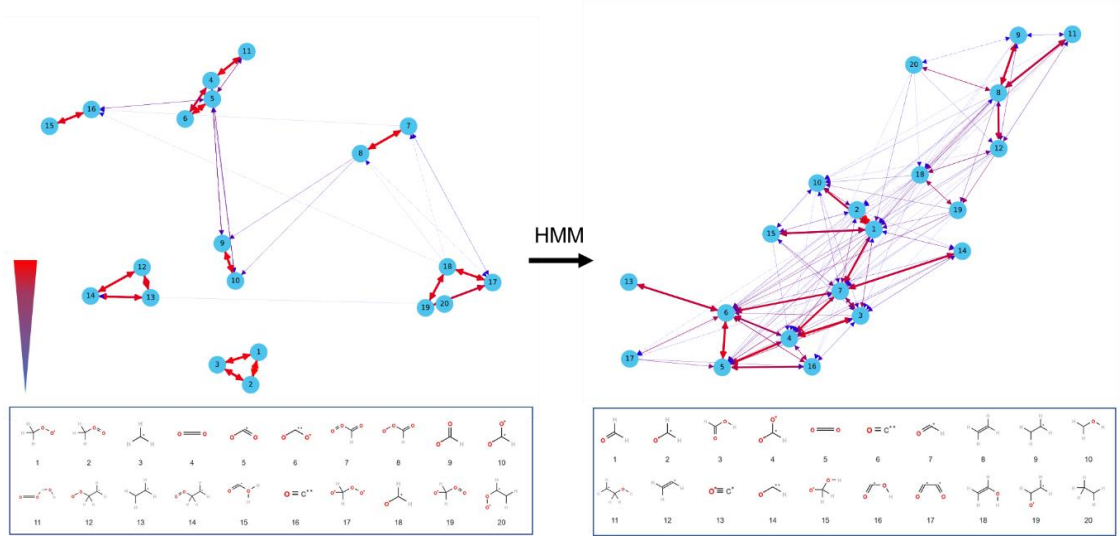

Figure 3. The reaction network for a 2.5ns reactive MD simulation of 4-component RP-3 oxidation with ReaxFF. The left panel is the network without HMM filtering and the right one is the network with HMM.

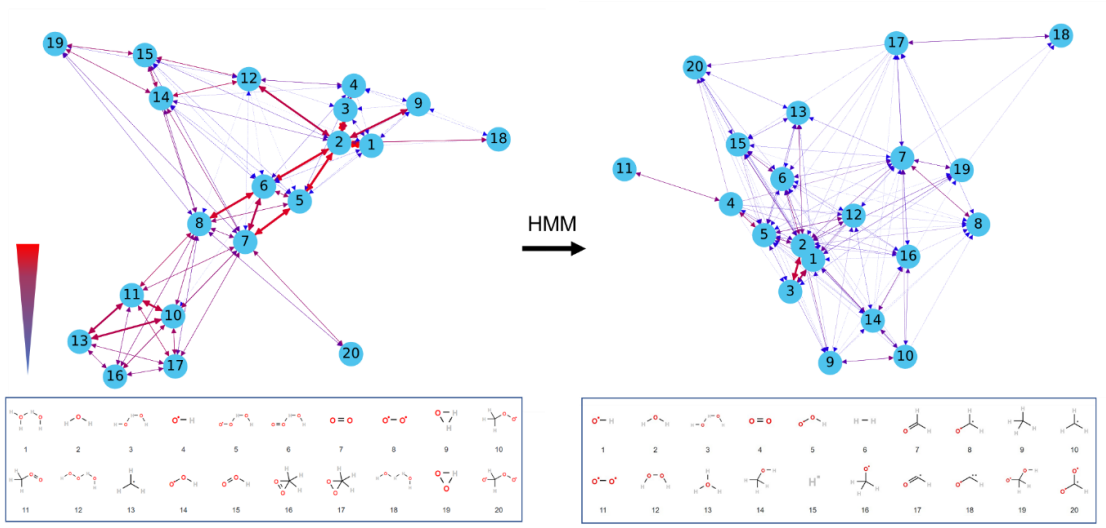

Figure 4. The reaction network for a 2.5ns reactive MD simulation of methane oxidation with ReaxFF. The left panel is the network without HMM filtering and the right one is the network with HMM.

Similar to Fig. 3, Fig. 4 depicts reaction networks with HMM and without HMM for the reactive MD trajectories of the oxidation of methane. In the network without HMM, there still a lot of insignificant oscillations. For example, the reaction between species 2, 6, 7 and 8 are just the oscillations between O_2 and water molecule. Compare with the network without HMM, the one with HMM is more reasonable. From the above discussion and comparison, a conclusion can be made that noise filtering by HMM is effective and necessary.

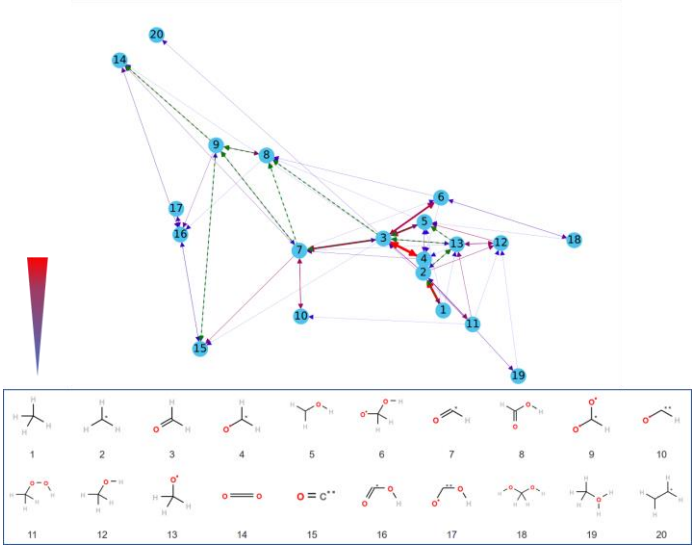

Figure 5. Reaction network starts from methane for the 2.5ns reactive MD simulation of methane oxidation with ReaxFF. The HMM was used to filter noises in the trajectory. The dashed green lines indicate the reaction paths found in Ref. 16.

ReacNetGen can give reaction network starts form any given species. Fig. 5 shows the reaction network starts from methane in a 2.5ns reactive MD simulation of methane oxidation. In order to compare the results, we used the same simulation conditions as the previous work of He et al[16]. In the simulation, a 3-dimension periodic box containing 50 CH_4 and 100 O_2 molecules with the density of 0.218g/cm^{-3} was used, and the temperature was set to 3000K. As shown in Fig. 5, all of the important species detected manually in Ref. 16, such as •CH_3, CH_3O• and HCHO etc. are included. And the main reaction route calibrated by QM calculation are also covered by the reaction network. Fig. 6 shows the reactive network for the same trajectory that used in Fig. 4, but the OpenBabel was used for the analysis of atomic connectivity instead of bond information given by ReaxFF.

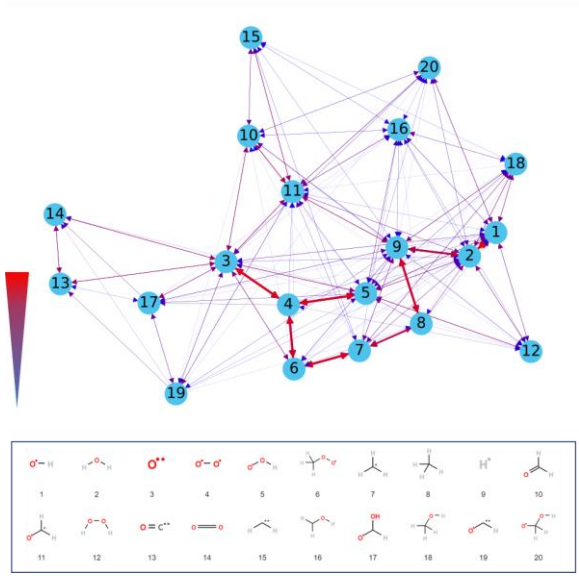

Figure 6. The reaction network for a 2.5ns reactive MD simulation of methane oxidation with ReaxFF. The OpenBabel was used to detect the atomic connectivity. The HMM was used to filter noises in the trajectory

The reaction network in Fig. 6 agrees very well with that in Fig. 4, which means that trajectories produced by reactive MD simulations which only output atomic coordinates, like AIMD, can also be used in ReacNetGen.

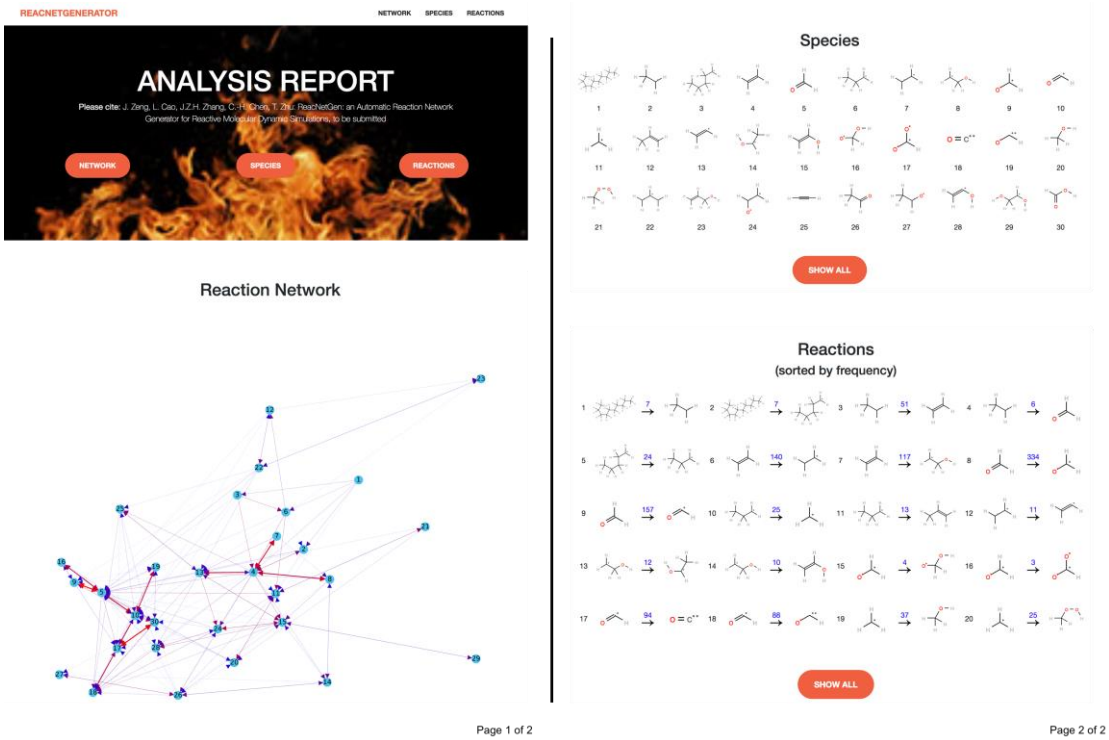

Figure 7. The web page contains all of the results given by ReacNetGen for the 2.5ns reactive MD simulation of 4-component RP-3 oxidation with ReaxFF.

For the convenience of the user, we put all the results generated by ReacNetGen (Network, Species, and Reactions) in an interactive web page. By default, 30 species with the most reactions are taken to draw the network. However, by clicking on a given species, one can check the special network which starts from it.

4. Conclusion

A new approach named ReacNetGen is developed to obtain reaction networks based on reactive MD simulations. Molecular species can be automatically generated from the 3D coordinates of atoms in the trajectory. The hidden Markov model is used to filter the noises in the trajectory, which makes the analysis process easier and more accurate. Compared with manual analysis, the advantage of this method in terms of efficiency is very obvious for large-scale simulation trajectories. It has been used in the analysis of the simulated oxidation of 4-component RP-3 and methane, and the results show good agreement with previous studies. Further study of this method will focus on two aspects, the first one includes parallel algorithms to further improve its efficiency and the second one is to introduce energy criteria to make the results of the analysis more accurate.

References

1. Aktulga, H. M.; Fogarty, J. C.; Pandit, S. A.; Grama, A. Y., Parallel reactive molecular dynamics:

Numerical methods and algorithmic techniques. *Parallel Computing* **2012,** *38* (4-5), 245-259.
2. Chenoweth, K.; Cheung, S.; van Duin, A. C.; Goddard, W. A.; Kober, E. M., Simulations on the thermal decomposition of a poly (dimethylsiloxane) polymer using the ReaxFF reactive force field. *Journal Of The American Chemical Society* **2005,** *127* (19), 7192-7202.
3. Zheng, M.; Li, X.; Liu, J.; Wang, Z.; Gong, X.; Guo, L.; Song, W., Pyrolysis of Liulin coal simulated by GPU-based ReaxFF MD with cheminformatics analysis. *Energy & Fuels* **2013,** *28* (1), 522-534.
4. Han, S.; Li, X.; Nie, F.; Zheng, M.; Liu, X.; Guo, L., Revealing the Initial Chemistry of Soot Nanoparticle Formation by ReaxFF Molecular Dynamics Simulations. *Energy & Fuels* **2017,** *31* (8), 8434-8444.
5. Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J., Discovering chemistry with an ab initio nanoreactor. *Nature Chemistry* **2014,** *6*, 1044.
6. Liu, J.; Li, X. X.; Guo, L.; Zheng, M.; Han, J. Y.; Yuan, X. L.; Nie, F. G.; Liu, X. L., Reaction analysis and visualization of ReaxFF molecular dynamics simulations. *J Mol Graph Model* **2014,** *53*, 13-22.
7. Wang, L.-P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J., Automated discovery and refinement of reactive molecular dynamics pathways. *Journal of chemical theory and computation* **2016,** *12* (2), 638-649.
8. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011,** *3* (1), 33.
9. Tarjan, R., Depth-first search and linear graph algorithms. *SIAM journal on computing* **1972,** *1* (2), 146-160.
10. Rabiner, L. R., A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **1989,** *77* (2), 257-286.
11. Forney, G. D., The viterbi algorithm. *Proceedings of the IEEE* **1973,** *61* (3), 268-278.
12. Landrum, G., RDKit: Open-Source Cheminformatics Software. **2016**.
13. Cordella, L. P.; Foggia, P.; Sansone, C.; Vento, M., A (sub) graph isomorphism algorithm for matching large graphs. *IEEE transactions on pattern analysis and machine intelligence* **2004,** *26* (10), 1367-1372.
14. Hagberg, A.; Swart, P.; S Chult, D. *Exploring network structure, dynamics, and function using NetworkX*; Los Alamos National Lab.(LANL), Los Alamos, NM (United States): 2008.
15. Fruchterman, T. M.; Reingold, E. M., Graph drawing by force-directed placement. *Software: Practice and experience* **1991,** *21* (11), 1129-1164.
16. He, Z.; Li, X.-B.; Liu, L.-M.; Zhu, W., The intrinsic mechanism of methane oxidation under explosion condition: A combined ReaxFF and DFT study. *Fuel* **2014,** *124*, 85-90.