

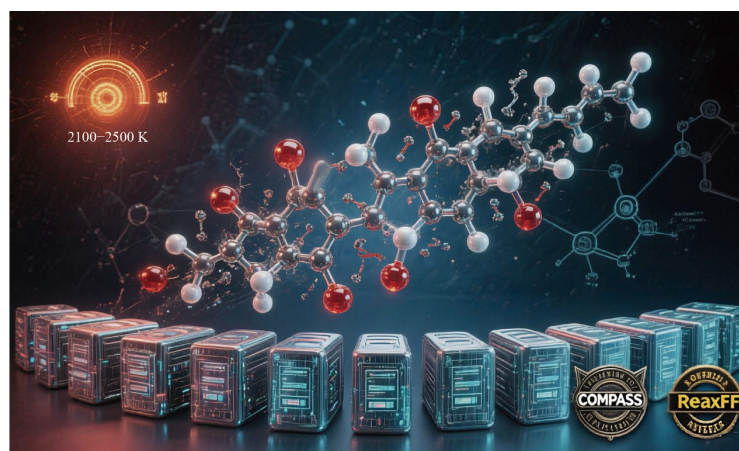
Molecular dataset based on paraffin oxidative pyrolysis

YANG Lin¹, YANG Yanfang¹, LU Kuan^{2,*}

¹School of Statistics, Shanxi University of Finance and Economics, Taiyuan 030006, China;

²State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan 030001, China

Abstract: As a key component of shale oil, petroleum fractions, and chemical products, the oxidative pyrolysis behavior of paraffin directly influences energy conversion efficiency and the direction of process optimization. A deep understanding of its oxidative pyrolysis mechanism is crucial for addressing wax deposition in oil and gas extraction, enhancing product selectivity in cracking processes, and advancing novel clean fuel technologies. Traditional experimental methods face challenges in capturing transient free-radical reaction pathways at



high temperatures, whereas molecular dynamics simulations offer a powerful approach to bridge the research gap in elucidating atomic-scale dynamic mechanisms. This database is constructed based on high-precision molecular dynamics simulations, comprising oxidative pyrolysis trajectory data for three paraffin models featuring different straight-chain hydrocarbon distributions within the temperature range of 2100–2500 K. The COMPASS force field was employed to optimize the initial structures, and the ReaxFF reactive force field was used to simulate the oxidative pyrolysis process. The database includes atomic trajectories, species evolution information, and reaction network analysis results for both heating and isothermal cracking processes, with a total data volume of approximately 141 GB (including 150000 atomic configuration frames). The data is stored in a hierarchical directory structure, supporting multi-scale oxidative pyrolysis mechanism studies and providing atomic-scale dynamic evidence for revealing carbon chain length effects and temperature sensitivity.

Key words: paraffin oxidative pyrolysis; molecular dynamics; ReaxFF force field; reaction network

In the indirect coal liquefaction process, the residual wax produced by the Fischer-Tropsch synthesis reaction is a key byproduct that restricts system efficiency and economic viability. These residual waxes are classified into three primary sources: heavy oil filter cakes, generated during periodic filtration of slurry-bed reactors, which contain deactivated iron-based catalysts and high-carbon-number heavy wax oils; heavy wax filter cakes, formed in the separation unit from liquid products, comprising residual catalyst particles and waxy components; and catalyst replacement filter cakes, produced in the filtration unit after

additive incorporation, primarily composed of heavy wax, additives, and trace catalysts. Efficient conversion of these residual waxes is one of the core challenges in optimizing the Fischer-Tropsch synthesis process. Notably, the oxidative pyrolysis behavior of the paraffin components (C_{14} – C_{45}) directly determines their energy conversion efficiency and resource recovery potential. Therefore, elucidating the oxidative pyrolysis mechanism of paraffins at the atomic scale holds dual significance for addressing wax deposition, regulating product selectivity, and developing clean fuel technologies—it pertains both to

Received: 2025-11-13; Revised: 2025-12-07

Foundation items: Supported by Natural Science Foundation of Shanxi Province (202203021221219), Research on the Construction of Scientific and Technological Innovation Think Tank of Shanxi Association for Science and Technology (KXKT202542), Planning Project under Commerce Statistical Society of China (2025STY122).

* Corresponding author. E-mail: lukuan@sxicc.ac.cn.

DOI: 10.1016/S1872-5813(26)60646-9 CSTR: 32157.14.2097-213X.2025.JFCT.0205 Associated Data DOI: 10.57760/sciencedb.31639

本文的英文电子版由 Elsevier 出版社在 ScienceDirect 上出版 (<http://www.sciencedirect.com/science/journal/18725813>)

process intensification and the high-value utilization of byproducts.

However, existing research faces two major bottlenecks. First, traditional experimental methods struggle to capture transient free-radical reaction pathways (e.g., nanosecond-scale chain reactions) at high temperatures (>2000 K), as the lifetimes of key intermediates (such as peroxy radicals) are extremely short (<1 ns), exceeding the spatiotemporal resolution limits of conventional characterization techniques. Second, current molecular dynamics (MD) simulations are mostly limited to single-component models or narrow temperature ranges, failing to systematically resolve the correlation mechanisms between complex carbon chain distribution (C_{14} – C_{45}) in actual residual wax and oxidative pyrolysis behavior across a wide temperature range. Reactive force field (ReaxFF), which was constructed based on bond order concepts, can accurately describe chemical bond breaking/formation processes, providing an atomic-scale dynamic perspective for paraffin oxidative pyrolysis. Meanwhile, high-precision MD trajectory data can drive machine learning models to optimize force field parameters and predict product distributions, offering a new paradigm for industrial intelligent regulation.

To address these challenges, the paraffin oxidative pyrolysis dataset (see Table 1) constructed in this study achieves three pioneering innovations: (1) Multi-scale model adaptation: Based on the actual compositional characteristics of Fischer-Tropsch synthesis residual wax, three carbon chain distribution models (Simu1–Simu3) were constructed to directly correlate with compositional variations in industrial residual wax. (2) Wide temperature range dynamic trajectory coverage: The full oxidative pyrolysis pathways within the 2100–2500 K temperature range were recorded in detail (0.25 fs time step), filling the gap in high-temperature transient reaction mechanisms. (3) Intelligent analysis compatibility: Data formats are pre-adapted for ReacNetGenerator reaction network analysis and OVITO visualization tools, supporting machine learning-driven reaction prediction and coking mechanism exploration.

As a core demonstration of hydrocarbon molecular dataset, this dataset provides a theoretical cornerstone for the efficient conversion of Fischer-Tropsch synthesis residual wax by sharing high-fidelity simulation data, driving a paradigm shift in the coal chemical industry from process optimization to mechanism-driven advancement.

Table 1 Data profile of molecular dataset based on paraffin oxidative pyrolysis

Data title (V0.1)	Molecular dataset based on paraffin oxidative pyrolysis
Author's name	YANG Lin, YANG Yanfang, LU Kuan
Corresponding author's name	LU Kuan
Data content description	This dataset system contains molecular dynamics simulation data of oxidative pyrolysis for three paraffin models (Simu1–Simu3) across a temperature range of 2100–2500 K. The trajectory data includes heating processes (sampled every 4000 atomic configuration frames) and isothermal oxidative pyrolysis processes (sampled every 6000 atomic configuration frames), cumulatively generating 150000 high-resolution atomic trajectory frames (calculated as 3 models $\times$ 5 temperature points $\times$ 10000 frames per simulation). Post-processed trajectory data comprises 10 file types per simulation, including reaction networks, species statistics, and reaction pathway analyses. These yield 150 standardized analysis files in multiple formats: interactive .html reports, .csv species concentration tables, and .json reaction topology files. The combined raw trajectory and post-processed data amount to approximately 141 GB, stored in a hierarchical directory structure (e.g., Simu1/2200K/thermostat/frames-thermostat/). This comprehensive dataset thoroughly documents the evolution of chemical bonds and radical reaction pathways in paraffins under high-temperature conditions.
Data generation time	2024–2025
Data volume	141 GB (including 150,000 atomic configuration frames)
Data format	.zip, .data
Data source location	https://cstr.cn/31253.11.sciencedb.31639 https://doi.org/10.57760/sciencedb.31639
Citation format of data	YANG Lin, YANG Yanfang, LU Kuan. Molecular dataset based on paraffin oxidation pyrolysis[J]. <i>J Fuel Chem Technol</i> , 2026, 54, 20250365. YANG Lin, YANG Yanfang, LU Kuan. Molecular dataset based on paraffin oxidation pyrolysis [DB/OL]. Science Data Bank, 2025. https://doiorg/10.57760/sciencedb.31639 . DOI:10.57760/sciencedb.31639.

1 Data value

Through its multidimensional design, this dataset delivers systematic data support and analytical frameworks for investigating the mechanisms of paraffin oxidative pyrolysis. Its core value is first demonstrated by a multi-model comparison framework, in which users can quantitatively assess the influence of carbon chain length on reaction pathways based on three initial structures with distinct carbon chain distributions (Simu1–Simu3). This capability offers mechanistic insights into the differences in reactivity among residual waxes. Second, the dataset enables comprehensive reaction trajectory analysis, recording atomic motion with a temporal resolution of 0.1 ps. This high-resolution data allows for the tracking of dynamic evolution processes of key intermediates—such as alkyl and peroxy radicals—thereby supplying dynamic evidence for elucidating the microscopic mechanisms of free-radical chain reactions. Third, the dataset covers five temperature points (2100–2500 K), forming a complete kinetic temperature series. Users can extract species concentration data from the “thermostat” folder to construct Arrhenius curves, accurately determine the activation energies of elementary reactions such as C–C bond cleavage, and identify potential reaction pathway transition thresholds around 2300 K. Examples include the shift from dehydrogenation-dominated mechanisms at lower temperatures to direct bond cleavage mechanisms at higher temperatures. These findings provide a theoretical basis for optimizing oxidative pyrolysis temperatures to mitigate excessive oxidation. In terms of analytical compatibility, the dataset’s output formats are directly supported by reaction network analysis tools such as ReacNetGenerator, and are also compatible with bond-order visualization modules in OVITO software. This enables intuitive identification of the spatiotemporal distribution characteristics of coking precursors—such as polycyclic aromatic hydrocarbons—facilitating cross-scale analysis from the atomic level to mesoscopic aggregation behavior. Furthermore, the dataset can be utilized to construct macroscopic kinetic models for residual wax oxidative pyrolysis or to train machine learning models for rapid prediction of product yields under various operating conditions, thereby directly supporting process optimization and reactor design. In summary, by integrating multiple models, high-resolution trajectories, a broad temperature range, and standardized analytical interfaces, this dataset offers a comprehensive and reliable data platform for studying reaction kinetics in paraffin oxidative pyrolysis and for developing coking inhibition strategies.

2 Data generation method or process

The data in this dataset were obtained via reactive molecular dynamics simulations. The construction of the

dataset involved an integrated workflow comprising three main stages: initial data preparation, data generation, and data processing. A schematic overview of the overall procedure is presented in Figure 1.

Based on the experimental data (see Supplementary Table S1), we constructed three three-dimensional periodic paraffin models (Simu1, Simu2, and Simu3), each exhibiting distinct hydrocarbon distributions. The initial simulation box dimensions were $126.88 \text{ \AA} \times 126.88 \text{ \AA} \times 126.88 \text{ \AA}$. Periodic boundary conditions were applied in all three spatial directions throughout the simulation to accurately represent bulk behavior. Only straight-chain alkane structures were considered in the paraffin models. The distribution of carbon chain lengths, ranging from C_{14} to C_{45} , was consistent with the experimental observations (see Table 2). Following the addition of 2000 O_2 molecules to each model, structural optimization was first performed using the COMPASS force field^[1–2]. The systems were then subjected to annealing by heating from 300 to 1000 K under an NVT ensemble. Finally, the optimized structures were equilibrated in an NPT ensemble at 1000 and 0.8 GPa. After relaxation, the model density was determined to be 0.92 g/cm^3 , and the final box dimensions were approximately $60.369 \text{ \AA} \times 60.39 \text{ \AA} \times 60.39 \text{ \AA}$.

All reactive molecular dynamics simulations were carried out using the LAMMPS^[3] software. The CHO reactive force field^[4] was employed with a time step of 0.25 fs. Temperature control was achieved via an NVT ensemble combined with a Nosé-Hoover thermostat, using a damping constant of 100 fs. The Verlet algorithm was used for MD integration. Initially, all models were structurally relaxed for 100 ps at 300 K. The relaxed configurations were subjected to a heating protocol, starting from 300 K and increasing to final temperatures of 2100, 2200, 2300, 2400 and 2500 K. The heating phase was conducted over 400 ps. The simulation temperatures were set significantly above the experimental values to accelerate reaction kinetics, thereby enabling the observation of a larger number of chemical events within the limited simulation timeframe.

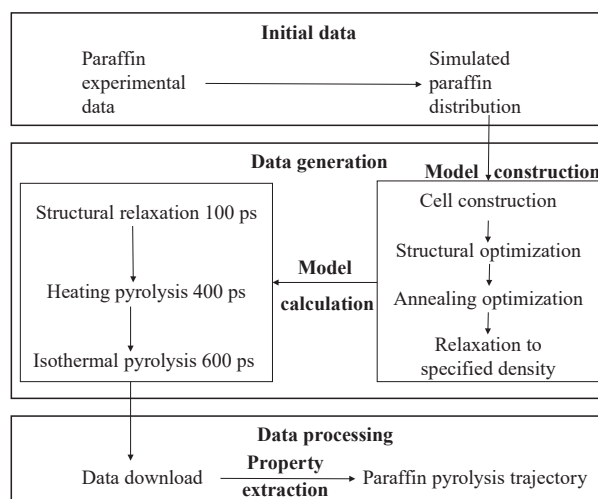


Fig. 1 Data collection process flowchart

Table 2 Hydrocarbon distribution of three paraffin models

Number of carbon	Simulation 1	Simulation 2	Simulation 3
14	0	7	0
15	0	9	0
16	0	10	0
17	0	11	0
18	0	11	0
19	4	11	0
20	5	10	4
21	5	10	5
22	6	12	6
23	7	0	6
24	7	10	7
25	8	9	8
26	9	8	8
27	0	8	9
28	0	7	0
29	9	7	0
30	8	6	8
31	8	6	8
32	8	5	7
33	7	5	7
34	7	4	7
35	6	4	6
36	6	4	6
37	5	3	5
38	5	3	5
39	4	3	4
40	4	2	4
41	3	2	4
42	3	0	3
43	2	0	3
44	2	0	2
45	0	0	2
Total atoms	12846	14066	12755
$\rho/(\text{g}\cdot\text{cm}^{-3})$	0.92		

Subsequently, the high-temperature structures obtained at each target temperature were maintained under isothermal conditions for 600 ps to simulate oxidative pyrolysis. Molecular products were identified using a bond-order cutoff of 0.3. Species information and trajectory data were recorded every 400 steps (corresponding to 0.1 ps). Reaction analysis was performed using ReacNetGenerator^[5], and molecular structures were

visualized with the OVITO^[6] software.

3 Data sample description

The data in this dataset are organized according to a three-level hierarchical structure, corresponding to the three paraffin component distributions (Simu1, Simu2, Simu3). For each component model, simulation results are provided for five temperature points: 2100, 2200, 2300, 2400, 2500 K. The data for each temperature point are further categorized into two phases: the heating phase (heat) and the constant-temperature oxidative pyrolysis phase (thermostat). Within each phase folder, a frame data directory (frames-heat or frames-thermostat) contains 4001 and 6001 structure files (named frame*.data), respectively. These files record atomic configurations and evolution of products at different time steps. In addition, the remaining 10 files with the ".dump.*" suffix store post-processing information derived from the frame data, including reaction network topologies and species statistics. Detailed descriptions of the data storage structure and analysis reports are provided in Table 3 and Figure 2–5. For definitions of numerical values and symbols, readers are referred to the methodology established in reference [5].

4 Data quality

To ensure the reliability and reproducibility of reactive molecular dynamics simulation data for paraffin oxidative pyrolysis, systematic quality control measures were implemented throughout the entire simulation workflow. For the force field, the high-accuracy CHO-ReaxFF reactive potential, parameterized using first-principles calculations, was employed to accurately describe the reaction pathways of carbon, hydrogen, and oxygen species at elevated temperatures. Prior to reaction simulations, all systems underwent structural relaxation for 100 ps to eliminate initial configuration deviations and ensure that both energy and structure reached equilibrium. Temperature control was maintained using a Nosé-Hoover thermostat with a damping constant of 100 fs. Fluctuations were verified to remain below $\pm 1\%$ at 2100 K (Figure 6), satisfying the stability requirements for reactive dynamics simulations. To guarantee the integrity of trajectory data, full atomic coordinates and bond-order information were recorded every 400 steps (0.1 ps).

Table 3 Dataset storage content based on paraffin oxidative pyrolysis

Data type	Data content	Data format	Software
Structural information	evolution of reactive species over time	.dump.data	LAMMPS
Statistics for all species	final species distribution after the reaction	.dump.csv	
Analysis report	reaction Network	.dump.html/.dump.json/.dump.svg	ReacNetGenerator
	reaction species	.dump.moname/.dump.species	
	reaction frequency	.dump.table/.dump.reaction	
	reaction pathway	.dump.reactionabcd/.dump.route	

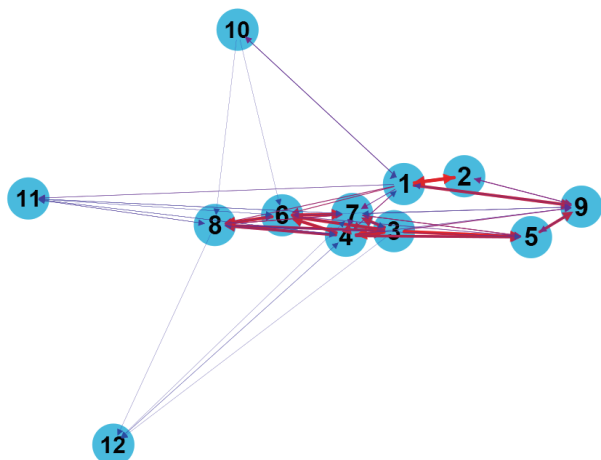


Fig. 2 Schematic diagram of the reaction network in the analysis report of Simu1_2200/thermostat

A uniform bond-order cutoff of 0.3 was applied to ensure consistent species identification across all simulations. Furthermore, the energy conservation error under a 0.25 fs integration timestep was maintained below 0.5% (Figure 6), meeting the numerical precision criteria for short-range dynamics simulations in LAMMPS. Collectively, these measures ensure that the simulation results fulfill the

requirements for scientific research in terms of physical plausibility and data reliability.

5 Data usage methods and suggestion

To maximize the utility of this dataset in investigating the mechanisms of paraffin oxidative pyrolysis, users are advised to adopt the following integrated analytical workflow. This workflow comprises a full-chain methodology spanning atomic-scale trajectory visualization, reaction network reconstruction, thermodynamic analysis, and machine learning model development, with the aim of systematically elucidating the oxidative cracking behavior and regulatory principles of paraffin at high temperatures. Atomic motion trajectories can be parsed and visualized using OVITO software (example file path: Simulation/Simu1_2100K/thermostat/frames-thermostat/frames*.data). By utilizing its built-in bond-order analysis function (recommended bond-order threshold: 0.3), users can dynamically track the bond-breaking and bond-forming sequences of key elements (C, H, and O) during the reaction process, thereby directly observing the evolution pathways and lifetimes of radical intermediates (e.g., alkyl radicals).

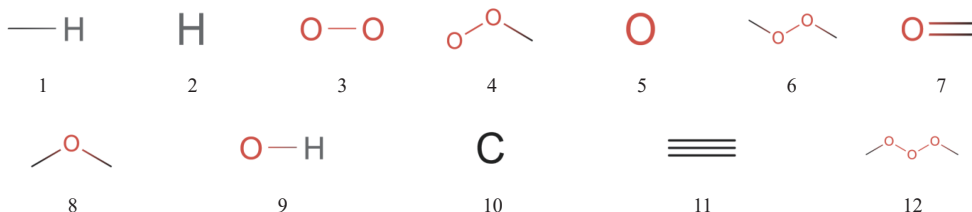


Fig. 3 Schematic diagram of reactive species in the analysis report of Simu1_2200/thermostat

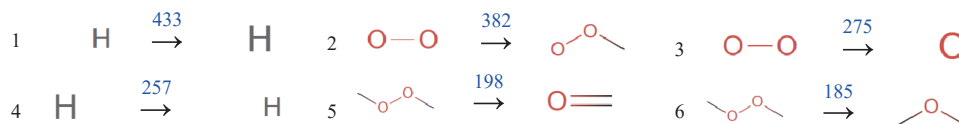


Fig. 4 Schematic diagram of reaction frequency in the analysis report of Simu1_2200/thermostat



Fig. 5 Schematic diagram of the reaction pathway in the analysis report of Simu1_2200/thermostat

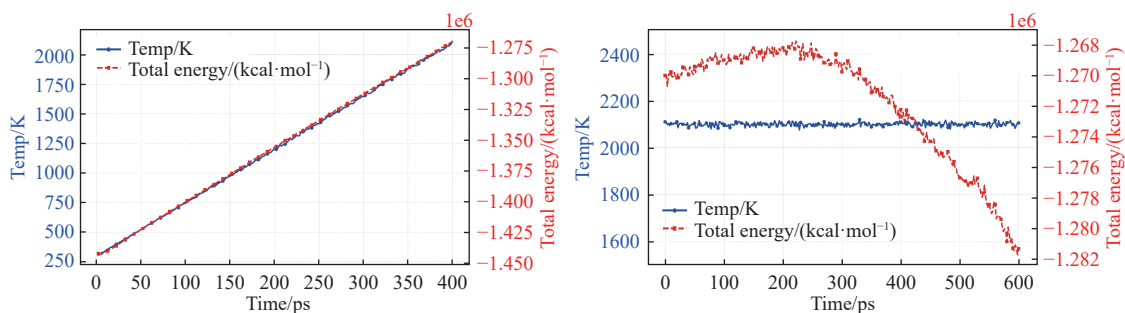


Fig. 6 Variation trend diagram of temperature and total energy over time during heating phase (400 ps) and isothermal phase (600 ps)

In the reaction network analysis stage, users can directly open the generated .html format report files in a web browser. These files, automatically generated by ReacNetGenerator, present reaction pathways, key species, and their occurrence frequencies in an interactive graphical format. By examining the topological changes in network structures under different temperatures or models—such as high-frequency reaction pathways and dominant reaction channels—users can identify the formation pathways of coking precursors and the competitive relationship between oxidation and oxidative pyrolysis reactions.

To quantify the effect of temperature on reaction pathways, it is recommended to extract and compare product distribution data files (e.g., processed_species-thermostat.csv) from folders corresponding to different temperatures (e.g., Simul/2100K and Simul/2500K). Analyzing the concentration variations of specific products (e.g., C_2H_4 , CO_2) with temperature enables the calculation of thermodynamic parameters such as apparent activation energies, thereby revealing mechanistic shifts with temperature (e.g., the transition from peroxy radical-dominated oxidation at lower temperatures to direct C–C bond cleavage at higher temperatures). For detailed analytical procedures, readers are referred to the literature “Atomic-scale insight into thermal decomposition behavior of polypropylene: A ReaxFF method”. All raw .data trajectory files in this dataset can serve as input for machine learning models. Users may convert these files into formats suitable for model training (e.g., feature vectors describing local atomic environments) and subsequently develop predictive models. Such models can be employed to rapidly predict the equilibrium product distributions of paraffin oxidative pyrolysis under varying initial compositions and temperature conditions, providing an efficient computational tool for materials design and process optimization.

6 Conclusion

This dataset constructs an oxidative pyrolysis trajectory dataset via molecular dynamics simulations, covering three paraffin components across a temperature range of 2100–2500 K, providing multiscale atomic dynamic information for elucidating hydrocarbon oxidative pyrolysis mechanisms. Its multi-model design supports quantitative

analysis of how carbon chain length influences reaction pathways; high-resolution trajectories comprehensively record the evolution of free radicals; wide temperature coverage facilitates the extraction of kinetic parameters and identification of mechanistic transition thresholds; and the standardized storage structure paired with compatible formats ensures the ready-to-use nature and reproducibility of the data. Not only does this dataset provide an atomic-scale basis for fundamental research on paraffin oxidative pyrolysis, but it also lays a data foundation for industrial process optimization and the exploration of machine learning-driven new paradigms.

Data availability statement

The data supporting the results in this study has been published in sciencedb dataset and can be obtained through DOI:10.57760/sciencedb.31639.

Declaration of Competing Interest

The authors declare no conflict of interest.

References

- [1] SUN H, JIN Z, YANG C, et al. COMPASS II: Extended coverage for polymer and drug-like molecule databases[J]. J Mol Model, 2016, 22(2): 47.
- [2] SUN H. COMPASS: An ab initio force-field optimized for condensed-phase applications overview with details on alkane and benzene compounds[J]. J Phys Chem B, 1998, 102(38): 7338–7364.
- [3] PLIMPTON S. Fast parallel algorithms for short-range molecular dynamics[J]. J Comput Phys, 1995, 117(1): 1–19.
- [4] ASHRAF C, VAN DUIN A C. Extension of the ReaxFF combustion force field toward syngas combustion and initial oxidation kinetics[J]. J Phys Chem A, 2017, 121(5): 1051–1068.
- [5] ZENG J, CAO L, CHIN C H, et al. ReacNetGenerator: An automatic reaction network generator for reactive molecular dynamics simulations[J]. Phys Chem Chem Phys, 2020, 22(2): 683–691.
- [6] STUKOWSKI A. Visualization and analysis of atomistic simulation data with OVITO—the open visualization tool[J]. Modell Simul Mater Sci Eng, 2010, 18(1): 015012.

基于石蜡氧化热解的分子数据集

杨 琳¹, 杨艳芳¹, 路 宽^{2,*}

(1. 山西财经大学 统计学院, 山西 太原 030006;

2. 中国科学院山西煤炭化学研究所 煤炭高效低碳利用全国重点实验室, 山西 太原 030001)

摘 要: 石蜡作为页岩油、石油馏分及化工产品的关键组分, 其氧化热解行为直接影响了能源转化效率与工艺优化方向。深入理解石蜡氧化热解机理对解决油气开采中的蜡沉积问题、提升裂解工艺产物选择性以及开发新型清洁燃料技

术具有关键意义。然而,传统实验手段难以捕捉高温下瞬态自由基反应路径,而分子动力学模拟可填补原子尺度动态机制的认知空白。本数据库基于高精度分子动力学模拟构建,包含三种不同直链烃分布的石蜡模型在 2100–2500 K 的氧化热解轨迹数据。采用 COMPASS 力场优化初始结构,利用 ReaxFF 反应力场模拟热解过程,数据集涵盖加热过程和恒温裂解过程的原子轨迹、物种演化信息及反应网络分析结果,总数据量约 141 GB(含 15 万原子构型帧)。数据采用层级目录存储,支持多尺度热解机理研究,为揭示碳链长度效应与温度敏感性提供原子尺度动态依据。

关键词: 石蜡氧化热解; 分子动力学; ReaxFF 力场; 反应网络

中图分类号: O64

文献标识码: A

作者简介:



杨琳(1989), 博士, 主要从事生存分析理论方法研究和深度学习、机器学习等统计方法的改进与应用, ylintj@126.com。



路宽(1991), 博士, 副研究员, 主要从事计算方法开发、计算化学模拟在化学、材料科学中的研究和应用工作, lukuan@sxicc.ac.cn。