PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2024, **26**, 5649

Generating a skeleton reaction network for reactions of large-scale ReaxFF MD pyrolysis simulations based on a machine learning predicted reaction class†

Shanwen Yang, [ab] Xiaoxia Li, *[abc] Mo Zheng, [abc] Chunxing Ren [abc] and Li Guo [abc]

The reactive molecular dynamics using ReaxFF provides an effective means to generate global reactions for pyrolysis of realistic fuel mixtures. The reactions from large-scale pyrolysis simulations of a fuel mixture may be characterized by multiple reaction sites, explosion of intermediate species structures, and scattered contribution of diversified pathways to product species. This work proposes an approach of SRG-Reax aiming at generating skeleton reaction networks based on reaction patterns or classes of reaction centers from huge reactions obtained from ReaxFF MD simulations of realistic fuel pyrolysis. SRG-Reax (Skeleton Reaction network Generation for ReaxFF MD) is implemented through building a semi-supervised machine learning model of tri-training for predicting the reaction classes of pyrolysis reactions based on an extended reaction center. Three different reaction center descriptions of reaction features and reaction transformation fingerprints are employed as inputs for developing the tri-training classifier. Major reaction pathways can be identified based on reaction class ratios and product species ratios calculated by merging reaction pathways of the same reaction class. The SRG-Reax approach was applied in skeleton reaction network generation for RP-3 pyrolysis based on the ReaxFF MD simulations of a high-fidelity 45-component RP-3 fuel model. The skeleton reaction networks for *n*-paraffins, iso-paraffins, cycloparaffins, olefins, and aromatics in RP-3 pyrolysis were obtained. The reaction class ratios and product species ratios in the obtained skeleton reaction network provide comprehensive intuitive insight into global pyrolysis chemistry. SRG-Reax has the potential to obtain relatively complete skeleton reaction networks for the pyrolysis of hydrocarbon fuel, polymers, biomass, coal, and more.

Received 6th December 2023,
Accepted 16th January 2024

DOI: 10.1039/d3cp05935a

rsc.li/pccp

1. Introduction

Pyrolysis chemistry is important both in engine combustion and industrial utilization of various fuels.[1–3] The acquisition and reduction of a complex global reaction network in hydrocarbon fuel pyrolysis have been important topical interests in the fuel community due to their essential roles in fuel pyrolysis. The capability of the computational methods for obtaining reactions and kinetics of fuel pyrolysis has improved progressively in recent years with the constantly increased power of computers.[4,5] Reactive molecular dynamics simulations (rMD) using the first-principles-based ReaxFF force field (ReaxFF MD) have quickly evolved into a new computational approach[6,7] for obtaining the dynamic evolution of reaction networks in pyrolysis for real hydrocarbon fuel molecules. ReaxFF MD can be performed several orders of magnitude faster in calculating energies and forces compared to quantum mechanical methods (QM) with close accuracy to the widely used DFT method.[8,9] Particularly, the significant speed-up of ReaxFF MD simulations allowed by the GPU-enabled code of GMD-Reax[10] or other

[a] *State Key Laboratory of Multiphase Complex Systems, Institute of Process Engineering, Chinese Academy of Sciences, Beijing 100190, P. R. China. E-mail: xxia@ipe.ac.cn*
[b] *School of Chemical Engineering, University of Chinese Academy of Sciences, Beijing 100049, P. R. China*
[c] *Innovation Academy for Green Manufacture, Chinese Academy of Sciences, Beijing 100190, P. R. China*

† Electronic supplementary information (ESI) available: It includes 46 defined reaction classes (RxC) of hydrocarbon pyrolysis, the ReaxFF MD pyrolysis simulation parameters and sampling intervals used for reaction analysis of fuel models for preparing the reaction data set. It also provides a sample input interface for manual check and labeling in the active learning of SRG-Reax, explanations of the reaction features in Input 1 and Input 3 of the tri-training classifier, details of 196 reaction features in the vector of Input 1, and computed results for the 18 types of reaction descriptors in Input 3 for some sample reactions. Additionally, there are feature importance of the top 50 among the 196 features of Input 1, evaluation results of reaction fingerprint (FP) candidates for the single classifier of the tri-training based on the random forest algorithm, along with the confusion matrix of all 46 classes after the tri-training reaction classifier refinement. For a better understanding of the labeled data and inputs for the tri-training classifier, two RAR files containing sample reaction data with typical 46 reaction classes manually labeled and 300 sample reaction data of the training set. See DOI: **https://doi.org/10.1039/d3cp05935a**

GPU-based code[11] makes it practical to directly perform simulations of the pyrolysis reactions of fuel mixture models with dozens of high carbon number components consisting of *n*-paraffins, isoparaffins, cycloparaffins, alkenes, and aromatics.[12–14] The reactions containing full information of reaction sites analyzed with the aid of VARxMD[15] from simulation trajectories demonstrate the advantage of ReaxFF MD in obtaining the dynamic evolution of the complex radical-driven reactions in real fuel pyrolysis, of which applications[2,16] can be found in pyrolysis simulations of many large systems like coal, biomass, polymers, energetics materials, hydrocarbon fuel, *etc.*

The ReaxFF MD simulations allow the overall reactions of radical lifetime species from real fuel molecules into initial radicals to be revealed, as well as the following chain radical growth and chain propagation, and chain termination to generate stable small gas products that may be detectable experimentally. For example, the product species in JP-10 pyrolysis obtained from the ReaxFF MD simulations are basically consistent with those detected using various experimental techniques such as batch reactor, jet-stirred reactor, electrically heated tube, annular tubular reactor, flow tube reactor, and shock tube, as reported in the literature. Particularly, the product evolution tendency of methane, ethane, ethylene, propylene, acetylene, allene, 1-butene, propyne, 1,3-butadiene, and cyclopentadiene with temperature is in qualitative agreement with the experimental results of single-pulse shock tube.[14] A summary of comparisons between the apparent first-order Arrhenius parameters fitted from ReaxFF MD simulations and experimental or continuum simulations for hydrocarbon fuels is provided in the literature.[2] These applications indicate that the large-scale ReaxFF MD simulation method is a substantial step forward in theoretical methodology development for unraveling the kinetic behavior of the radical-driven pyrolysis reactions for real hydrocarbon fuel.[2,14]

The featured advantage of the ReaxFF MD method is that no prior knowledge of reaction pathways is required, making it an alternative method for exploring the chemistry space. The processing scheme of ChemTraYzer-TAD (CTY) proposed and developed by Kai Leonhard *et al.*[17] showcases ReaxFF MD simulations as a useful method in searching for possible elementary reactions of relatively simple fuel systems since ReaxFF MD is more generally applicable in generating all relevant reactions than rule-based methods for chemistry space exploration. For large fuel systems close to real fuel of high carbon numbers and many components, the calculation step based on higher levels of theory within the ChemTraYzer-TAD scheme may not be feasible due to the high computational cost of the necessary *ab initio* calculations for the large number of diversified intermediate species involved.

Large-scale ReaxFF MD simulations with its obtained complete reaction list of unique species and full reaction site information[2,15] provide a good theoretical basis for studying the chemistry of fuel mixture species and tracking their evolution to final products for the quantification of reaction kinetics in pyrolysis. However, linking the molecular-level reaction network of the complex pyrolysis process to the macroscopic kinetics of fuel pyrolysis remains a very challenging task.[9] To obtain a simplified kinetic view of a complex pyrolysis process from ReaxFF MD simulation, it is highly desirable to reduce the complex reaction network into a small-scale network of major pathways.

There have been numerous efforts to develop methodologies for mechanism reduction based on elementary reactions.[18–20] One representative and popular method is the skeletal reduction method of DRG or its derivatives[19,21,22] that eliminates unimportant species and reactions based on detailed mechanisms mostly consisting of elementary reactions with kinetic parameters A (frequency parameter) and E (activation energy) of the Arrhenius equation obtained using quantum mechanics calculation.[21,22] The resulting skeletal mechanism or the retained reactions, is a subset of the detailed mechanism in elementary form. Simulation and mechanism reduction of elementary reactions using ReaxFF MD is possible in principle, which was demonstrated by Sun's group,[19] that simplifies the complex reaction network into a skeletal network through analyzing ReaxFF MD simulation data for hydrogen combustion simulations of 400 H_2 and 200 O_2 at 3500 K for 3 ns. One of the key steps in Sun's work for the mechanism reduction is the identification and rate constant calculations of elementary reactions, where the rate constant calculation is achieved by employing the method proposed by Leonhard's group.[23] The efforts of Leonhard's group for rate constant calculation and Sun's group in establishing a skeletal network through directly introducing available reduction methods of DRGEP and CSP-Index to ReaxFF MD simulation results made a step forward to extract chemical kinetics directly from atomistic simulation data of ReaxFF MD. However, obtaining global elementary reactions for large-scale pyrolysis simulation of a complex hydrocarbon fuel mixture is not that feasible. The reactions from large-scale pyrolysis simulations of a complex hydrocarbon fuel mixture may be characterized by multiple reaction sites due to a large sampling interval frequently used in the output trajectory of ReaxFF MD simulations for practical computational cost considerations. Moreover, the explosion of intermediate species and their structures, and the consequent diversified pathways that make their contributions to pyrolysis product species scattered. The situation motivates us to develop an alternative approach for obtaining a skeleton reaction network merely based on the reactions from ReaxFF MD simulation through reaction classification of reaction centers in order to peak at a simplified kinetic view of a complex pyrolysis process.

The idea of obtaining a skeleton network based on reactions of ReaxFF MD alone is inspired by the fact that the reactions already occurred in a simulation box carry the system chemical kinetics implicitly in the form of sampled reactions over time. Lumping reaction pathways based on reaction classification of reaction centers should work on the basis that the bond changes reflect the very core information of a reaction, and the characteristics of a reaction and the product structures formed from a reaction largely depend on its reaction centers. To lump similar pathways, particularly the pathways with scattered contributions to targeted products, the diversified

reaction pathways with similar reaction centers can be considered as similar reactions that can be lumped into one reaction class.[24,25] Therefore, important reaction pathways can be identified through reaction classification from huge reactions of ReaxFF MD simulations for fuel pyrolysis. By taking advantage of today's machine learning methods in dealing with a large number of independent variables,[26,27] a machine learning classification method can be developed to automatically classify the reactions from ReaxFF MD simulations into small-scale reaction classes. The skeleton reaction network can be built through merging the reaction pathways of the same class.

With the intention to reduce the complex and large detailed reaction network from ReaxFF MD simulations into skeleton reaction networks for fuel pyrolysis, this work presents a new approach titled SRG-Reax (Skeleton Reaction network Generation for ReaxFF MD) based on reaction class prediction using a semi-supervised machine learning (tri-training) classification model. Section 2 focuses on the methods and processing algorithms in the SRG-Reax approach, including the construction strategy of a machine learning classifier of tri-training for reaction class prediction, three inputs of different reaction descriptions for training the classifier, classification accuracy evaluation and improvement, automatic reaction subnetwork searching strategy, reaction network reduction based on the reaction class ratio and product species ratio through the combination of automated merging pathways of the same predicted reaction class and manual reducing species/pathways of negligible reaction classes. Section 3 presents the SRG-Reax application results of the obtained skeleton reaction networks for the five representative structure categories (*n*-paraffins, iso-paraffins, cycloparaffins, olefins, and aromatics) in the pyrolysis of the high-fidelity RP-3 fuel model containing a 45-component. The last section gives the conclusion.

2. Methods

There are three basic steps in SRG-Reax to obtain the skeleton reaction network from the reactions obtained from ReaxFF MD simulations and VARxMD analysis. The first step is to prepare the reaction dataset labeled with reaction class for reaction classification. The second step is to build a suitable automatic classification model of reaction classes for fuel pyrolysis, of which the model in essence can map specific chemical structures of reaction centers to a given reaction class. The third step is the generation of a reaction network with the information of reaction class ratios and product species ratios that forms the very basis to simplify reaction pathways and species to a certain extent, and finally to obtain the skeleton reaction network. The processing scheme of SRG-Reax established in this work is shown in Fig. 1.

The first step of SRG-Reax for preparation of the reaction dataset consists of a number of sub-steps, including defining reaction classes based on reaction sites, manually labeling reaction classes, reaction data preprocessing, and generating reaction descriptions. Considering the limited availability of

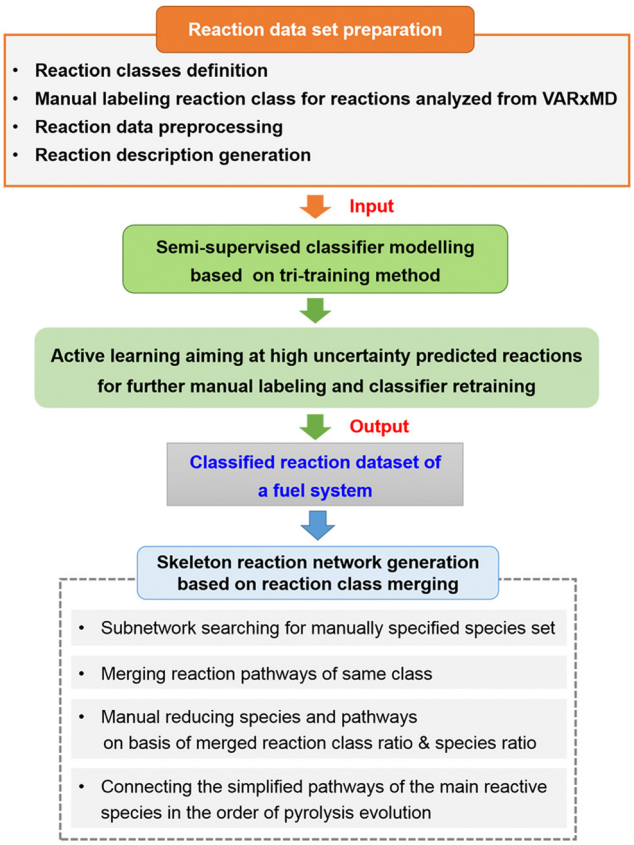

Fig. 1 Processing scheme of SRG-Reax for skeleton reaction network generation of fuel pyrolysis based on ReaxFF MD simulations and machine learning predicted reaction class.

manually labeled reaction data, semi-supervised machine learning[28] based on the tri-training approach[29] was adopted in the second step in order to utilize a large amount of unlabeled data and a small amount of labeled data, which was incorporated with an active learning strategy to pick out high uncertainty[30] predicted reactions for further manual labeling that in return improve the learning performance (prediction accuracy) of the semi-supervised classifier. The automatically predicted classes of all reactions of a fuel system by the trained reaction classification model are the very input for obtaining the skeleton reaction network.

To control the complexity of reaction networks and to facilitate the reduction of reaction networks, a skeleton reaction network can be built based on connecting the subnetworks of main reactive species in the order of pyrolysis reaction evolution. Subnetworks can be obtained by manually specified species that may be a set of source fuel molecules, intermediates, or final products in a fuel pyrolysis system. By merging reaction pathways of the same class, both the reaction class ratios and product species ratios are readily available for the subnetwork. Based on the two ratios, manual reducing pathways in the reaction subnetwork can be performed. Finally, the goal of obtaining a skeleton reaction network can be achieved in principle by connecting simplified subnetworks.

2.1. Machine learning for reaction classification

2.1.1. Machine learning model. To the best of our knowledge, there is no precedent for automatic classification for voluminous reactions of large-scale ReaxFF MD simulations. In particular, the fact there is no available labeled reaction dataset for modeling the classifiers of reaction classes makes it challenging to create an automated reaction classifier. This work employed the semi-supervised machine learning method, where the reaction classifiers were built based on the tri-training approach[29] for making the best use of a small amount of labeled data. The reaction data were derived from analysis results of ReaxFF MD simulations of representative hydrocarbon fuels using VARxMD, including *n*-dodecane, typical fuel surrogate models of RP-3 (Rocket Propellant-3) containing 3 and 4 components,[31,32] and real fuel representative mixture models of RP-3 containing 45 components[33] and RP-1 (Rocket Propellant-1) containing 24 components.[34] These fuel models cover the five representative structure categories of *n*-paraffins, iso-paraffins, cycloparaffins, olefins, and aromatics in typical hydrocarbon fuels. Most of the pyrolysis simulations for modeling the classifier in this work were validated with literature reported experimental data in terms of pyrolysis kinetics in our previous works; for details please refer to the Arrhenius plots both of the RP-1 and RP-3 pyrolysis kinetics and its comparison with experimental results.[12,13] The reaction data were labeled manually with class tags of the 46 reaction classes defined based on the featured structures of reaction centers. The definitions of these reaction classes can be found in Table S1.1 of the ESI.† The simulation details of these pyrolysis reactions are provided in Table S1.2 (ESI†). 7862 labeled reactions and 9386 unlabeled reactions data were used for tri-training[29] classifier modeling. 70% labeled reactions and all unlabeled reactions form the pool of the training dataset for reaction classification. The remaining 30% of the labeled reaction data are used as the testing dataset. The details for building the reaction data set will be published in a separate paper.

The adopted tri-training method proposed by Zhou and Li[29] for reaction automatic classification is a co-training style semi-supervised learning algorithm that generates three classifiers from a single- or multi-view input of the labeled training set. The important feature of the co-training[35] process of the three classifiers can be used to refine the reaction classifier using the unlabeled reaction data with pseudo-labels-voted from the three classifier predictions in each round of the tri-training process. Thus, the "divergence" among the three classifiers can make better use of unlabeled data. Combining three classifiers should in principle result in a more reasonable decision boundary for reaction classification that reduces the overall error of the final predicted reaction classes.[36–38] Fig. 2 represents a schematic diagram of tri-training modeling of the reaction classifier and active learning in SRG-Reax for automated classifying reactions obtained from ReaxFF MD simulations. The tri-training code was implemented following the pseudocode available in ref. 29. The active learning code was implemented by referencing the code.[39]

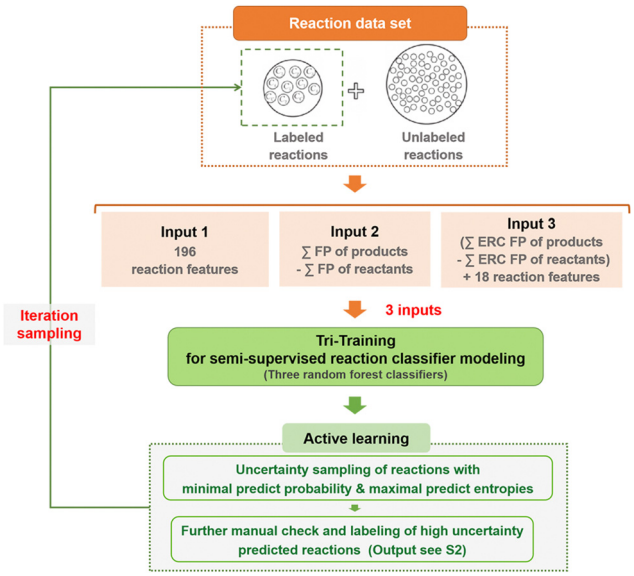

Fig. 2 Schematic diagram of tri-training modeling of the reaction classifier and active learning in SRG-Reax for pyrolysis reactions simulated with ReaxFF MD simulations.

Due to its better performance in reaction classification, the random forest method[40] is employed as the basic classifier of the tri-training model. The code of the random forest classification method was implemented by referencing the code.[41] Actually, the three classifiers of tri-training are characterized by the descriptions of their inputs. Typical reaction descriptions are reaction fingerprints and descriptors derived from structures of reactants and products of a reaction.[42–46] Three different inputs were employed in this work for classifier training to complement the three reaction descriptions with each other as overviewed in Fig. 2. The three inputs are a vector of 196 reaction features, the reaction transformation fingerprint of the full reaction, and reaction transformation fingerprint of the extended reaction center plus 18 reaction features respectively.

The active learning strategy shown in Fig. 2 is combined with the tri-training semi-supervised classification algorithm in SRG-Reax that can filter out the most uncertain samples from the unlabeled reactions, which are critical for further improvement of the tri-training classifier. The uncertainty sampling based on minimal predicted probability and maximal predicted entropies allows for a more targeted human annotator effort with less labeling while complementing the labeled training dataset. The program output of a sample reaction for manual checking and labeling in the active learning of SRG-Reax is shown in the ESI,† S2. The strategy of combining semi-supervised tri-training and active learning can maximize in principle the utilization of labeled and unlabeled reaction data while ensuring manual labeling quality for retraining.

2.1.2. Reaction description. Reaction description is the key to reaction classifier training, which is by no means an easy job, however. In this work, three different inputs representing reaction characteristics were used to generate divergence in the co-trained classifiers of tri-training, which is necessary for

reaction classification where a highly uneven sample distribution exists. Adopting different reaction descriptions is equivalent to capturing reaction features from different perspectives, which is critical for approaching a complete description of reactivity features as possible for a reaction. In principle, the use of different reaction description methods is a complementary strategy to the description of reactive structure features that hopefully improves the utilization of the labeled reaction data for automatic reaction classification. Two strategies for reaction description were employed. One is a combination of multi-level structural and physicochemical reaction descriptions of reactive species both in reactants and products. The other is a kind of reaction fingerprint proposed by N. Schneider, *et al.*[44] calculated from the chemical fingerprints (FP) of reactants and products. The three specific reaction descriptions as inputs were defined for the tri-training classifier on the basis of the two reaction description methods.

The calculation flowchart for calculating the vectors of three inputs of the tri-training classifier is presented in Fig. 3. The flowchart in Fig. 3 provides an overview of the calculation of all features and fingerprints in the three input vectors. Calculation of the hierarchical 4-level reaction description (see Table S3.1, ESI†) in terms of 196 features of Input 1 (Table S3.3, ESI†) are shown in pink in Fig. 3. In addition, the 18 features in Input 3 (see the symbol list in Tables S3.2 and S3.4 (ESI†) for sample data of the 18 features) are calculated based on the 4-levels features.

The first input of tri-training is highlighted in Fig. 4, where the overview of the vector of 196 features is given. Input 1 is composed of hierarchical reaction descriptions of 4-levels in view of the reaction center (bond features as Level 1 and atom features as Level 2), extended reaction center (features of all neighboring atoms and function groups of a reaction center as Level 3) and full reaction (features of a full reaction as Level 4). The reaction features of Input 1 can be obtained from the reaction description in atom-mapped SMARTS[47] by obtaining the corresponding unique atom IDs[15] of reaction sites analyzed from VARxMD. All these reaction features and their descriptions in Input 1 are listed in the ESI,† S3. The features of bond order, number of broken bonds and formed bonds, number of lone pair electrons, and net charge are taken from ReaxFF MD trajectory files. The other features of reactions are calculated using RDKit[48] for aromaticity, number of free radicals, atom mass, valence, *etc.* The 4-level reaction description is crucial for distinguishing the classes of reactions. For simplicity, RxC is used in the representation of reaction class. For example, the NumBridgeRings in bond features of the reaction center (Level 1) describes how many rings the broken or formed bond

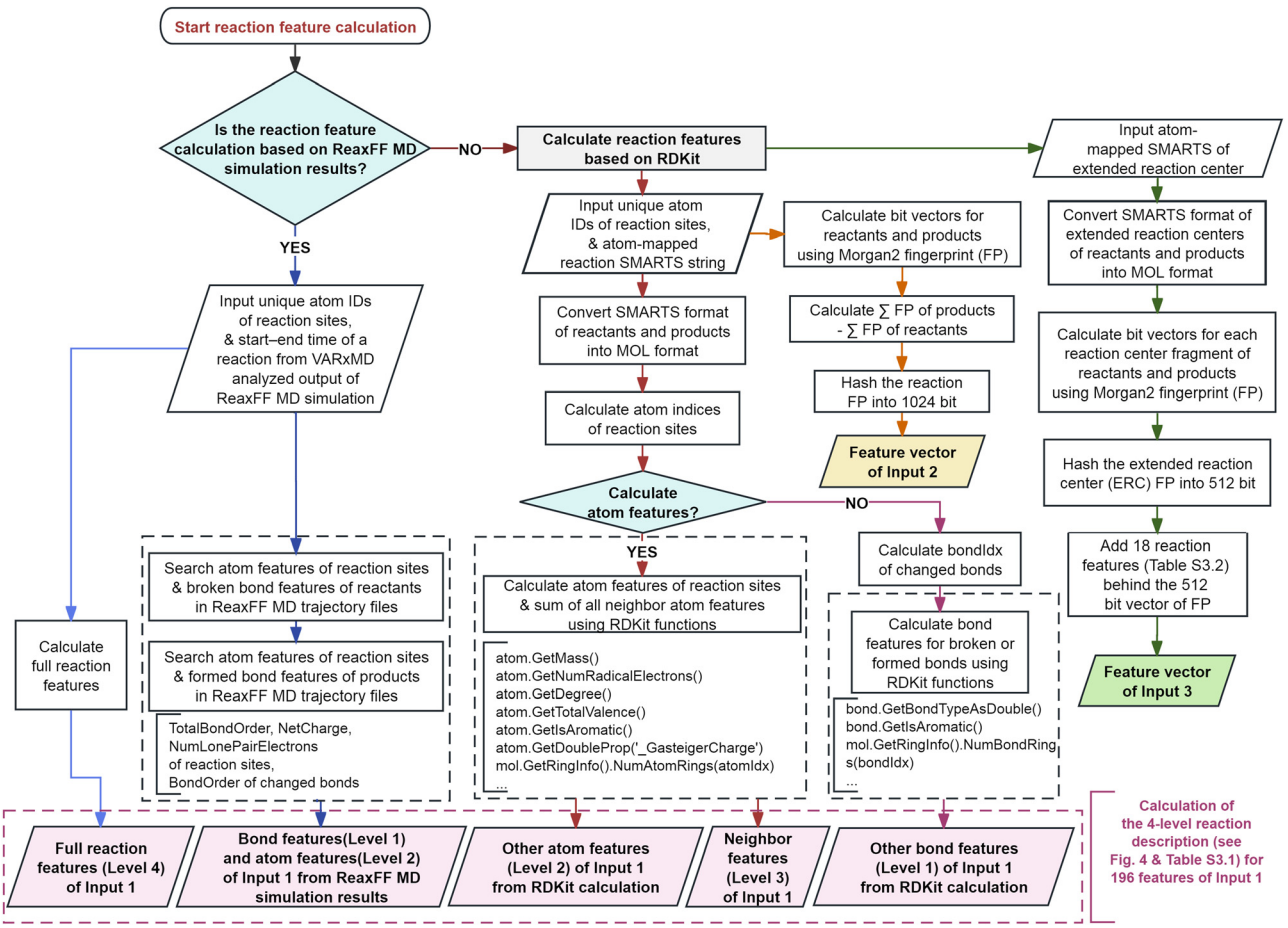

Fig. 3 Calculation flowchart for all features and fingerprints in the three input vectors of the tri-training classifier.

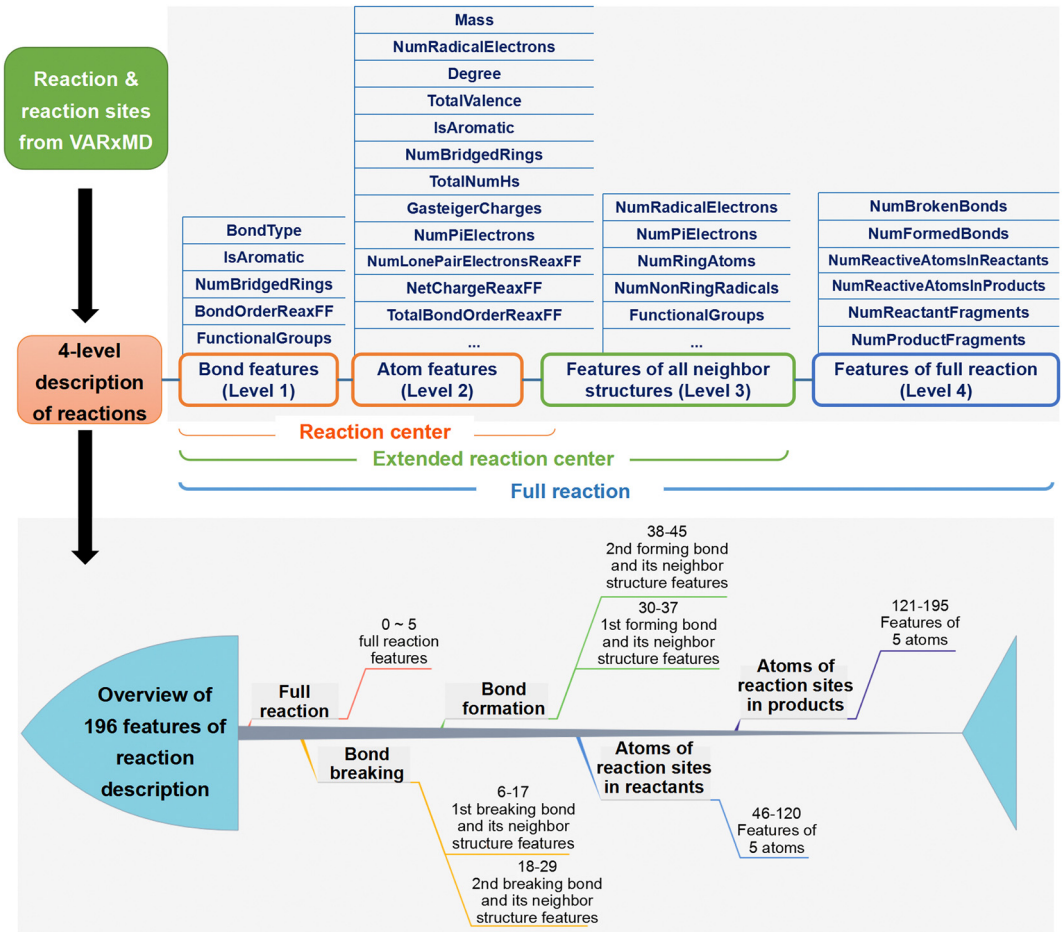

Fig. 4 Reaction description of Input 1 with 196 features for the tri-training classifier modeling.

belongs to, which is very important for discriminating the reaction classes of bridge bond breaking (RxC 20) and bridge bond formation (RxC 40) in polycycle. Another example is related to the features of neighboring structures (Level 3) of a reaction center. The values of NumRadicalElectrons, NumNonRingRadicals, and NumRingAtoms are necessary to indicate the difference among β-breaking reactions of chain hydrocarbons (RxC 3), ring-opening reactions induced by ring radicals (RxC 22), and ring-opening reactions induced by ring branched radicals (RxC 23). The importance of the features in Input 1 can be measured by the single random forest classifier (RFC) adopted in tri-training based on mean decrease impurity.[40,49] For reference, the importance of the top 50 features among the 196 features of Input 1 are provided in the ESI,† Fig. S4.

Input 2 is the reaction transformation fingerprint (FP) in the form of ΣFP of products-ΣFP of reactants, which describes the overall reaction structure changes with the 1024-bit Morgan2[48] fingerprint (equal to ECFP4[50]) of reactants and products.[44] Input 3 is based on a combination of the reaction fingerprint of the extended reaction center of reactants and products with features of reaction transformation. To save computational cost, 512-bit Morgan2 fingerprints of the extended reaction center (ERC FP) were used. Furthermore, with the 18 full reaction descriptors, the 512-bit Morgan2 fingerprint is able to achieve good classification performance in the classifier training process. The 18 features and sample reaction data of the 18 types of reaction descriptors are listed in the ESI,† S3. The evaluation results of reaction fingerprint (FP) candidates (Morgan2,[48,50] AtomPairs,[51] and TopologicalTorsions[52]) of Input 2 and Input 3 for the random forest single classifier of tri-training are listed in the ESI,† Table S4. The evaluation results in Table S4 (ESI†) indicate that both the fingerprint type and FP bit size of a fingerprint affect classification performance. Significant effects can be found for Input 3 with or without the 18 reaction features. The appropriate combination of reaction features and reaction fingerprints can help complement each other for sufficient reaction character description, thus improving the reaction classification performance.

2.1.3. Evaluation and improvement of the classification performance of the machine learning model. One of the most significant concerns of a machine learning classifier is its actual performance for classifying reactions. In this work, the confusion matrix and the *F*1-score are used to evaluate the classification results of each specific reaction class and the overall classification accuracy of the reaction classifier in hydrocarbon fuel pyrolysis.

For conveniently inspecting the classification accuracy of each reaction class, the confusion matrix[53] between the true labeled class and the predicted class was used. The confusion matrix provides a distinct means to identify wrongly classified reactions, and with this class a reaction is confused by the reaction classifier. This type of confusion matrix plays an important role in helping improve the reaction description and correcting reaction labeling. Reaction classification performance can be refined based on the confusion matrix of a classifier as displayed in Fig. 5(a), where the misclassification rates are particularly high for reactions of chain isomerization (RxC 5), detachment of isopropyl radical (RxC 16), and β-ring opening induced by a branched carbon radical (RxC 23). It can be observed that the reactions of each 'true class' in these poorly classified reaction classes are often mispredicted into several specific classes. By checking the tags of the manually labeled original reaction data and the reaction description definition, the classifier improvement can be made by relabeling the corresponding reactions and introducing new reaction descriptions in features of Input 1 as detailed in Table 1. The two poorly classified reaction classes of RxC 16 and RxC 23 in Table 1 are associated with the absence of some important reaction features that influence the predictive accuracy of the reaction classifier. Thus, feature improvement of Input 1 was achieved *via* the thread of feature performance prompted in the confusion matrix through introducing reaction descriptions of the function group in Input 1 and introducing non-ring radicals of neighboring structures of a reaction center into the environment features of Input 1 (NumNonRingRadicals in the level 3 of reaction description) as shown in Table 1. The effectiveness of feature improvement is indicated for RxC 16 and RxC 23 through the column values of prediction accuracy before/after refinement in Table 1. Better classification performance can be confirmed with the confusion matrix of the refined classifier as shown in Fig. 5(b). Fig. 5(b) shows the prediction results of 29 classes with more than 20 labeled samples. The complete classification results of the total 46 reaction classes are provided in the ESI,† S5. Among the reaction classification results, the classes with poor classification accuracy are those with very few samples in the training set.

*F*1 score[53] is an index used to measure the accuracy of classification models in statistics, which takes into account both the accuracy and recall of classification models. In multi-class classification problems, Micro-F1 gives equal weight to each sample when calculating the average, while macro-F1 gives equal weight to each class, which is useful for determining poorly classified categories, especially for datasets with unbalanced samples.

The *F*1 score is used to characterize the reaction classification performance of 3 individual random forest classifiers that serve as the basic classifiers of the tri-training model, the tri-training classifier of the combination of the 3 individual random forest classifiers, and the final tri-training classifier after multiple rounds of active learning through iteration sampling and re-labeling of the reaction data. Table 2 shows a case of the application of the *F*1 score in evaluating the

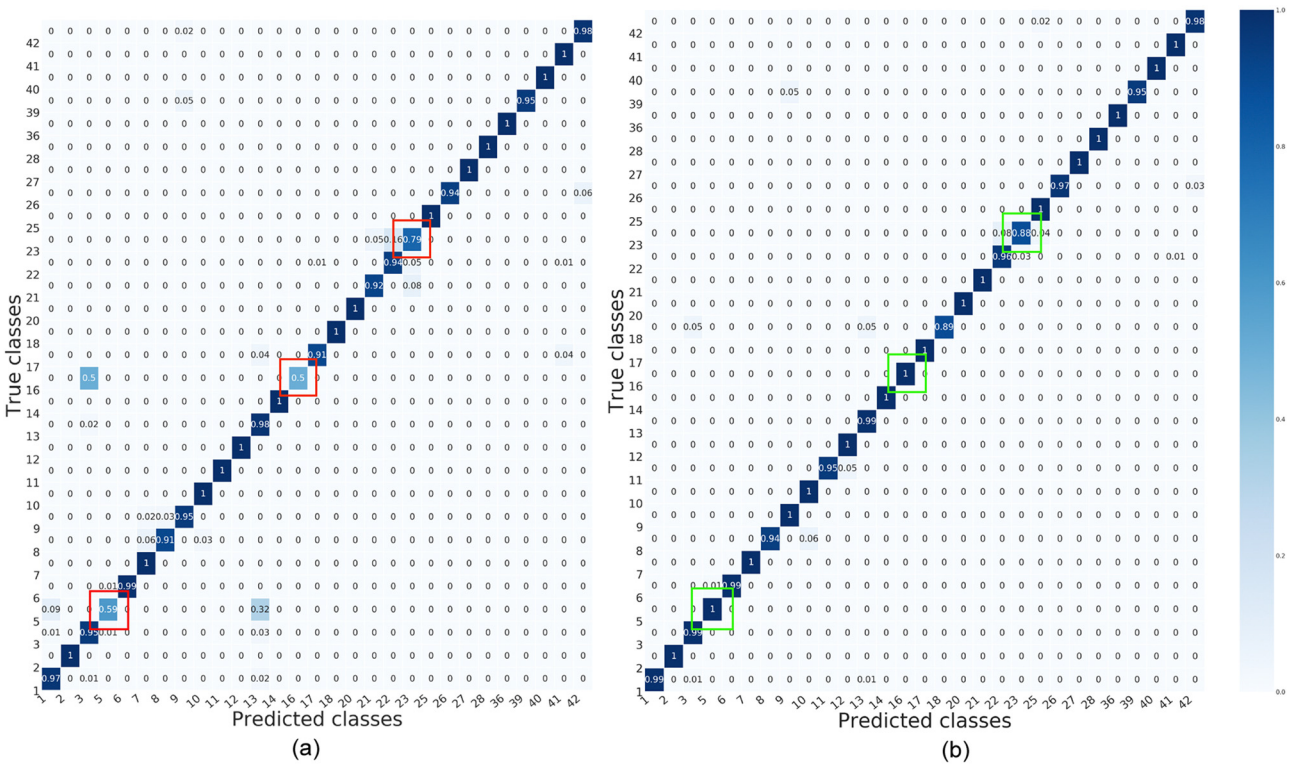

Fig. 5 Confusion matrix for refined evaluation of the reaction classifier in SRG-Reax (a) before reaction class refinement; (b) after reaction class refinement.

Table 1 Summary of classifier improvement by adding new reaction descriptions and relabeling the corresponding reaction

Poorly predicted reaction class	Prediction accuracy before/after refinement	Wrongly predicted classes	Note for the wrong prediction	Refinement performed
RxC 5 (chain isomerization)	0.59/1	32% RxC 5 reactions misclassified as RxC 13 (recombination of C radicals), 9% RxC 5 as RxC 1 (C–C bond homolysis)	Incorrect labeling of the original reactions in RxC 5	Correcting wrong reaction labels
RxC 16 (detachment of isopropyl radical)	0.5/1	50% RxC 16 reactions misclassified as RxC 3 (β-scission)	 The incomplete feature description of Input 1 leads to the identification failure of isopropyl structures with multiple radicals. Original SMARTS for structure identification: '*-[C;D3]([C;D1;H2])([C;D1;**H3**])'	The function group identification of isopropyl structures was extended to include any number of radicals on a branch in Input 1 as depicted in the improved SMARTS description: '*-[C;D3]([C;D1;H,H2])([C;D1;**H,H2**,H3])'
RxC 23 (ring opening induced by β-branched carbon radical)	0.79/0.88	16% RxC 23 reactions misclassified as RxC 22 (β-ring opening induced by ring carbon radical)	The features of Input 1 in level 3 only include statistics of free radicals on the ring but on the ring branch	Adding statistics of non-ring radicals (NumNonRingRadicals) of neighboring structures of a reaction center into the environment features in level 3

classification performance improvement among the 3 individual random forest classifiers with three different inputs, the tri-training classifier before active learning and the tri-training after multi-round active learning iteration. The optimized parameters for the best performance of the single random forest classifier in Table 2 were obtained by performing a GridSearch on each RandomForestClassifier() within the tri-training method. A 5-fold cross-validation (cv = 5) on the training set was performed within a GridSearch using parameters: '*n*_estimators' within the range of 50 to 200 in steps of 10, 'max_features' in the set [0.2, 0.3, 0.4, 0.5, 0.6, 0.7], and 'max_depth' within the range of 5 to 40 in steps of 5.

2.2. Generation of a skeleton reaction network on the basis of classified reactions

The well-trained tri-training reaction classifier in the SRG-Reax method can predict reaction classes for each of the reactions obtained from ReaxFF MD simulations of hydrocarbon fuel pyrolysis. The reaction class predictions can be served directly for the automated construction of a complete reaction class tagged network (RxCN) that includes all reaction class tagged pathways between reactant and product species in a pyrolysis system. The RxCN differs from a conventional reaction network. Each reaction pathway of RxCN is assigned a predicted reaction class label by SRG-Reax. Fig. 6 shows an automatically generated RxCN of the 45-component RP-3 model, where the network is reduced directly based on the occurrence frequencies of reaction pathways using Gephi (visualization and exploration software for graphs and networks).[54] The node of RxCN in Fig. 6 is depicted with the chemical formula of a species, where the integer in parenthesis of the formula is the unique species ID[15] of a ReaxFF MD simulation system generated by VARxMD. All the species in RxCNs in Fig. 8–12 are represented in the format of chemical formula (unique species ID).

Because the 45 components in the RP-3 model account for 98% of the composition of real RP-3 fuel,[13] the reaction pathway understanding from the pyrolysis RxCN identified in its simulations is relatively complete and is closer to the understanding reported from pyrolysis experiments. Unfortunately, the complete pyrolysis RxCN of the 45-component RP-3 fuel is extremely complex and has 722 reactive species and 2205 reaction pathways, as shown in Fig. 6(a). Based on the occurrence frequencies of reaction pathways, the simplified RxCN with the top 15% of reaction pathways is still very complex (120 reactant species and 329 reaction pathways) as shown in Fig. 6(b). The simplified RxCN with the top 5% reaction pathways shown in Fig. 6(c) is relatively

Table 2 Performance comparison of a single classifier (random forest) adopted in tri-training, tri-training classifier, and tri-training classifier after iterative active learning

	Best performance and parameters of the single classifier (random forest) adopted in tri-training	Tri-training classifier before active learning		Tri-training classifier after active learning	
Reaction descriptions	Micro-*F*1	Micro-*F*1	Macro-*F*1	Micro-*F*1	Macro-*F*1
Input 1 196 reaction features	0.962 ('*n*_estimators': 200, 'max_depth': 10, 'max_features': 0.3)	0.968	0.907	0.989	0.927
Input 2 $\sum$ FP of products $-\sum$ FP of reactants	0.959 ('*n*_estimators': 160, 'max_depth': 40, 'max_features': 0.3)				
Input 3 ($\sum$ ERC FP of products $-\sum$ ERC FP of reactants) + 18 full reaction features	0.953 ('*n*_estimators': 50, 'max_depth': 30, 'max_features': 0.2)				

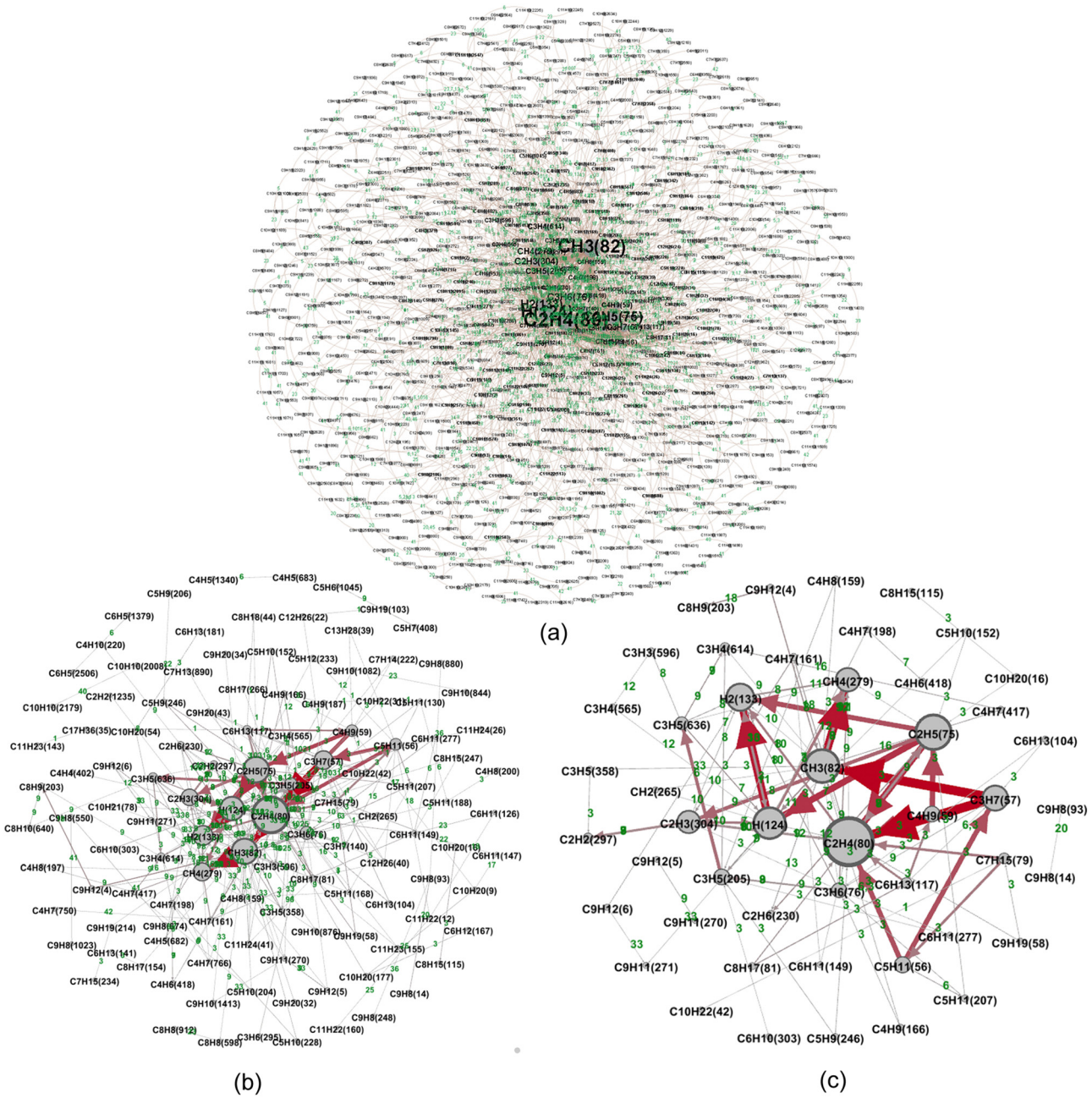

Fig. 6 Detailed reaction class tagged network (RxCN) generated automatically by SRG-Reax for pyrolysis reactions of the 45-component RP-3 model based on predicted reaction classes: (a) detailed RxCN of all reaction pathways; (b) and (c) simplified RxCN obtained by selecting the top 15% pathways and top 5% of reaction pathways.

simple with 48 reactant species and 118 reaction pathways, but it no longer has good coverage of the important reaction pathways for the pyrolysis of original fuel components, where most of the reaction pathways of cycloparaffin and aromatics have been filtered out. The results of Fig. 6 indicate that it is not practical to obtain a reasonable reduced reaction network by directly simplifying the pyrolysis reaction network using tools like Gephi based on the occurrence frequencies of reaction pathways.

Due to the diversity of pyrolysis reactions among a large number of reactive species and the molecular model size limitations of ReaxFF MD simulations, many of the reaction pathways with similar reaction centers have poor statistical significance, resulting in the indiscriminate filtering within the simple pruning process of a reaction network. Thus, simple pruning of reaction pathways may lead to the loss of meaningful pathways for understanding the pyrolysis reaction mechanisms.

The approach to obtain the skeleton reaction network proposed in this work is mainly based on merging reaction pathways of the same reaction class predicted with the machine learning method. The most direct presentation of reactions with predicted class after reaction classification in SRG-Reax is

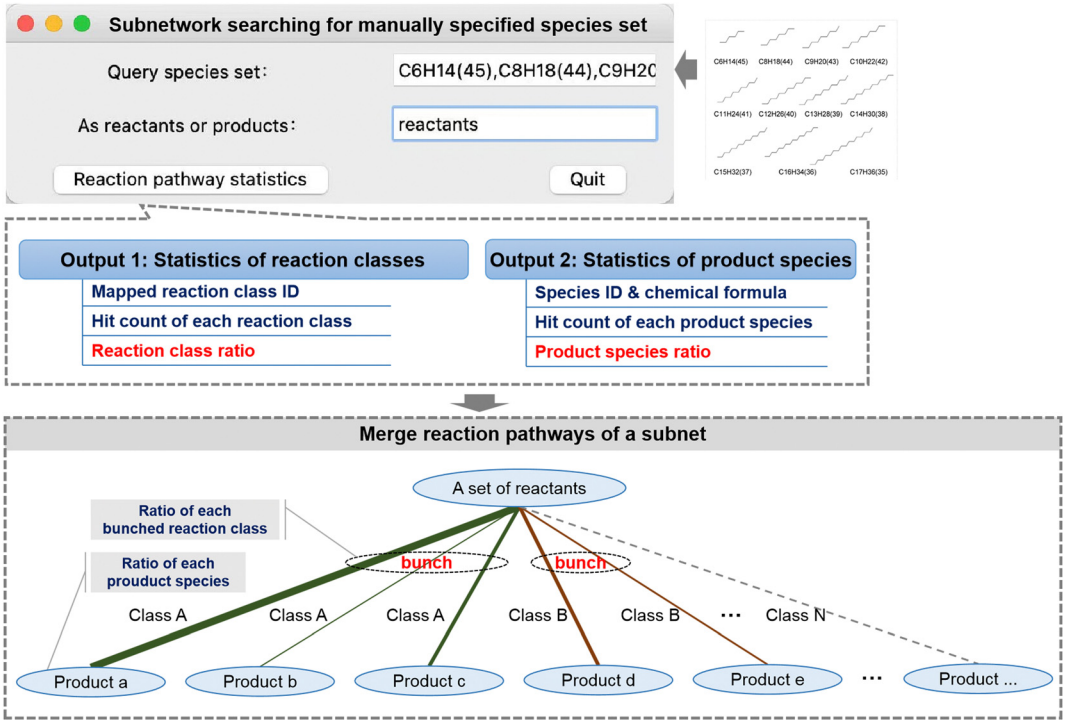

Fig. 7 Automatic reaction subnetwork search method for a set of pyrolysis species specified.

a detailed reaction class tagged network (as RxCN), while the skeleton reaction network can only be obtained after further systematic statistics of merged reaction pathways and manual reaction reduction based on reaction class ratio and relevant product species ratio. The reaction class ratio is defined as the number of pathways for a specific reaction class of the reactant species divided by the number of pathways of all reaction classes for the reactant species of the sub-network to measure the contribution of a specific reaction class in a sub-network. Thus, the reaction class ratio represents the weighting of a single reaction class among all reaction classes in a subnetwork. To further account for the contribution of a single reaction class to the formed products in a sub-network, the product species ratio for a single reaction class is defined as the number of the corresponding product species divided by the total number of all products of the reactant species.

The key step of automatically obtaining the statistics of reaction classes and intermediates/product species of a reaction network node is to build an automatic reaction subnetwork searching method for reactant species. As shown in Fig. 7, the first step to obtain the statistics of reaction classes and intermediates/product species of a reaction network node is to prepare manually a set of targeted reactant species in the form of species ID. To clearly exhibit the relationship among reactant species, reaction pathways, and reaction classes, the method for constructing this type of skeleton reaction network in this work involves traversing and searching all reaction pathways starting from the specified set of reactant species ($N \geq 1$). The reactant species can be a single fuel species, a representative category of fuel species, or all fuel species.

The starting point of skeleton reaction network construction is the fuel component species. The skeleton reaction network of RP-3 pyrolysis is constructed by subnetwork searching along the direction of pyrolysis reaction evolution. Accordingly, the reaction class ratio of each pathway class and the corresponding product species ratio of each product will be calculated. The pathways of the same reaction class will be merged as one bunched pathway. The pyrolysis product species (intermediates) obtained in the current round of subnetwork searching will be the search input of the next round. The subnetwork searching uses a breadth-first search strategy by traversing all reactant species combined with manual reduction based on the calculated reaction class ratios and corresponding product species ratios. A manual reduction strategy is introduced by removing pathways of less important class and less important product species of current round results before the next round of subnetwork searching, which avoids the loss of meaningful pathways in the simple pruning of reaction pathways as discussed in the results of Fig. 6. The choice of introducing manual reduction for a skeleton reaction network is based on practical consideration because the diversity of product species associated with a certain reaction pathway in RP-3 pyrolysis makes it hard to choose appropriate cut-offs in pruning reaction pathways along the complex pyrolysis reaction evolution.

3. Results and discussion

Typical hydrocarbon fuels generally cover five representative structure categories of *n*-paraffins, iso-paraffins, cycloparaffins, olefins, and aromatics. The contribution of each component

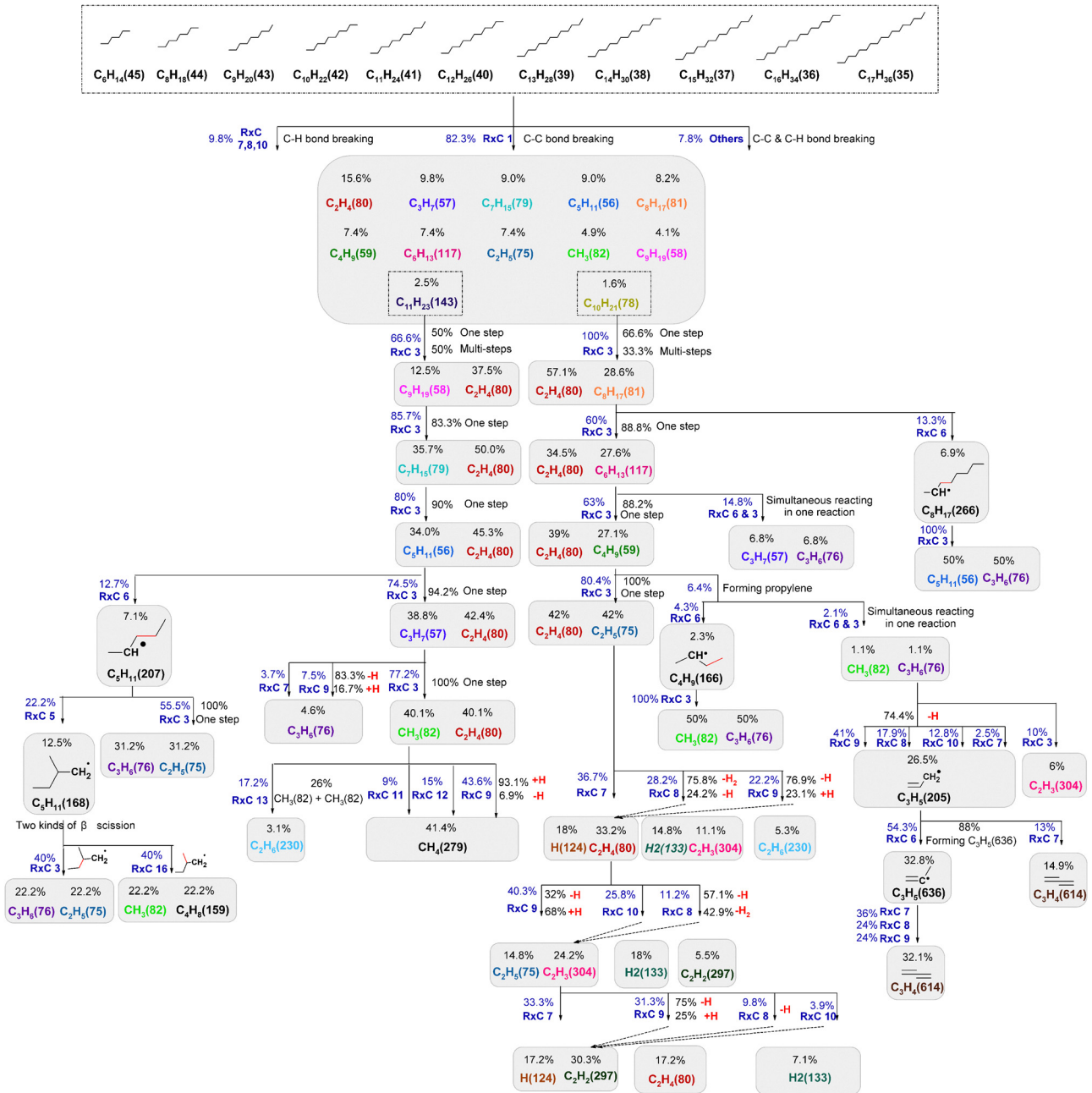

Fig. 8 Skeleton reaction network of *n*-paraffins in pyrolysis of a 45-component RP-3 model obtained on the basis of pathway merging using a predicted reaction class by tri-training classifier in SRG-Reax, where the integer in parenthesis of a species formula is the ID of a unique species.

category to reaction pathways forms most of the reaction space of fuel pyrolysis. SRG-Reax was applied to the pyrolysis reactions of the 45-component RP-3 model which is a high-fidelity molecular model of RP-3 fuel. In order to facilitate analysis of the contributions and characteristics of representative component structure categories to the skeleton reaction networks in RP-3 pyrolysis, the generation of a skeleton reaction network was performed on the basis of component structure categories. The five component structure categories in the 45-component RP-3 fuel model are shown in Table 3. The reaction classes or species in the skeleton pyrolysis reaction networks generated by SRG-Reax agree with the available mechanism in terms of reaction classes and some species detected experimentally in the literature, which is summarized in Table 4.

The skeleton pyrolysis reaction networks obtained by SRG-Reax are characterized by the reaction class ratios and corresponding intermediates/product species ratios. For simplicity of the skeleton reaction network representation, it is designed in this work that the specific pathways of a reactant species appear only once in a prominent position of the reaction network. The prominent position was chosen to exhibit the stepwise characteristics of the chain reaction process in pyrolysis.

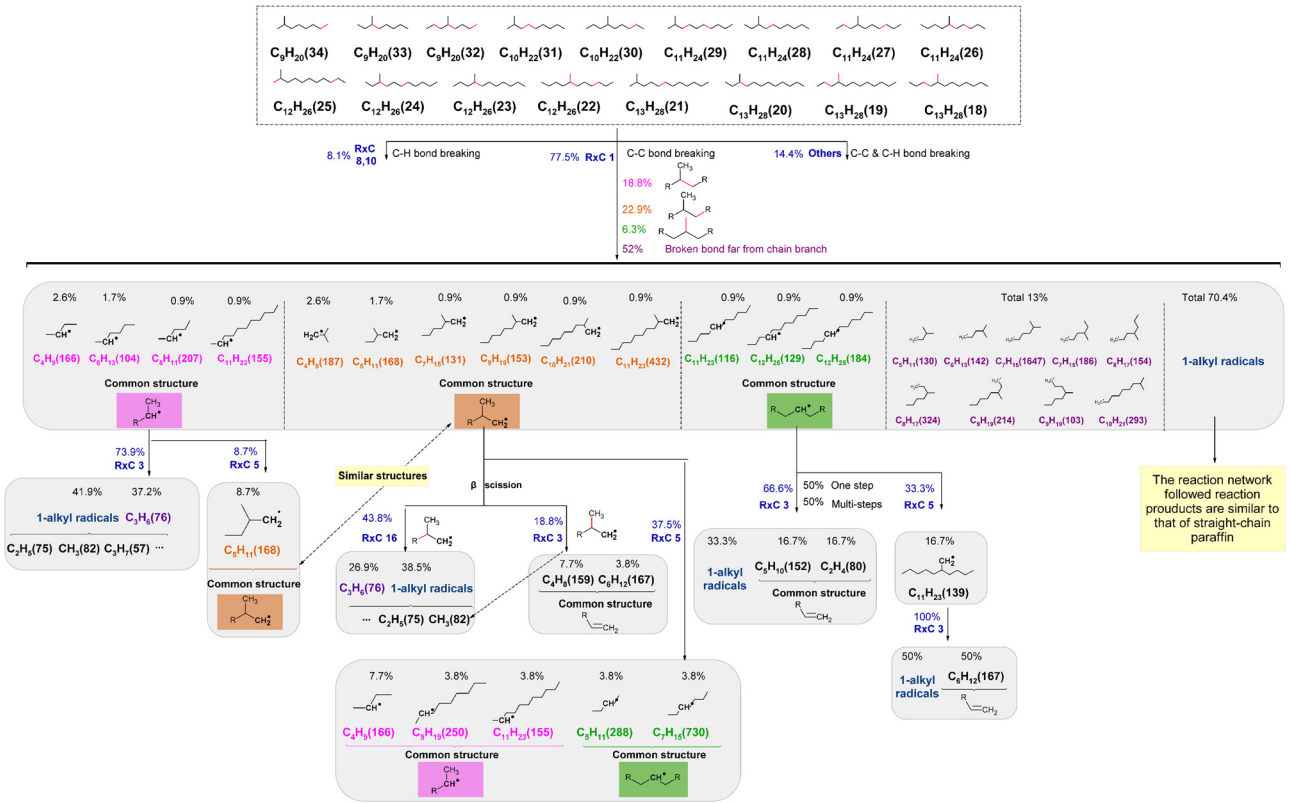

Fig. 9 Skeleton reaction network of iso-paraffins in pyrolysis of the 45-component RP-3 model obtained on the basis of pathway merging using the predicted reaction class by tri-training classifier in SRG-Reax, where the integer in parenthesis of a species formula is the ID of a unique species.

3.1. Skeleton reaction network of *n*-paraffins in RP-3 pyrolysis

The identified skeleton reaction network of *n*-paraffins in pyrolysis of the 45-component RP-3 model is depicted in Fig. 8. The hierarchical layout of the reaction network follows the pyrolysis evolution pathways of reactive species. The simplified reaction network in Fig. 8 illustrates the main reaction pathways, reaction classes, and species involved in the pyrolysis of *n*-paraffin components in RP-3 fuel.

As shown in Fig. 8, most of the initial pyrolysis reactions of *n*-paraffins are C–C bond homolysis, with a reaction class label of RxC 1. The class ratio of RxC 1 is 82.3% which accounts for 82.3% of all reactions during initial pyrolysis. The obtained products with the highest carbon numbers through RxC 1 are $C_{11}H_{23}$ (143) and $C_{10}H_{21}$ (78) shown in the small dotted rectangle, where the 143 and 78 in parenthesis are the unique species IDs for $C_{11}H_{23}$ and $C_{10}H_{21}$. There are 9.8% reactions of C–H bond breaking through reaction classes of H detachment (RxC 7), H_2 formation *via* dehydrogenation (RxC 8), and H-abstraction of H_2 by C (RxC 11). Due to the large time intervals (1 ps) used in the reaction analysis and the high-temperature reaction conditions of simulation, 7.8% reactions of multiple classes may occur simultaneously between a set of reactants and products.

The subsequent pyrolysis process after the initial C–C bond homolysis (RxC 1) can be categorized into two major reaction classes of merged pathways (RxC 3 and RxC 6), corresponding to the generation of the major products ethylene and propylene. The most common pathway identified for the intermediates with terminal radicals is the continuous β-scission (RxC 3) at the end of the carbon chain to produce ethylene. Additionally, many intermediates with terminal radicals undergo an intra-molecular H-shift (RxC 6, accounting for about 6–15%) before β-scission, resulting in the production of propylene, and eventually generating C_{1-2} small fragments.

The small fragments of C_{1-2} generated from the RP-3 pyrolysis process mainly undergo C–H bond cleavage reactions (RxC 7–13). C_2H_5(75) mainly undergoes reactions of H detachment (RxC 7, 36.7%) and inter-molecular H-abstraction by C radical (RxC 9, 22.2%) to produce ethylene. It is worth mentioning that ~30% of C_2H_5(75) directly undergoes H_2 formation *via* dehydrogenation (RxC 8) to form C_2H_3(304), which can be followed by a H detachment (RxC 7) reaction to form an important precursor of C_2H_2(297) for the formation of aromatics. CH_3(82) generated in the pyrolysis process mainly abstracts H atoms from a carbon chain (RxC 9, 43.6%) to produce CH_4(279), while directly adding H radical to CH_3(82) to form CH_4(279) only accounts for 15% of CH_3 pathways. CH_3(82) can also combine with each other producing ethane (RxC 13, 17.2%). Major pyrolysis products such as ethylene and propylene, can further undergo various types of dehydrogenation reactions (RxC 7–10) in high-temperature pyrolysis systems to generate carbon soot precursor structures with higher C/H ratios, such as C_2H_2 and C_3H_4.

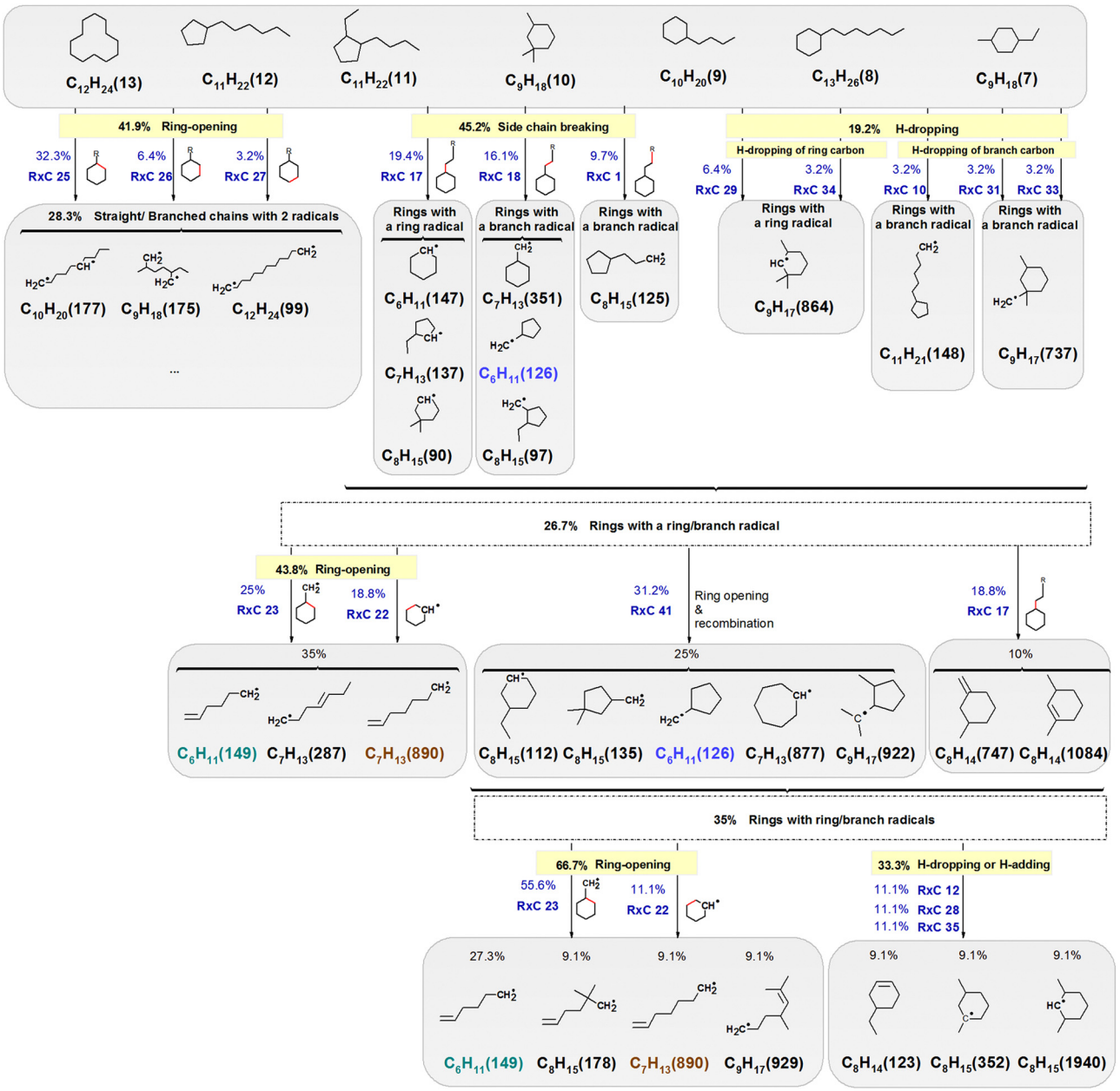

Fig. 10 Skeleton ring-opening reaction network of cycloparaffins in pyrolysis of the 45-component RP-3 model obtained on the basis of pathway merging using the predicted reaction class by tri-training classifier in SRG-Reax, where the integer in parenthesis of a species formula is the ID of a unique species.

The SRG-Reax method allows for the skeleton network in *n*-paraffin pyrolysis to be conveniently obtained with the revealed major reaction classes and relevant species therein. It enables a clear understanding of the main reaction pathways that most frequently occur based on the explicit branch ratio of reaction classes of each reactant species in the skeleton reaction network. Meanwhile, by considering the ratios of product species included in each class of merged reaction pathways, the major intermediate/product species of each reaction class can be identified, and the significance of each specific reaction pathway within the merged pathways can also be traced. These findings for *n*-paraffins are consistent with the knowledge of pyrolysis mechanisms of hydrocarbon fuels obtained from experimental techniques and other computational methods.[55–58,68] The major reaction classes for decomposition (RxC 1, 3), isomerization (H-shift, RxC 6), and H-dropping (RxC 7–10, including H abstraction) of fuel molecules/alkyl radical species in the pyrolysis skeleton network of RP-3 *n*-paraffins in Fig. 8 agree with the important reaction classes outlined by Curran *et al.*[55] in the high-temperature kinetic mechanism of *n*-heptane oxidation, which is validated by experimental techniques of stirred reactors and shock tubes and is well accepted. The main products including methane, ethylene, ethane, and propylene in Fig. 8 show good agreement with the experimental results of *n*-heptane

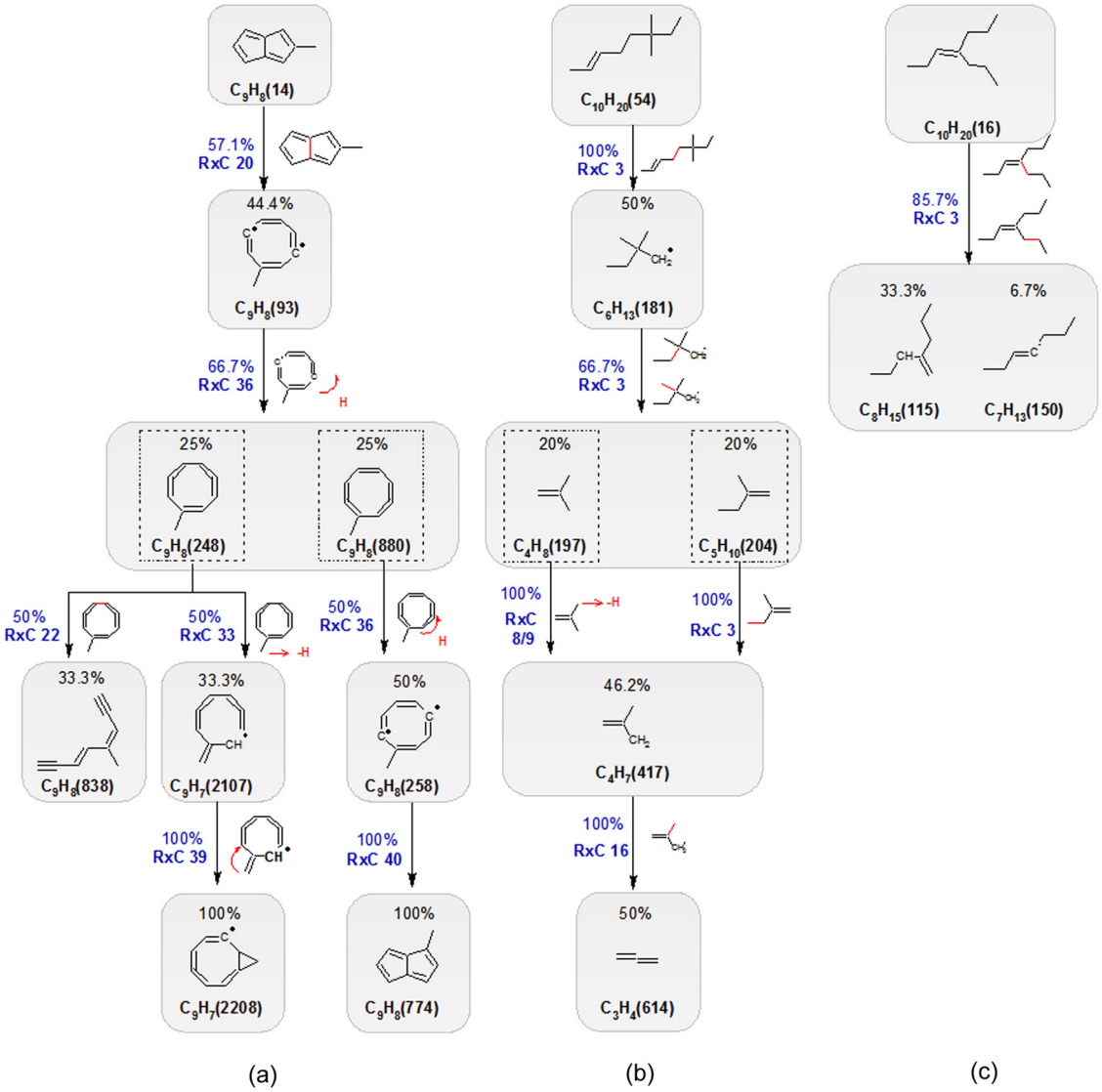

Fig. 11 Skeleton reaction network of olefins in the pyrolysis of the 45-component RP-3 model obtained on the basis of pathway merging using the predicted reaction class by the tri-training classifier in SRG-Reax, where the integer in parenthesis of a species formula is the ID of a unique species. (a) C_9H_8(14) of bicyclic olefin, (b) $C_{10}H_{20}$(54) of chain olefin with 2 methyl branches, (c) $C_{10}H_{20}$(16) of chain olefin with a long branch.

pyrolysis in tubular reactors.[56,57] In particular, the important role of 1,2-H-shift (RxC 6) reactions for propylene generation exhibited in the obtained skeleton network agrees well with the theoretical explanation based on the calculated ring-strain energy of the cyclic transition state of H-shift in Vinu's review[68] and also supported by Aribike, D. S. and Susu, A. A.'s experimental dectection[57] and detailed mechanistic modeling[58] of the pyrolysis of *n*-heptane. More importantly, the skeleton reaction network generated by SRG-Reax not only has good coverage of the characterized reaction classes in *n*-paraffin pyrolysis, but also provides a clear relationship between these reaction classes and important species.

3.2. Skeleton reaction network of iso-paraffins in RP-3 pyrolysis

Iso-paraffins account for 24 mol% of the 45-component RP-3 fuel, and their side chains are all methyl. Due to the low proportion and diverse structures of iso-paraffins, the number of individual pyrolysis intermediate structures generated is very small, and the statistical significance of the merged reaction pathways for individual intermediates is poor. To determine the pyrolysis character of iso-paraffins in its skeleton reaction network, a set of intermediate species with similar reaction centers is collectively used as the input of subnetwork searching, which differs from the input of *n*-paraffins where a single intermediate species is used. In other words, the merging and statistic counting of reaction pathways in the skeleton reaction network for iso-paraffins is the basis of a single node that actually represents a set of intermediate species with similar structural characteristics. The simplified skeleton reaction network for the pyrolysis of iso-paraffins in the 45-component RP-3 model is depicted in Fig. 9.

It can be observed in Fig. 9 that the presence of the methyl branch in the iso-paraffins of RP-3 fuel leads to the formation

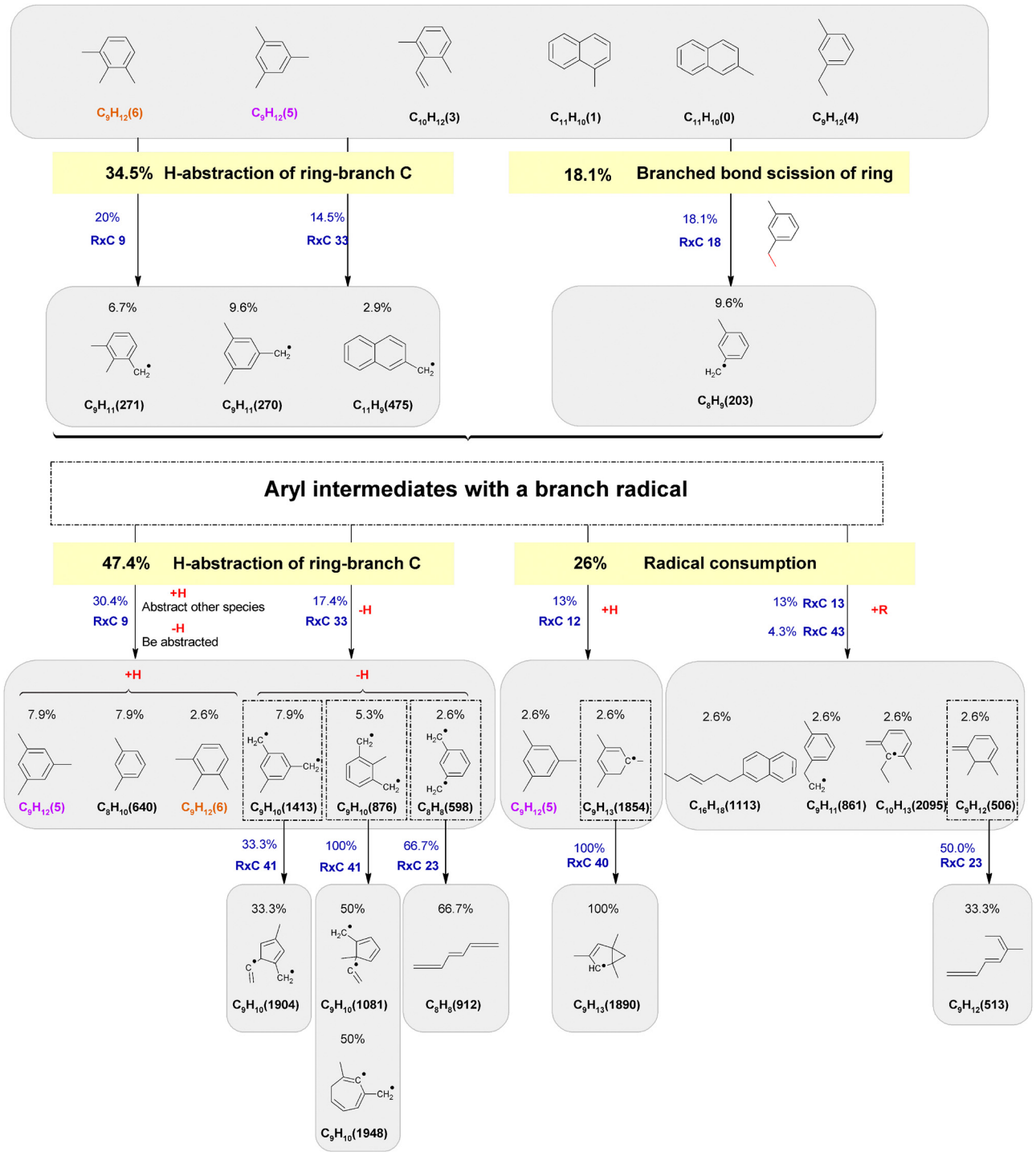

Fig. 12 Skeleton reaction network of aromatic components in the pyrolysis of the 45-component RP-3 model obtained on the basis of pathway merging using the predicted reaction class by the tri-training classifier in SRG-Reax, where the integer in parenthesis of a species formula is the ID of a unique species.

of typical intermediate species of four groups after initial C–C bond homolysis. The structures of the four group species are closely related to the specific position of initial C–C bond cleavage. The first and the second C–C bonds adjacent to the methyl branch break most easily, accounting for 18.8% and 22.9% of the merged reaction pathway class (RxC 1), while the cleavage position at the methyl-branch bond only accounts for 6.3%. These three initial bond cleavage positions correspond to the generation of the three category species of 2-alkyl radical, 2-methylene radical iso-alkane, and non-1,2-alkane radical (CH radical positioned at x other than 1 and 2). For clarity, the formula of the three group species and their reaction center

Table 3 Quantitative proportions of representative structure categories in the 45-component RP-3 fuel model

Molecular count	Component structure ratio (mol%)				
	N-paraffins	Iso-paraffins	Cycloparaffins	Olefins	Aromatics
202	24	24	16	9	27

structures are highlighted in pink, orange, and green respectively in Fig. 9. The remaining 52% C–C bond cleavage positions far from the methyl branch will produce intermediate species of purple formula that leads to the production of intermediates with the same three reaction center structures (pink, orange and green) mentioned above by undergoing more pyrolysis steps, of which the subsequent pyrolysis pathways are not repeated in Fig. 9.

The dominant subsequent pyrolysis reaction class of intermediate 2-alkyl radical (pink) is β-scission (RxC 3, 73.9%) to produce propene and 1-alkyl radical. The subsequent pyrolysis pathways of 1-alkyl radical are similar to those of *n*-paraffins in Fig. 8 and are not repeated in Fig. 9. Besides, the small portion (8.7%) of the subsequent pyrolysis reaction of the 2-alkyl radical (pink) is chain isomerization (RxC 5) that will produce intermediate species with similar structure to 2-methylene radical iso-alkane intermediates (orange) generated from the homolysis at the second C–C bond adjacent to the methyl branch. Accordingly, the subsequent major merged pyrolysis pathway class of 2-methylene radical iso-alkane (orange) is β-scission (total accounting for ~60%) that involves either β-scission (RxC 3, 18.8%) to produce a methyl radical and 1-alkene, or β-scission with detachment of isopropyl radical (RxC 16, 43.8%) to generate propene and 1-alkyl radical. The remaining ~40% of the following pyrolysis pathways of the orange formula intermediates are intra-molecular chain isomerization reactions (RxC 5) to generate the same structures with the background of pink and green as shown in Fig. 9. The pyrolysis reactions of non-1,2-alkane radical with green formula mainly correspond to two classes of reaction pathways. About 66.6% of the pathways will undergo β-scission (RxC 3) to produce 1-alkene and 1-alkyl radical. While 33.3% will undergo intra-molecular chain isomerization (RxC 5) to produce *x*-methylene radical iso-alkane, which will subsequently undergo β-scission (RxC 3) to generate the same product structures obtained directly by β-scission (RxC 3) of the green structures.

Compared to the skeleton reaction network of *n*-paraffin pyrolysis, three characterized intermediates (pink, orange, and green) during iso-paraffin pyrolysis were observed from the skeleton reaction network in Fig. 9. The pathways for the identified three category intermediates that lead to the formation of alkenes, especially propylene, agree with the mechanisms proposed both by Dryer and Brezinsky based on flow reactor study of the oxidation of iso-octane[61] and by Kevin on the basis of experimental species detection in the pyrolysis of various iso-paraffins.[62] The characterized reaction class of iso-paraffin pyrolysis identified is intra-molecular chain isomerization (RxC 5) involving chain cleavage and recombination. These results demonstrate the effectiveness of the SRG-Reax method in gaining a comprehensive understanding of the skeleton pyrolysis pathways and representative intermediate species of iso-paraffins.

3.3. Skeleton reaction network of cycloparaffins in RP-3 pyrolysis

Cycloparaffins account for 16 mol% of the 45-component RP-3 fuel. The identified skeleton reaction network of the cycloparaffin component in pyrolysis is shown in Fig. 10. Due to the low proportion and diverse structures of cycloparaffins, reaction pathways of poor statistical significance are merged based on subnetworks of ring structure fuel/intermediate species set as done for iso-paraffins in Section 3.2.

As shown in Fig. 10, the initial pyrolysis reaction pathways of cycloparaffins in RP-3 pyrolysis can be classified into three superclasses of direct ring opening, side chain breaking, and H-dropping. The dominant initial reactions are ring-opening reactions (41.9%) and side chain breaking reactions (45.2%). The high portion of side chain breaking reactions is related to the presence of long side chains in most cycloparaffin components of the 45-component RP-3 fuel model. Besides, 19.2% of the initial pyrolysis reactions for cycloparaffins are H-dropping reactions of dehydrogenation and H-abstraction from the ring or ring branch. The proportions of the three reaction superclasses identified are roughly consistent with the results obtained by manual statistics in our previous work.[13] In partiuclar, the reaction class ratio ranking (RxC 25 > RxC 17 > RxC 18 > RxC 26 > RxC 27) identified from pyrolysis of various cycloparaffin structures by SRG-Reax for different initial cleavage positions agrees well with the ranking of DFT-calculated bond dissociation energies (BDEs) for ring-opening of ethylcyclohexane (ECH) and bond cleavage of side chains at different positions.[13] This finding provides support for the reasonableness of the core of our method that similar reaction center structures have similar reactivity. Furthermore, it is found that the radicals generated on the ring or ring branch during the initial pyrolysis can induce ring opening (43.8%) in the subsequent pyrolysis to produce 1-alkene radical species through RxC 22 and RxC 23. There are 31.2% ring isomerization reactions through pathways of RxC 41 of ring opening and recombination to produce five/seven-member-ring intermediates, which differs from the reaction classes of the three superclasses during initial pyrolysis of cycloparaffins. The ring isomerization reactions (RxC 41) identified in this work are supported by the formation of the five-member-ring intermediates in the pyrolysis of ethylcyclohexane (ECH) detected using photoionization mass spectrometry (PIMS) and gas chromatography (GC) in Wang's study,[63] as well as using a batch reactor and GC-MS in Dai's work[64] that validates the existence of cyclization processes in cycloparaffin pyrolysis. The generated ring species from the ring isomerization will undergo a similar ring opening process.

3.4. Skeleton reaction network of olefins and aromatics in RP-3 pyrolysis

RP-3 fuel has a low content of olefins. There are three olefin components in the 45-component RP-3 fuel model, namely C_9H_8(14) of bicyclic olefin, $C_{10}H_{20}$(54) of chain olefin with

Table 4 Agreement between reaction classes/species in the skeleton reaction network of RP-3 pyrolysis generated by SRG-Reax and the available mechanism

Component category of RP-3 fuel	Major reaction classes and corresponding species by SRG-Reax	Reported literature and method
N-Paraffins		
	• C–C bond homolysis (RxC 1)	Oxidation/*n*-heptane/shock tubes/GC[55]
	n-Paraffins → 2 1-alkyl radicals	
	• H-shift (RxC 6)	Pyrolysis/*n*-heptane/tubular reactor/GC[56]
	1-Alkyl radical → 2-alkyl radical (1-alkyl radicals are likely to isomerize before they decompose)	Pyrolysis/*n*-heptane/annular stainless steel reactor/GC[57,58]
	• β-scission (RxC 3)	
	1-Alkyl radical → ethylene	
	2-Alkyl radical → propylene	
	• H-abstraction (RxC 9, 10)	
	H-abstraction by H, CH_3, C_2H_3, C_2H_5 → H_2, CH_4, C_2H_4, C_2H_6	
	• H detachment (RxC 7)	
	C_2H_5, C_2H_3, C_3H_5, C_3H_7 → C_2H_4, C_2H_2, C_3H_4, C_3H_6	
	• H radical addition to C (RxC 12)	Pyrolysis/*n*-heptane/GC[57,58]
	CH_3 + H → CH_4	Pyrolysis/*n*-dodecane/PFR/GC-MS[59]
	• Recombination of C radicals (RxC 13)	
	CH_3 + CH_3 → C_2H_6	
	• More complete classes and species see Fig. 8	—
Iso-paraffins		
	• C–C bond homolysis (RxC 1)	Pyrolysis/iso-dodecane/tubular reactor/GC[60]
	Iso-paraffins → CH_3 (methyl branch) + Non-1-alkyl radical	
	• β-scission (RxC 3)	Oxidation/iso-octane/flow reactor/GC-MS[61]
	2-Alkyl radical → propylene + 1-alkyl radical	Pyrolysis/various iso-paraffins/tubular reactor/GC[62]
	2-Methylene radical iso-alkane → 1-alkene + CH_3 (methyl branch)	
	• Detachment of isopropyl radical (RxC 16)	
	2-Methylene radical iso-alkane → propylene + 1-alkyl radical	
	• Intra-molecular chain isomerization (RxC 5)	—
	n-Alkyl radical → iso-alkyl radical	
	Iso-alkyl radical → *n*-alkyl radical	
	• More complete classes and species see Fig. 9	—
Cycloparaffins		
	• Ring-opening (RxC 25–27)	Pyrolysis/ethylcyclohexane/flow reactor/VUV/PIMS/GC[63]
	α-ring bond scission of ring branch (RxC 25)	Pyrolysis/ethylcyclohexane/batch reactor/GC-MS[64]
	β-ring bond scission of ring branch (RxC 26)	
	Non-α,β-ring bond scission of ring branch (RxC 27)	
	• Side chain breaking (RxC 1, 17–18)	
	α-branch bond scission of ring carbon (RxC 17)	
	β-branch bond scission of ring carbon (RxC 18)	
	Non-α,β-branch bond scission of ring carbon (RxC 1)	
	• H-dropping of ring carbon (RxC 29, 34)	
	Ring dehydrogenation → H2 (RxC 29)	
	H-abstraction by acyclic radical (RxC 34)	
	• H-dropping of ring branch carbon (RxC 10, 31, 33)	
	H-abstraction of non-a-C on ring branch by H (RxC 10)	
	Dehydrogenation of a-C of ring branch (RxC 31)	
	H-abstraction of a-C of ring branch by H (RxC 33)	
	• Ring isomerization with ring opening and recombination (RxC 41)	
	Ring with ring/ring branch radical → five-member-ring intermediate	
	• More complete classes and species see Fig. 10	—
Aromatics		
	• H-dropping on aryl ring branch (RxC 9, 33)	Pyrolysis/aromatics in shale oil/GC-MS/FTIR/nonylbenzene CBS-QB3 calculations[65]
	H-abstraction of a-C of ring branch by C radical (RxC 9)	Pyrolysis/*n*-butylbenzene/sealed isobaric gold tubes/GC-MS/GC-FID/CBS-QB3 calculations[66]
	H-abstraction of a-C of ring branch by H (RxC 33)	
	• Side chain breaking (RxC 18)	
	β-branch bond scission of ring carbon (RxC 18)	
	• Aryl ring opening (RxC 23)	
	β-ring opening of branched carbon radical (RxC 23)	
	• Aromatic derivative formation (RxC 9, 12)	
	H-abstraction by aryl ring with branch radical → aromatics (RxC 9)	
	Aryl ring with branch radical + H → aromatics (RxC 12)	
	• Recombination of C radicals (RxC 13)	

Table 4 (*continued*)

Component category of RP-3 fuel	Major reaction classes and corresponding species by SRG-Reax	Reported literature and method
		Pyrolysis/*n*-butylbenzene/sealed isobaric gold tubes/GC-MS/GC-FID/CBS-QB3 calculations[66]
	Combination of aryl ring branch radical and C radical fragment	Pyrolysis/benzene with C_2–C_3/single pulse shock tube/GC-MS[67]
	• Ring isomerization with ring opening and recombination (RxC 41) Aryl ring with ring/ring branch radical → five-member-ring intermediate	Pyrolysis/benzene with C_2–C_3/single pulse shock tube/GC-MS[67]
	• More complete classes and species see Fig. 12	—

2 methyl branches, and $C_{10}H_{20}$(16) of chain olefin with a long branch. Their reaction classes and pyrolysis intermediates exhibit certain specificity, as shown in the skeleton reaction network of olefins in Fig. 11. The relatively stable bicyclic-olefin species of C_9H_8(14) experience initial C–C bond cleavage at the bridge C–C bond (RxC 20) and intra-ring H-shift (RxC 36). Three major classes of merged pathways followed in the subsequent pyrolysis, including a small portion undergoes ring-opening to form unsaturated *n*-paraffins (RxC 22), H-abstraction (RxC 33) of the methyl branch of the ring and linking with the ring atom to form a bicyclic structure (RxC 39), and H-shift (RxC 36) of the ring atom and reconnecting of ring atoms to reform a bridge bond (RxC 40) to achieve ring isomerization. It should be noted that the short-branched olefin, $C_{10}H_{20}$(54), mainly produces propadiene of C_3H_4(614) by undergoing β-scission (RxC 3, 16) and dehydrogenation (RxC 8 or RxC 9).

Although aromatics account for 27 mol% of the 45-component RP-3 fuel, it is difficult for aromatics to break down into very small molecular fragments in the pyrolysis process due to the stability of aromatic ring structures. The identified pyrolysis reaction pathways of aromatics components in RP-3 are lengthy. The skeleton reaction network of aromatics in Fig. 12 only displays the initial cleavage pathways for simplicity. The initial pyrolysis reactions of aromatics usually involve the H-abstraction of aryl ring-branch carbon (RxC 9, 33), and the β-C–C bond cleavage of the aryl ring branch (RxC 18). All initial pyrolysis reactions of aromatics lead to the formation of aryl intermediates with a branch radical. Among the C–H and C–C bond cracking of aromatic side chains, the identified easier cracking of β-C–C branch bonds (RxC 18) and α-C–H branch bonds (RxC 9, 33) over that of α-C–C branch bonds (RxC 17) are supported by the BDEs calculated using the CBS-QB3 method.[65,66]

The subsequent observed pyrolysis pathways of aryl radical intermediates are two superclass reactions. One involves H-abstraction (47.4%), including H-abstraction from other radical fragments (RxC 9) to form stable aromatic derivatives, or continued H-abstraction by another radical fragment (RxC 9, RxC 33) to form aryl intermediate species with more radicals. Another is radical consumption reactions (26.0%) where the H radical (RxC 12) or a species fragment directly links to an aryl ring or an aryl ring branch (RxC 13, RxC 43). These activated intermediate species may lead to aromaticity destruction through aryl ring opening (RxC 23) and aryl ring transformation (RxC 40, RxC 41) as shown in Fig. 12.

The identified skeleton reaction network of aromatic components in RP-3 pyrolysis not only confirms the stability of the aryl ring structure on one hand, but also unravels the mild activation of aryl rings existing in RP-3 pyrolysis through generating radicals on the aryl ring or aryl ring branch, or through H/species fragments linking to the aryl ring (pyrolysis reactions of RxC 40–41 in Fig. 12). The slow progressive aromaticity destruction may lead to the aryl ring opening gradually in RP-3 pyrolysis.[65,67]

4. Conclusions

This work proposes an alternative approach of SRG-Reax for skeleton reaction network generation of the complex fuel pyrolysis reactions of ReaxFF MD simulations based on reaction classification of reaction centers using a machine learning method. SRG-Reax is achieved through building a semi-supervised machine learning model of tri-training with three different inputs of reaction center descriptions for predicting the reaction class of each of the pyrolysis reactions, combined with performing pathway reduction for a specified subnetwork of reactant species set based on reaction class ratios and product species ratios.

The SRG-Reax approach was applied in obtaining the skeleton reaction networks of 45-component real RP-3 fuel pyrolysis for its component categories of *n*-paraffins, iso-paraffins, cycloparaffins, olefins, and aromatics in this work. Some interesting findings were obtained from the skeleton pyrolysis reaction networks generated in this work. There are two major classes of merged reaction pathways found in the skeleton pyrolysis reaction network of *n*-paraffins. In addition to the dominant continuous terminal β-scission through RxC 3 to produce ethylene, there is a possibility to undergo an intra-molecular H-shift (RxC 6, accounting for 6–15%) before β-scission, resulting in the production of propylene. The characteristic reaction class of branched paraffin pyrolysis is intra-molecular chain isomerization (RxC 5) involving chain cleavage and recombination, and the products are alkenes especially propylene from further pyrolysis of the intermediates of 2-CH radical alkane, 2-methylene radical and x-CH radical alkane (non-1,2-alkane radical). The initial pyrolysis reaction pathways of cycloparaffins in RP-3 can be classified into three major characteristic superclasses of direct ring opening (RxC 25–27), side chain breaking (RxC 1, 17, 18), and

H-dropping (RxC 10, 29, 31, 33, 34). A pathway of ring isomerization was observed in the pyrolysis network of bicyclic olefin through initial bridge C–C bond cleavage, H-shift of a ring atom, reconnecting of ring atoms to reform a new bridge bond for gradual ring isomerization (RxC 20 → RxC 36 → RxC 40). The skeleton reaction network of aromatics in RP-3 pyrolysis indicates that the aryl ring structures are stable in pyrolysis before the H-abstraction of aryl ring-branch carbon, and β-C–C bond cleavage of the aromatic ring branch take place. However, the slow progressive aromaticity destruction may exist that leads to the aryl ring opening gradually in RP-3 pyrolysis.

The reaction network obtained using the SRG-Reax tool is basically a reduced network of reaction pathways lumped in terms of reaction class and their contribution to a set of targeted pyrolysis products. SRG-Reax helps in capturing a global network scenario for complex chemistry of realistic RP-3 fuel pyrolysis. SRG-Reax should be potentially applicable for the more complex fuel pyrolysis of coal, biomass, polymers, and more.

Data and software availability

Representative reaction data with 46 typical reaction classes manually labeled have been uploaded as part of the ESI,† for a better understanding of the labeled reaction data set supporting this article. The information supplied includes their source model information, reaction descriptions in atom-mapped SMARTS strings, atom IDs of reaction sites, the number of broken and formed bonds, extended reaction center structures in atom-mapped SMARTS strings, and their manually labeled reaction classes. Besides, 300 sample reaction data in the training set were uploaded as an item of the ESI,† to help generate a sense of the training data in the format of a tuple containing three arrays of Input 1, Input 2, and Input 3 for the tri-training classifier.

Author contributions

Shanwen Yang: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, writing – review & editing. Xiaoxia Li: conceptualization, methodolgy, funding acquisition, project administration, resources, supervision, writing – review & editing. Mo Zheng: conceptualization, supervision. Chunxing Ren: conceptualization, supervision. Li Guo: methodology, conceptualization, supervision.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by a grant from the National Natural Science Foundation of China (22173106). The authors thank Prof. Lin Ji for the mechanism reduction method discussion, Dr Song Han for the helpful discussion of hydrocarbon fuel and machine learning methods, and Yujie Tang for helping develop a utility code to output reaction data from VARxMD.

References

1 P. E. Savage, *J. Anal. Appl. Pyrolysis*, 2000, **54**, 109–126.
2 X. X. Li, M. Zheng, C. X. Ren and L. Guo, *Energy Fuels*, 2021, **35**, 11707–11739.
3 V. Burkle-Vitzthum, R. Bounaceur, R. Michels, G. Scacchi and P. M. Marquaire, *J. Anal. Appl. Pyrolysis*, 2017, **125**, 40–49.
4 A. Shiroudi, K. Hirao, K. Yoshizawa, M. Altarawneh, M. A. Abdel-Rahman, A. B. El-Meligy and A. M. El-Nahas, *Fuel*, 2020, **281**, 118798.
5 J. Z. Zeng, L. F. Zhang, H. Wang and T. Zhu, *Energy Fuels*, 2021, **35**, 762–769.
6 A. C. T. van Duin, S. Dasgupta, F. Lorant and W. A. Goddard, *J. Phys. Chem. A*, 2001, **105**, 9396–9409.
7 T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. X. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama and A. C. T. van Duin, *npj Comput. Mater.*, 2016, **2**, 15011.
8 M. Chen, W. Li, H. Zhang, M. Liu, J. Zhang, X. Li and Y. Han, *Energy Adv.*, 2023, **2**, 54–72.
9 Q. Mao, M. Y. Feng, X. Z. Jiang, Y. H. Ren, K. H. Luo and A. C. T. van Duin, *Prog. Energy Combust. Sci.*, 2023, **97**, 101084.
10 M. Zheng, X. X. Li and L. Guo, *J. Mol. Graphics*, 2013, **41**, 1–11.
11 S. B. Kylasa, H. M. Aktulga and A. Y. Grama, *J. Comput. Phys.*, 2014, **272**, 343–359.
12 S. Han, X. X. Li, M. Zheng and L. Guo, *Fuel*, 2018, **222**, 753–765.
13 P. Zhao, S. Han, X. X. Li, T. Zhu, X. F. Tao and L. Guo, *Energy Fuels*, 2019, **33**, 7176–7187.
14 H. Liu, J. H. Liang, R. N. He, X. X. Li, M. Zheng, C. X. Ren, G. J. An, X. M. Xu and Z. Zheng, *Combust. Flame*, 2022, **237**, 111865.
15 J. Liu, X. X. Li, L. Guo, M. Zheng, J. Y. Han, X. L. Yuan, F. G. Nie and X. L. Liu, *J. Mol. Graphics*, 2014, **53**, 13–22.
16 T. B. Y. Chen, A. C. Y. Yuen, B. Lin, L. Liu, A. L. P. Lo, Q. N. Chan, J. Zhang, S. C. P. Cheung and G. H. Yeoh, *J. Anal. Appl. Pyrolysis*, 2021, **153**, 104931.
17 L. Krep, I. S. Roy, W. Kopp, F. Schmalz, C. Huang and K. Leonhard, *J. Chem. Inf. Model.*, 2022, **62**, 1–13.
18 A. L. De Bortoli and F. N. Pereira, *J. Math. Chem.*, 2019, **57**, 812–833.
19 Y. Z. Wu, H. Sun, L. Wu and J. D. Deetz, *J. Comput. Chem.*, 2019, **40**, 1586–1592.
20 H. Wang and M. Frenklach, *Combust. Flame*, 1991, **87**, 365–370.
21 L. Heberle, P. Sharma and P. Pepiot, *Combust. Flame*, 2021, **234**, 111682.
22 T. F. Lu and C. K. Law, *Proc. Combust. Inst.*, 2005, **30**, 1333–1341.
23 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.

24 E. S. Blurock, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 505–510.
25 E. S. Blurock, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 607–616.
26 M. Meuwly, *Chem. Rev.*, 2021, **121**, 10218–10239.
27 S. Stocker, G. Csanyi, K. Reuter and J. T. Margraf, *Nat. Commun.*, 2020, **11**, 5505.
28 Z. H. Zhou and M. Li, *Knowl. Inf. Syst.*, 2010, **24**, 415–439.
29 Z. H. Zhou and M. Li, *IEEE Trans. Knowl. Data Eng.*, 2005, **17**, 1529–1541.
30 D. D. Lewis and W. A. Gale, presented in part at the SIGIR '94, 1994.
31 J. Q. Xu, J. J. Guo, A. K. Liu, J. L. Wang, N. X. Tan and X. Y. Li, *Acta Phys.-Chim. Sin.*, 2015, **31**, 643–652.
32 D. Zheng, W. M. Yu and B. J. Zhong, *Acta Phys.-Chim. Sin.*, 2015, **31**, 636–642.
33 H. W. Deng, C. B. Zhang, G. Q. Xu, Z. Tao, B. Zhang and G. Z. Liu, *J. Chem. Eng. Data*, 2011, **56**, 2980–2986.
34 T. J. Bruno and B. L. Smith, *Ind. Eng. Chem. Res.*, 2006, **45**, 4381–4388.
35 A. Blum and T. Mitchell, *Combining labeled and unlabeled data with co-training*, in Proceedings of the 11th annual conference on Computational learning theory, Association for Computing Machinery, New York, NY, USA, 1998, 92-100.
36 S. A. Goldman and Y. Zhou, *Enhancing Supervised Learning with Unlabeled Data*, in Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, 2000, 327-334.
37 D. Angluin and P. Laird, *Mach. Learn.*, 1988, **2**, 343–370.
38 S. Dasgupta, M. L. Littman and D. McAllester, *PAC generalization bounds for co-training*, in Proceedings of the 14th International Conference on Neural Information Processing Systems, Natural and Synthetic, MIT Press, Cambridge, MA, USA, 2001, 375–382.
39 Demonstrates an active learning technique to learn handwritten digits using label propagation, **https://scikit-learn.org/stable/auto_examples/semi_supervised/plot_label_propagation_digits_active_learning.html**, (accessed January, 2024).
40 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
41 A random forest classifier, **https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier**, (accessed January, 2024).
42 D. Probst, P. Schwaller and J.-L. Reymond, *Digital Discovery*, 2022, **1**, 91–97.
43 D. T. Ahneman, J. G. Estrada, S. S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
44 N. Schneider, D. M. Lowe, R. A. Sayle and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 39–53.
45 F. Sandfort, F. Strieth-Kalthoff, M. Kuhnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379–1390.
46 P. Schwaller, A. C. Vaucher, T. Laino and J. L. Reymond, *Mach. Learn.-Sci. Technol.*, 2021, **2**, 015016.
47 Daylight Chemical Information Systems. SMARTS - A Language for Describing Molecular Patterns., **https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html**, (accessed January, 2024).
48 RDKit: Open-Source Cheminformatics Software, **https://www.rdkit.org/**, (accessed January 2024).
49 C. Strobl, A. L. Boulesteix, A. Zeileis and T. Hothorn, *BMC Bioinf.*, 2007, **8**, 25.
50 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
51 R. E. Carhart, D. H. Smith and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64–73.
52 R. Nilakantan, N. Bauman, J. S. Dixon and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1987, **27**, 82–85.
53 Y. Wang, S. Cang and H. N. Yu, *Exp. Syst. Appl.*, 2019, **137**, 167–190.
54 Gephi: The Open Graph *Viz.* Platform, **https://gephi.org/**, (accessed January, 2024).
55 H. J. Curran, P. Gaffuri, W. J. Pitz and C. K. Westbrook, *Combust. Flame*, 1998, **114**, 149–177.
56 J. P. Chakraborty and D. Kunzru, *J. Anal. Appl. Pyrolysis*, 2009, **86**, 44–52.
57 D. S. Aribike and A. A. Susu, *Thermochim. Acta*, 1988, **127**, 247–258.
58 D. S. Aribike and A. A. Susu, *Thermochim. Acta*, 1988, **127**, 259–273.
59 K. D. Dahm, P. S. Virk, R. Bounaceur, F. Battin-Leclerc, P. M. Marquaire, R. Fournet, E. Daniau and M. Bouchez, *J. Anal. Appl. Pyrolysis*, 2004, **71**, 865–881.
60 R. P. Jiang, G. Z. Liu, X. Y. He, C. H. Yang, L. Wang, X. W. Zhang and Z. T. Mi, *J. Anal. Appl. Pyrolysis*, 2011, **92**, 292–306.
61 F. L. Dryer and K. Brezinsky, *Combust. Sci. Technol.*, 1986, **45**, 199–212.
62 K. D. Dahm, PhD Dissertation/Thesis, Massachusetts Institute of Technology, 1998.
63 Z. D. Wang, L. Zhao, Y. Wang, H. T. Bian, L. D. Zhang, F. Zhang, Y. Y. Li, S. M. Sarathy and F. Qi, *Combust. Flame*, 2015, **162**, 2873–2892.
64 Y. T. Dai, W. Q. Zhao, H. J. Xie, Y. S. Guo and W. J. Fang, *J. Anal. Appl. Pyrolysis*, 2020, **145**, 104723.
65 Y. W. Wang, X. X. Han and X. M. Jiang, *Energy*, 2023, **279**, 127998.
66 N. C. L. Guerra, J. C. L. Huerta, C. Lorgeoux, R. Michels, R. Fournet, B. Sirjean, A. Randi, R. Bounaceur and V. Burkle-Vitzthum, *J. Anal. Appl. Pyrolysis*, 2018, **133**, 234–245.
67 A. Hamadi, W. Y. Sun, S. Abid, N. Chaumeix and A. Comandini, *Combust. Flame*, 2022, **237**, 111858.
68 R. Vinu and L. J. Broadbelt, in *Annual Review of Chemical and Biomolecular Engineering*, ed. J. M. Prausnitz, 2012, vol. 3, pp. 29–54.

