Chemical Engineering Science 284 (2024) 119461

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

A knowledge-driven approach for automatic generation of reaction networks of methanol-to-olefins process

Junyi Yu [a,b], Hua Li [a,*], Mao Ye [a,*], Zhongmin Liu [a,b]

[a] National Engineering Research Center of Lower-Carbon Catalysis Technology, Dalian National Laboratory for Clean Energy, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, 116023 Dalian, China
[b] University of Chinese Academy of Sciences, 100049 Beijing, China

ARTICLE INFO

Keywords:
Knowledge-driven
Automatic generation
Reaction networks
Reaction rules topological matrix representation (RTMR)
Methanol to olefin (MTO)

ABSTRACT

In recent years, automatic generation of reaction networks for chemical processes has attracted considerable attention. However, most of the automatic generators focused on the construction of reaction networks by computer and are essentially dependent upon the predefined reaction rules or categories. The reaction rules for constructing reaction networks are derived manually based on the individual's understanding of reaction mechanism equation as well as the corresponding experiences, which hinders the automatic generation of reaction network directly from the complex reaction mechanisms or equations. In this article, we proposed a strategy to automatically identifies the reaction rules directly from the reaction equation using a *Reaction Rules Topological Matrix Representation* (RTMR) method. Taking the induction period of methanol to olefins (MTO) process as an example, we showed how to use RTMR to extract the information of reaction mechanism equation in the literatures, and then establish the reaction networks of the multiphase catalytic process. The results suggest that RTMR not only provides an effective approach for automatic reaction mechanisms equation extraction, but also demonstrates new reactions prediction ability after training with the literature information. It is expected RTMR can be promisingly implemented in knowledge-driven reaction network generation in multiphase catalytic processes.

1. Introduction

Reaction kinetic models have been used as a common practice in chemical engineering for process development and reactor design. Despite the differential–algebraic equations of related kinetic parameters and thermodynamic data, reaction networks also constitute a key component in reaction kinetic models (Vernuccio and Broadbelt, 2019). Reaction network deals with the connections between reactant and product molecules, as well as various intermediates, via potential reactions based on the reaction mechanisms. Usually, for a complex heterogeneous catalytic reaction process, it is a non-trivial task even for an experienced chemical engineer to establish the reaction network as the hand-drawn reaction routines would not incorporate all the mechanisms. Rapid development of computational power in recent decades makes it possible to automatically generate the reaction network of a complex reaction process by the computer (Froment, 2013; Suleimanov and Green, 2015; Thybaut and Marin, 2013; Unsleber, 2023; Unsleber and Reiher, 2020).

The automatic reaction network generators have been continuously developed (Turtscher and Reiher, 2023), and typically systems such as NETGEN (Broadbelt et al., 1994), (De Witt et al., 2000), EXGAS (Buda et al., 2005), RING (Rangarajan et al., 2010), Genesys (Vandewiele et al., 2012), RNG (Karaba et al., 2013) (Zámostný et al., 2014), RMG (Gao et al., 2016), and CRN-ML (Wen et al., 2023), have been successfully applied to catalytic cracking, isomerization, combustion and other reaction systems. In addition, they are also used in heterogeneous catalytic reactions (Steiner and Reiher, 2022) and automated reaction monitoring of complex chemical systems (Puliyanda et al., 2022). Most of the automatic generators focused on the construction of reaction networks by computer and are essentially dependent upon the predefined reaction rules. The state of art of the reaction network construction concentrates on the structural expressions of species and reactions using computer processible symbols on the one hand, and automatic reaction network generation algorithm on the other hand. Both linear symbols and graphical algorithms have been widely used as the representation methods. Linear symbols provide an effective way to

* Corresponding authors.
E-mail addresses: lihua@dicp.ac.cn (H. Li), maoye@dicp.ac.cn (M. Ye).

https://doi.org/10.1016/j.ces.2023.119461
Received 4 April 2023; Received in revised form 27 September 2023; Accepted 30 October 2023
Available online 4 November 2023

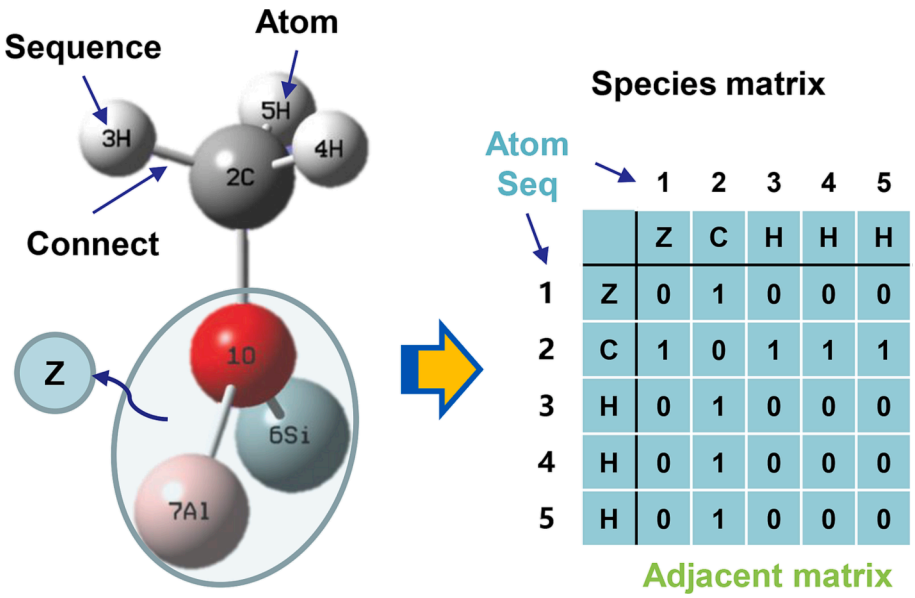


Fig. 1. A schematic representation of "Z-CH_3" by species matrix.

represent the structural formula of organic matters. For example, SMILES use simplified linear symbols to express atoms, charges, chemical bonds, branch chains and rings in chemical structures (Weininger, 1988; Weininger et al., 1989). SMART, an improved version of SMILES, allows for the use of wildcards to represent atoms and chemical bonds, and symbols to represent certain species (Vandewiele et al., 2012). SOL and InChI are also two frequently used linear symbols methods (Heller et al., 2015). Different from linear symbols, it is also possible to express the species using graph algorithm based on graph theory. The structure of species can be expressed by the adjacency matrix corresponding to the structural graph of species. Since this matrix is only composed of 0 and 1, it is also called Boolean adjacency matrix or Boolean matrix (Clymans and Froment, 1984; Hillewaert et al., 1988). After using computer processible symbols to represent all possible species, it is possible to automatically generate reaction network according to the predefined reaction rules.

The reaction rules predefined for constructing reaction networks are normally derived manually based on the individual's understanding of reaction mechanism and experiences, which is only suitable for specific reaction systems and hinders the automatic generation of reaction network directly from the complex reaction mechanisms. For unknown reaction systems, the reaction network constructed via the aforementioned way might be incomplete as hardly all possible reaction rules can be implemented solely based on the researcher's personal judgment. Automatic extraction of reaction rules is necessary. For instance, Gupta and Vlachos developed an automated first-principles reaction rule generation framework which learns the reaction rules from DFT data by generating molecular graphs of common subgraph from reactants to products and defining all bonds that could change during a reaction (Gupta and Vlachos, 2021). Despite that the reaction mechanism study is time-consuming and laborious for a given reaction process, in real practice the reaction process indeed contains many elementary reaction steps that have been previously discovered elsewhere. In this sense, the reaction rules based on reaction mechanism for a given reaction process can be mined from available knowledge based on the available knowledge in published literature. However, automatic generation of reaction rules based on the available knowledge has seldom been studied in the automatic generation of reaction networks (Clymans and Froment, 1984; Gupta and Vlachos, 2021; Heller et al., 2015; Hillewaert et al., 1988; Weininger, 1988; Weininger et al., 1989).

In this study, we proposed a knowledge-driven computer-assisted automatic generation of heterogeneous catalytic reaction network approach for the zeolite-catalyzed methanol to olefins process. In particular, the reaction mechanism information available in the literature (or from researcher's experiences) such as elementary reactions and transition state changes was collected and expressed as species matrix, namely *Reaction Rules Topological Matrix Representation* (RTMR), to represent the reaction rules. Thus, a reaction network can be generated according to the extracted reaction rules via RTMR. It will be shown that RTMR is an efficient carrier of reaction rules, which can be extracted from reaction mechanism knowledge from the open literature rather than solely relying on the past experiences of researchers. It is expected RTMR can be promisingly implemented in knowledge-driven reaction network generation in multiphase catalytic processes.

2. Species matrix and operations definition

2.1. Species matrix

In this work, all species involved are represented by a species matrix, which includes a sequence of elements and an adjacency matrix (in Fig. 1). The process of using species matrix to represent a specified species consists of numbering the atoms, groups or clusters, obtaining the element vector sequence, and establishing the adjacency matrix of the species. The sequence of elements is a list of symbols that correspond to atoms, groups, or clusters in the molecular structure of the species (for example, the symbol "Z" is used to describe molecular sieve clusters). The adjacency matrix represents their connection information ("1" for single bonds, "2" for double bonds, and "0" for unconnected bonds). Considering the isomorphism of molecular graph, species are standardized to ensure the one-to-one correspondence between species and adjacency matrix (Golender et al., 1981; Walters and Yalkowsky, 1996).

2.2. Operation definition of species matrix

2.2.1. Contragradient transformation

Note that the representation of a specified species with species matrix might lead to several different species matrices if, for example, the numbering of atoms starts with different atoms, in this work we also define contragradient transformation operation. Thus, after contragradient transformation the species matrices, if representing the same molecule, can eventually become a unified matrix.

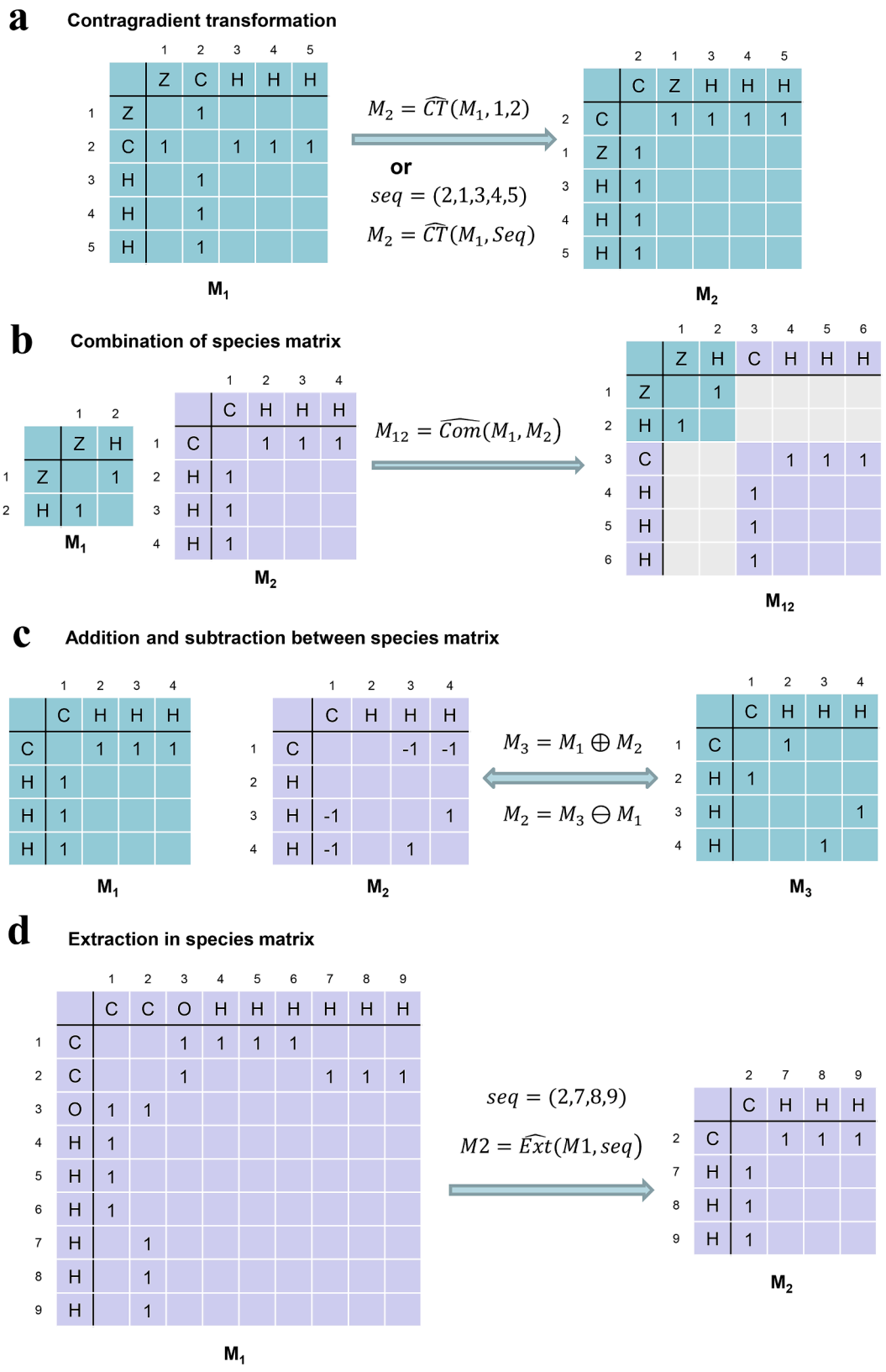


Fig. 2. Definitions of the operators of species matrices and corresponding transformation.

In contragradient transformation, we exchange the corresponding rows or columns of the species matrix according to the predefined atom number sequences. The contragradient transformation can be written as

$$M_2 = \widehat{CT}(M_1, a, b) \tag{1}$$

or

$$M_2 = \widehat{CT}(M_1, seq) \tag{2}$$

where $\widehat{CT}$ refers to the contragradient transformation operator, M_1 is

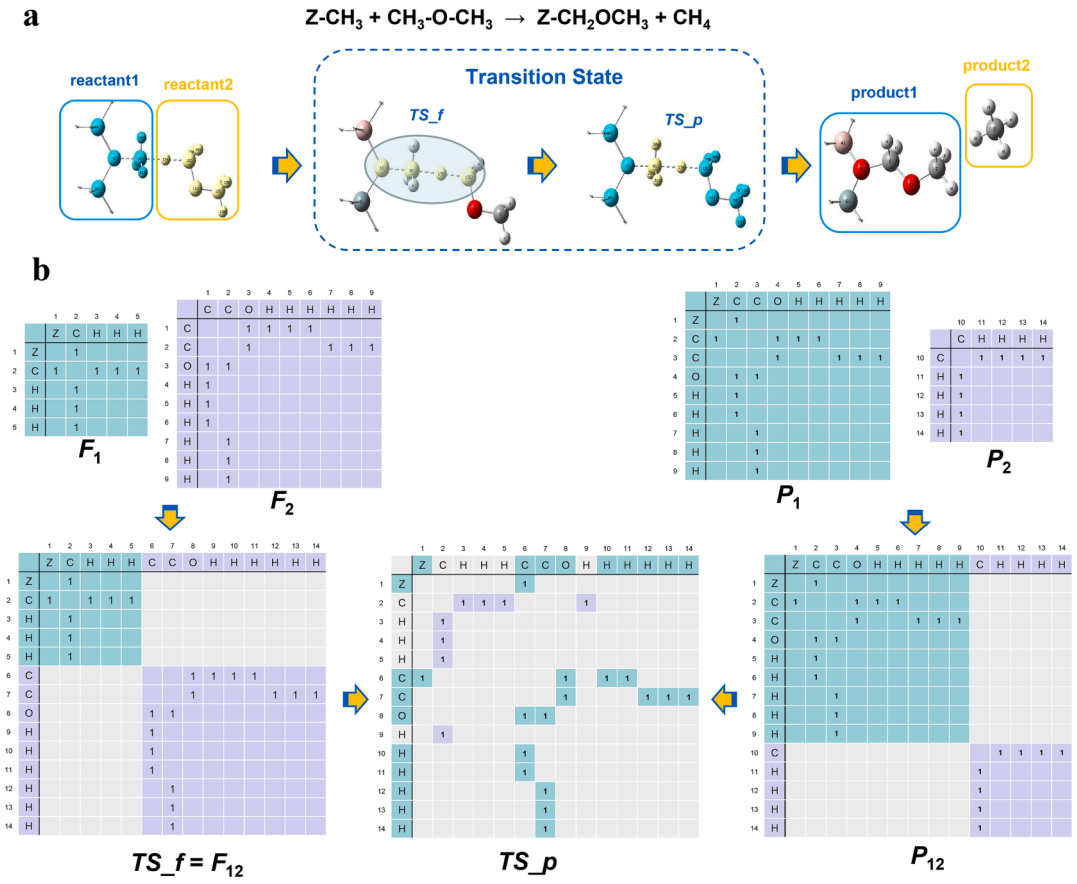


Fig. 3. The corresponding transformation of species matrices for an example of bimolecular reaction "Z-CH_3 + CH_3OCH_3 → Z-CH_2OCH_3 + CH_4". a: reaction formula; b: species matrices transformation of this reaction (omitting "0" in the matrices).

the original species matrix and M_2 is the transformed species matrix. In the Equation (1) two variables a and b are used, which denotes that in the contragradient transformation the row and column corresponding to the atom a will exchange with that to atom b. Equation (2) is another formula of Equation (1), in which a vector seq representing the elements sequence of M_2 is instead used. In Fig. 2**a** we showed an example for the contragradient transformation with atom 1 and 2 being exchanged, which is the same transformation that sets seq to (2,1,3,4,5).

2.2.2. Combination of species matrices

We define the combination operation for two species matrices, which refers to the merging of two species matrices into one matrix, as shown in the Fig. 2**b**. The combination of two species matrices can be written as

$$M_{12} = \widehat{Com}(M_1, M_2) \tag{3}$$

where M_1 and M_2 are two species matrices, and M_{12} is the species matrix after combination.

2.2.3. Addition and subtraction of species matrices

For two species matrices with same matrix dimension and same atomic sequence, we define addition or subtraction operation, as shown in the Fig. 2**c**. The addition and subtraction of two species matrices can written as

$$M_3 = M_1 \oplus M_2 \tag{4}$$

$$M_2 = M_3 \ominus M_1 \tag{5}$$

where M_1, M_2 and M_3 are three species matrices with same matrix dimension and same atomic sequence.

2.2.4. Topological distance in species matrix

Here we define the topological distance between two atoms in a molecule as the number of atoms of shortest distance in between these two atoms. The definition of topological distance can be written as

$$m_{a,b} = \widehat{Td}(M_1, a, b) \tag{6}$$

where $m_{a,b}$ is the topological distance between atoms a and b in the matrix M_1. (see more details in the **Supplemental material**).

2.2.5. Extraction of species matrix

In this work, we define an extraction operation to establish a new species matrix based on the specified sequence in an existing species matrix, as shown in the Fig. 2**d**. It can be expressed as

$$M_2 = \widehat{Ext}(M_1, seq) \tag{7}$$

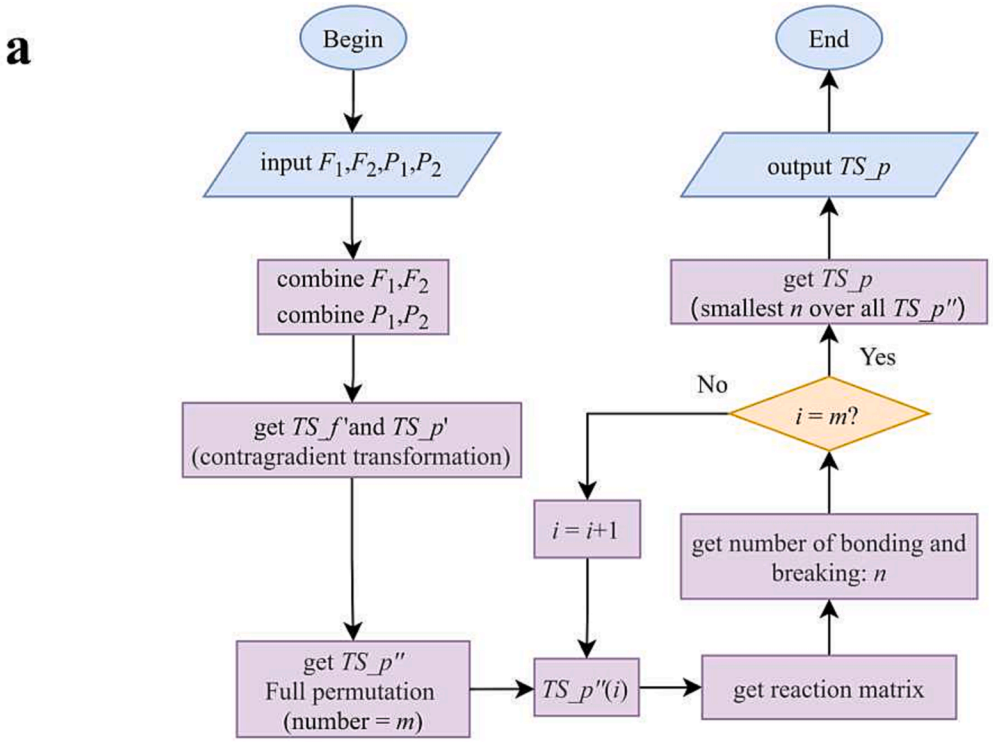
Here M_1 is the initial species matrices, *seq* is the specified sequence of atoms in M_1 and M_2 is the species matrix extracted from M_1 according to *seq*.

3. RTMR model

The automatic generation of reaction network includes three steps: using species matrix to describe species and reactions, representing reaction rules through RTMR, and generating reaction network based on reaction rules.

3.1. Representation of reactions by species matrix

For the elementary reaction process, it can be classified into



b

Seq(TS_f)

Atom	Seq
Z	1
C	2
H	3
H	4
H	5
C	6
C	7
O	8
H	9
H	10
H	11
H	12
H	13
H	14

Seq(TS_f')

Atom	Seq
Z	1
C	2
C	6
C	7
O	8
H	3
H	4
H	5
H	9
H	10
H	11
H	12
H	13
H	14

Seq(TS_p'')

Atom	Seq_1	Seq_2	Seq_3	...
Z	1	1	1	...
C	2	2	3	...
C	3	10	2	...
C	10	3	10	...
O	4	4	4	...
H	5	5	5	...
H	6	6	6	...
H	7	7	7	...
H	8	8	8	...
H	9	9	9	...
H	11	11	11	...
H	12	12	12	...
H	13	13	13	...
H	14	14	14	...

Seq(TS_p')

Atom	Seq
Z	1
C	2
C	3
C	10
O	4
H	5
H	6
H	7
H	8
H	9
H	11
H	12
H	13
H	14

Seq(P_{12})

Atom	Seq
Z	1
C	2
C	3
O	4
H	5
H	6
H	7
H	8
H	9
C	10
H	11
H	12
H	13
H	14

Fig. 4. Searching algorithm for..*TS_p*

monomolecular, bimolecular and multi-molecular reactions according to the number of molecules enclosed in the reacting species. In this work, only monomolecular and bimolecular reactions are considered: $F_1 \rightarrow P_1$; $F_1 \rightarrow P_1 + P_2$; $F_1 + F_2 \rightarrow P_1 + P_2$; $F_1 + F_2 \rightarrow P_1$ (here F_1 and F_2 are reactants, and P_1 and P_2 the products).

Taking the reaction formula "Z-CH_3 + CH_3OCH_3 → Z-CH_2OCH_3 + CH_4" shown in Fig. 3**a** as an example, we can find in this bimolecular reaction there are two reactants F_1 and F_2, which are respectively Z-CH_3 (with Z representing the active site of the molecular sieve) and CH_3-O-CH_3, and the two product molecules formed during reaction P_1 and P_2, which are respectively Z-CH_2OCH_3 and CH_4. As shown in Fig. 3**b**, the matrices of the reactants F_1 and F_2 are first combined via the operator $F_{12} = \widehat{Com}(F_1, F_2)$. In the same way, matrices of the products P_1 and P_2 can also be combined via $P_{12} = \widehat{Com}(P_1, P_2)$. The transition state that

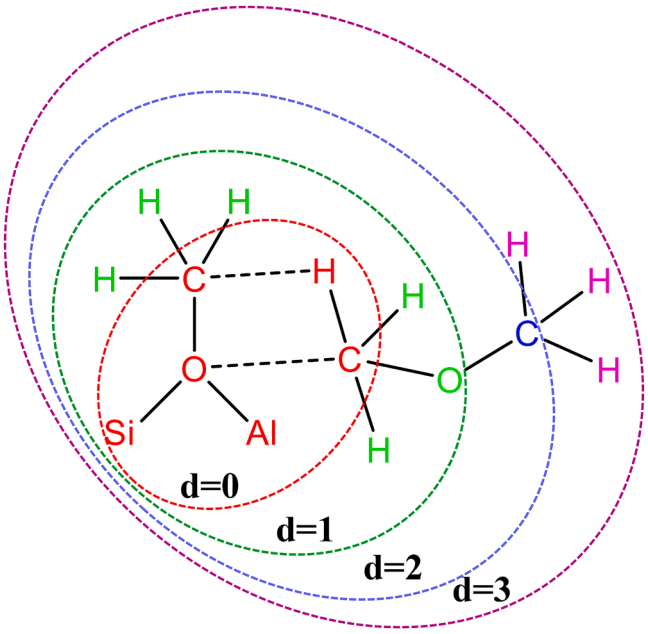


Fig. 5. The definition of RTMR distance.

defines the reaction is described by the change of the matrix TS_f to TS_p, where $TS\_f = F_{12}$. The matrix TS_p is derived through a search algorithm (This algorithm will be introduced later in this section).

Some methods have been proposed in the literature for reaction mapping, for example, maximum common subgraph between reactants and products (Raymond et al., 2002), minimum reaction energy by weighted maximum common edge subgraph energy (Apostolakis et al., 2008; Korner and Apostolakis, 2008), and integer linear optimization for identifying a mapping with the fewest number of bond changes (First et al., 2012). In this work we use a relatively simple semi-exhaustive algorithm to get TS_p, inspired by the maximum common subgraph method (Raymond et al., 2002). As shown in the Fig. 4, The algorithm for obtaining TS_p includes the following steps:

(1) Based on the species matrices of reactants F_1 and F_2 (as shown in Fig. 3), we can obtain the combination $F_{12} = \widehat{Com}(F_1, F_2)$, and get $TS_f = F_{12}$. Then we get the sequence of TS_f, which can be denoted as $Seq(TS\_f)$. According to the order of Z-C-O-H, we rearranged the sequence of TS_f and obtained a new sequence which is denoted as $Seq(TS\_f')$ shown in Fig. 4**b**). Then, we further derived the new species matrix TS_f' via $TS\_f' = \widehat{CT}(TS\_f, Seq(TS\_f'))$.

(2) Follow the same method, we rearranged the sequence of P_{12} according the to the order of Z-C-O-H. And obtained a new sequence which is denoted as $Seq(TS\_p')$ shown in Fig. 4**b**). Then, we further derived the new species matrix TS_p' via $TS\_p' = \widehat{CT}(P_{12}, Seq(TS\_p'))$.

(3) Based on the species matrix TS_p', we performed a full arrangement of the sequences of each element, that is, for each element there might be different sub-arrangement of atoms of this element. In this case, the number of sequence matrices $Seq(TS\_p'')$ is m (for example, $m = A_1^1 A_3^3 A_1^1 A_9^9$ in Fig. 4**b**), and we can obtain each potential species matrix TS_p'' via $TS\_p'' = \widehat{CT}(TS\_p', Seq(TS\_p''))$.

(4) For each TS_p'', obtain the reaction matrix through $TS\_p'' \ominus TS\_f'$, and count the number of bond formation and bond breaking, recorded as n.

(5) Find the smallest n over all TS_p'', and take the TS_p'' corresponding to the smallest n as TS_p.

3.2. The establishment of RTMR model

The extraction and operation of RTMR is critical for reaction network generation. The reaction described in Fig. 5 is "Z-CH_3 + CH_3OCH_3 → Z-CH_2OCH_3 + CH_4". The set of atoms (or a group of atoms "Z") involved in the bonding or breaking reaction in the reactant molecule is defined as the reaction center. The definition of topological distance: $m_{a,b} = \widehat{Td}(M1, a, b)$. RTMR distance refers to the minimum topological distance to any atom in the reaction center. The RTMR range is determined after selecting the RTMR distance, which is the sum of all atoms less than or equal to the RTMR distance (including the reaction center, the RTMR distance of the reaction center is 0). When $d = 1$, which means RTMR distance = 1, the range of the RTMR contains adjacent atoms of the reaction center (in the dotted green line in the figure); similarly, when $d = 2$, the RTMR contains the blue region, and when $d = 3$, it contains the purple region.

As an example of the bimolecular reaction "Z-CH_3 + CH_3OCH_3 → Z-CH_2OCH_3 + CH_4" in the previous section, Fig. 6**a** shows the TS_p matrix, in which the transformation from TS_f to TS_p is marked with special colors (bond formation, bond breakage, reaction center). Fig. 6**b** is the structure of RTMR, which includes three parts: structure1, structure2, and Reaction-Mat. The two topological matrices on the left represent the substructures of the reactants, including the reaction center and the substructure scope. The Reaction-Mat on the right is reaction matrix which describing the reaction process, the bond formation and bond breaking process of the transition state. The RTMR structure to Fig. 6**b** is extracted based on the TS_p corresponding to Fig. 6**a**. The extraction process is: (1) Select the RTMR distance. In the example in the figure, the RTMR distance is selected as 1; (2) Find the substructure range, which means $RTMRrange$, $RTMRrange_{F1} = (1, 2, 3, 4, 5)$, $RTMRrange_{F2} = (6, 8,$

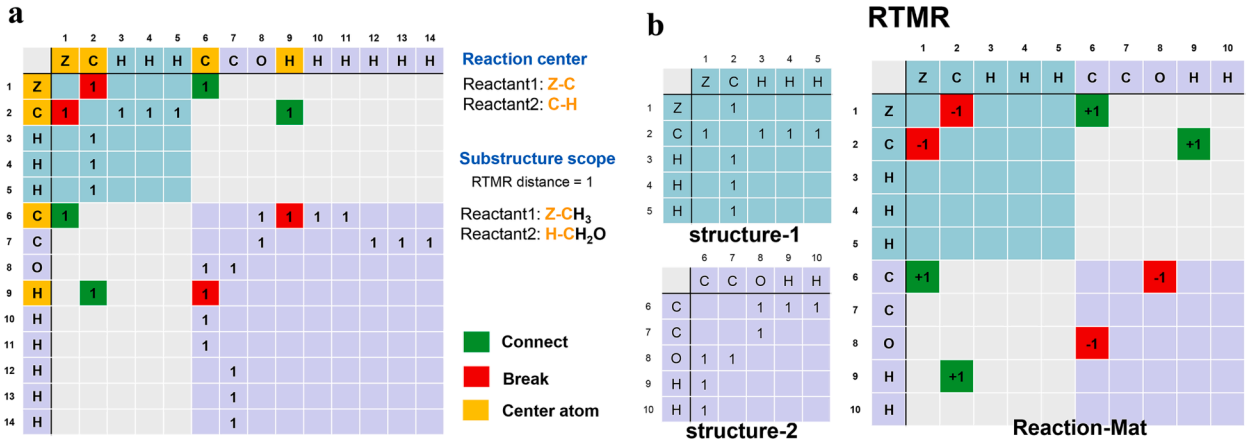


Fig. 6. Reaction rules base on RTMR (a: details of the bond breaking and connection process in the reaction; b: the components of RTMR).

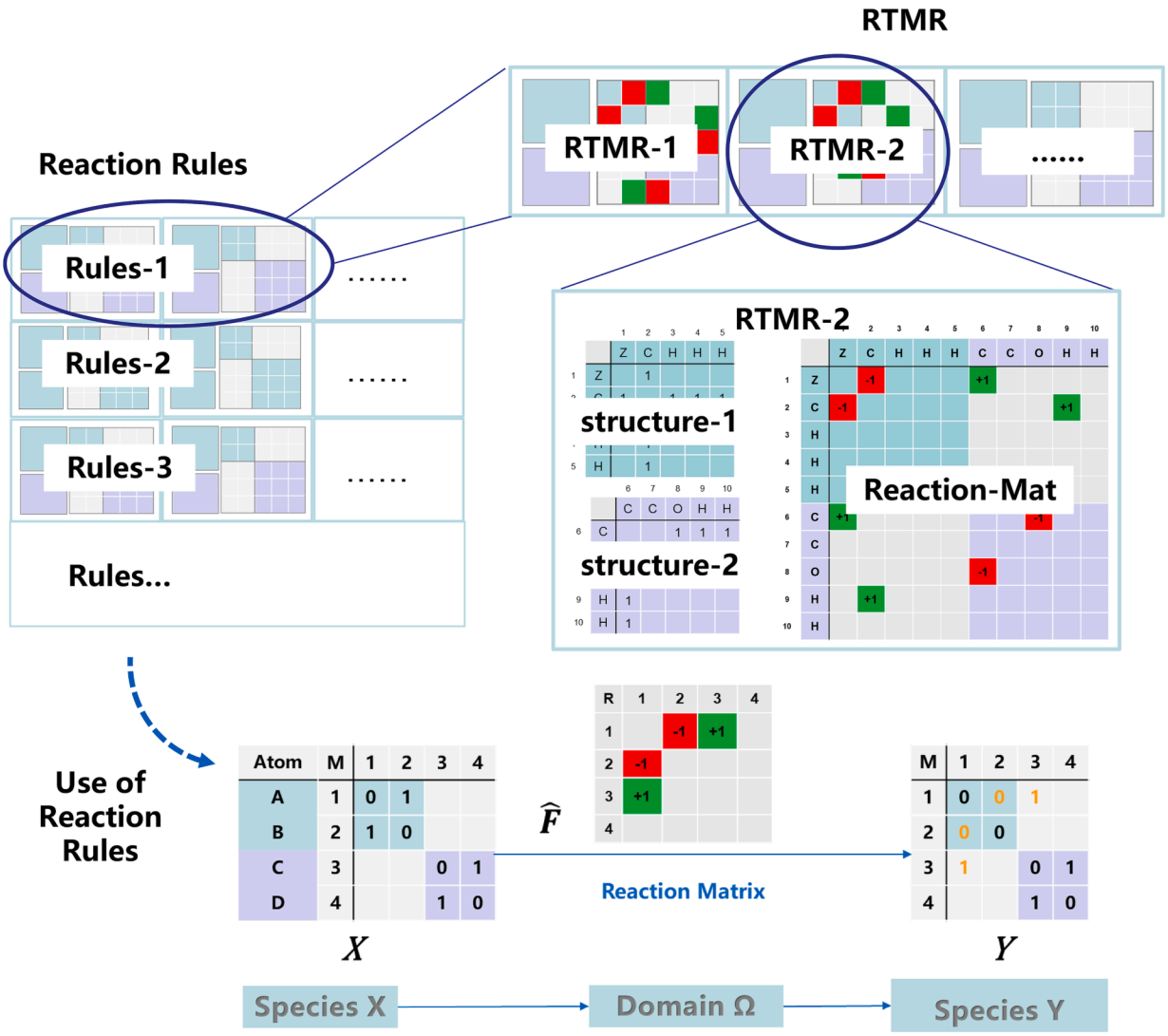


Fig. 7. Extraction and use of reaction rules.

9, 10, 11) , *RTMRrange* = (1, 2, 3, 4, 5, 6, 8, 9, 10, 11); (3) Extract the substructure in *TS_p* (the substructure is the intersection of the reactant and the *RTMRrange*, $structure1 = \widehat{Ext}(F_1, RTMRrange_{F1})$, $structure2 = \widehat{Ext}(F_2, RTMRrange_{F2})$); (4) Extract the *Reaction_Mat* by $Reaction\_Mat = \widehat{Ext}(TS\_p \ominus TS\_f, RTMRrange)$.

In particular, the order of the substructures in RTMR does not affect the reaction rules. The RTMR in Fig. 6**b** corresponds to a bimolecular reaction, and the RTMR corresponding to a single-molecule reaction process will only have two parts, i.e., structure-1 and Reaction-Mat. RTMRs with the same reaction center are defined as the same reaction rules, so when the extracted reactions have the same reaction center, they are all stored in the different RTMR but the same reaction rule (see **Figure S3** in **Supplemental material**). As can be seen from Fig. 7, the reaction rule may contain multiple RTMRs, and each RTMR contains complete substructures and Reaction-Mat.

When the reaction rules established by RTMR are used, it is necessary to determine whether the reaction species contains the topology structure of corresponding structure 1 and structure 2 to the RTMR of the reaction rule (see **Figure S4** in **Supplemental material**). In particular, bimolecular reactions are not affected by the order of each other, because the matrix of reactants (F_1 and F_2) can exchange with each other and the outcome of the reaction would not be influenced. If so, according to the corresponding relationship of the topological structure, we can obtain *TS_p* by Reaction-Mat, and then obtain the species matrix of the product (P_1 and P_2) through contragradient transformation and matrix separation; if not, the reaction species cannot react accordingly. (more details in **Supplemental material**).

$$\widehat{F}(X) = \begin{cases} Y, X \in \Omega \\ \varnothing, X \notin \Omega \end{cases} \tag{8}$$

$$Y = X \oplus R \tag{9}$$

The use of RTMR in describing the reaction rules can be abstracted as mathematical functions. As shown in the Fig. 7, for a well-established reaction rule, species X (the species matrix of species reactant1 and species reactant2 is combined, if it is a single molecule reaction, then X = reactant1, no reactant2) as an independent variable. In Eq. (8) and (9) The reaction rule is denoted as $Y = \widehat{F}(X)$, and the domain of $\widehat{F}(X)$ is denoted as Ω, which represents substructure and other information of the RTMR (if the independent variable species X contains the corresponding block, it belongs to the domain). The independent variable X that belongs to the domain can have an output, otherwise, it outputs ∅. R is the reaction matrix of the RTMR, and Y is the product, which is obtained by adding the independent variable X to the reaction matrix of the RTMR. The whole process in the picture is the realization form of a reaction rule.

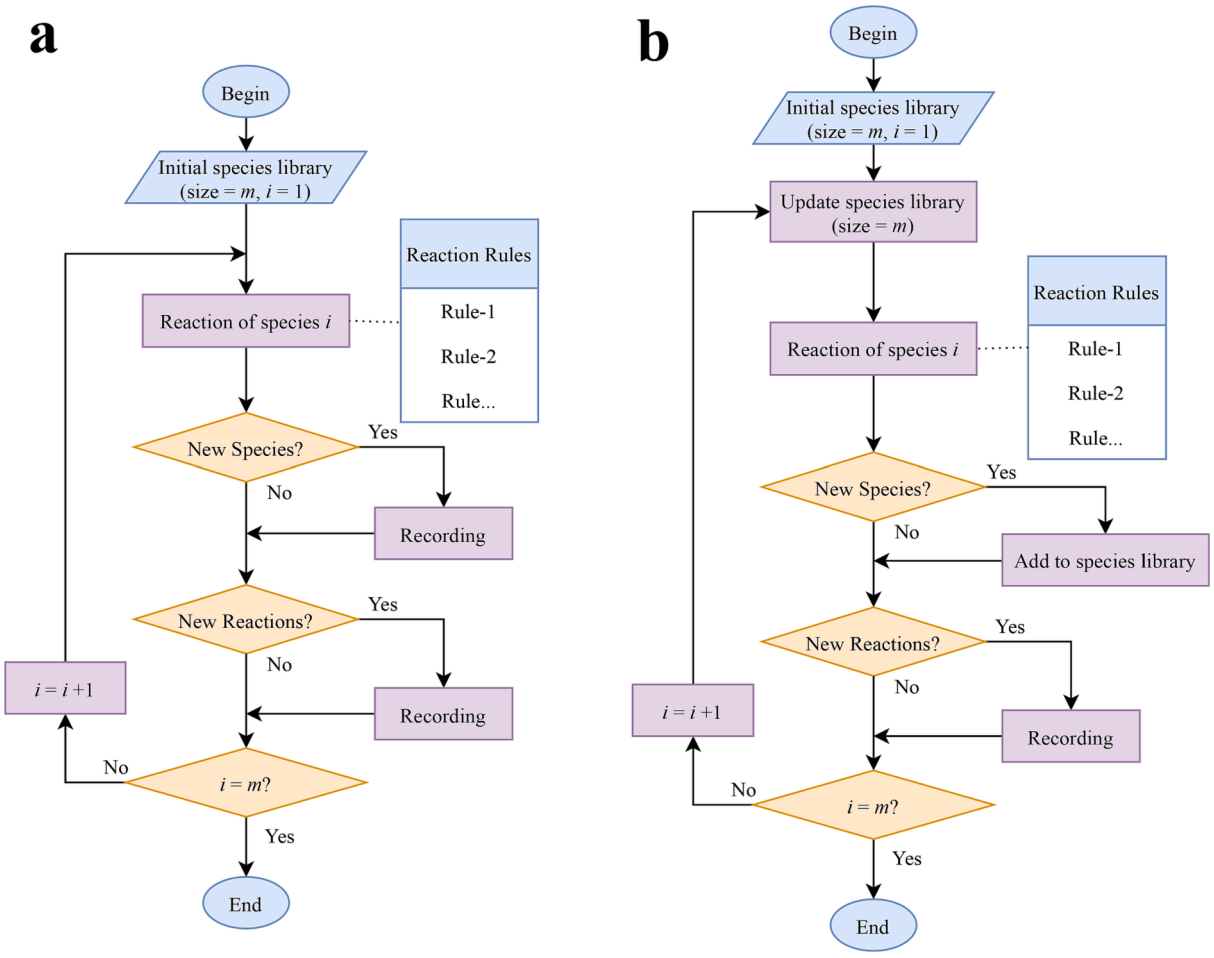


Fig. 8. Procedure for reaction network construction (a: Once-Reaction-Network; b: Cyclic-Reaction-Network).

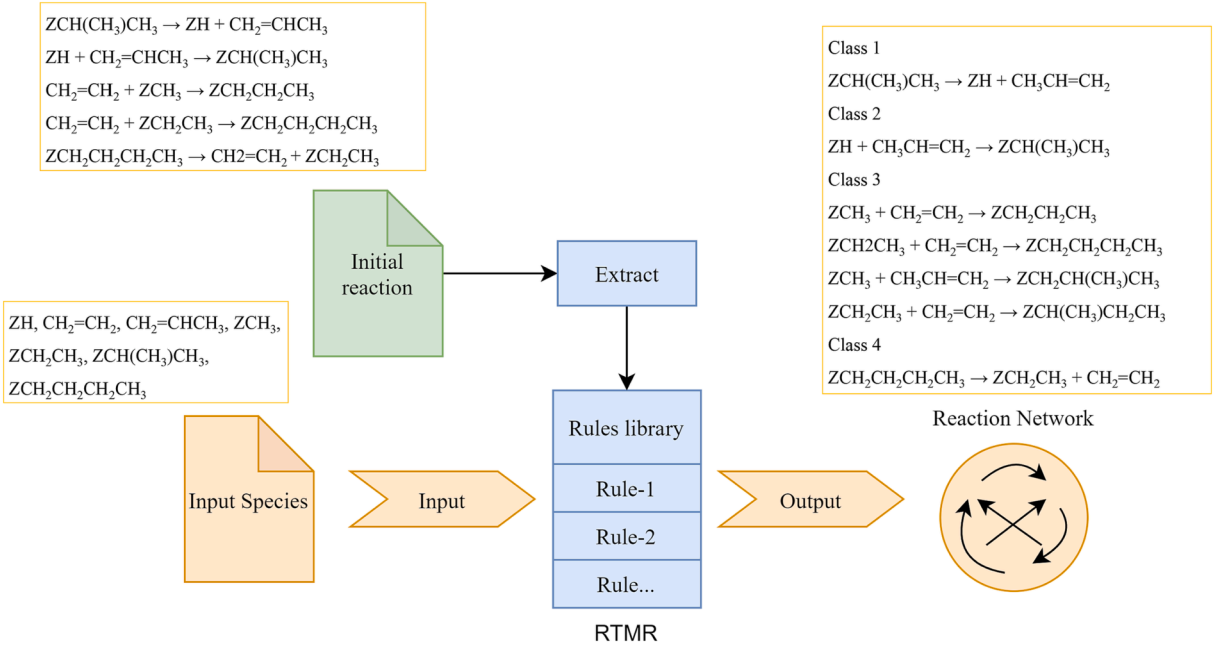


Fig. 9. Verify of extraction and generation of existing reaction networks.

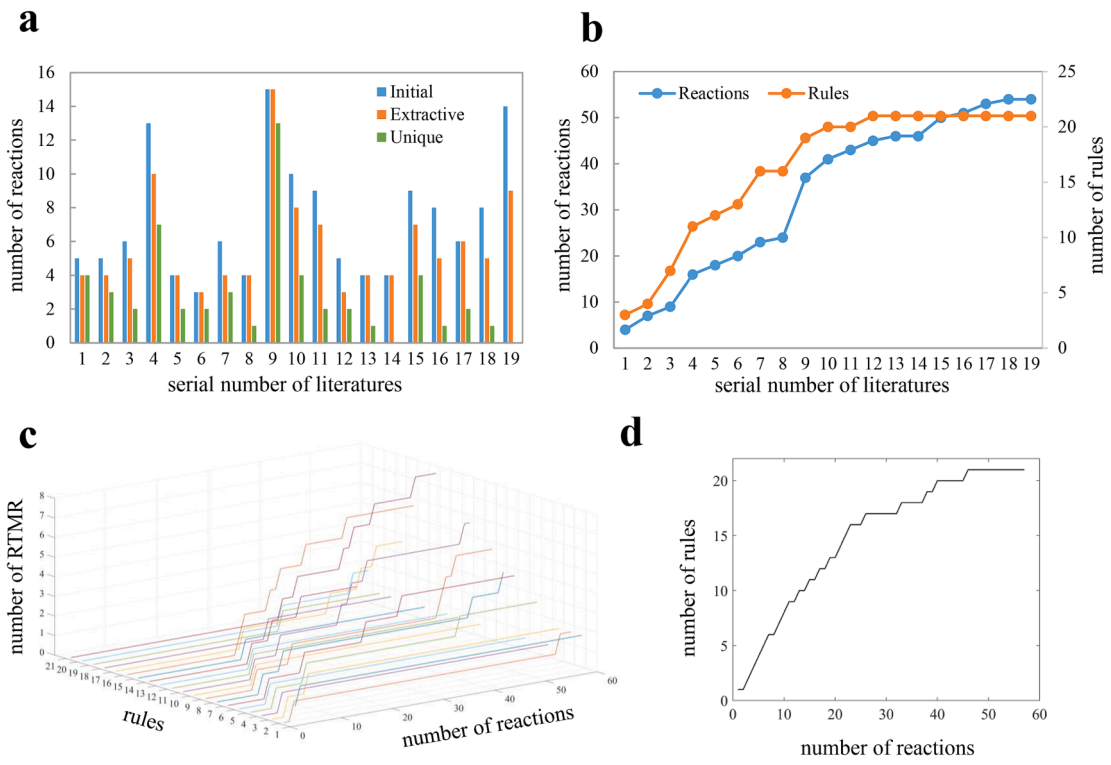


Fig. 10. Number of reactions, RTMR, and reaction rules in chronological order (a: number of reactions with the number of literatures, b: number of rules with the number of literatures, c: number of RTMR with the number of reactions, d: reactions rules vary with the number of reactions).

3.3. Reaction networks based on RTMR

According to the reaction rules extracted from RTMR, a reaction network can be established. The automatic generation of reaction networks is divided into two methods according to whether the new species generated by the reaction are used as reactants. As shown in Fig. 8**a** reaction network: first, the initial species were input and successively selected from the initial species library as reactants, and the corresponding products were generated according to the reaction rules (skip when there is no corresponding reaction). These reaction processes were recorded simultaneously, and new species and reaction correspondences were added to the species library and reaction library. When the entire initial species library is traversed, the program is completed, and the reaction network for this process can be obtained. However, for the reaction network shown in Fig. 8**b**, the newly generated species of the reaction can again be used as reactants until the program ends when all species have been traversed. The left side was defined as the Once-Reaction-Network and the right side as the Cyclic-Reaction-Network.

3.4. Verification

In order to check the effectiveness of the method, a few typical catalytic reactions were selected for verification. Note that catalytic reactions occurring at acidic sites, such as protonation, deprotonation, methylation, alkylation, β-scission, and so on, should go through the elementary steps of carbenium. The reaction rules of these reactions can be transformed into species matrix through the aforementioned method, and finally converted to RTMR.

Fig. 9 shows an example, in which five reactions including protonation, protonation, methylation, alkylation and β-scission were considered. These five reactions were selected as initial reactions and used to extract reaction rules that were represented by RTMR. Here the five reactions were automatically classified into four categories, since the reaction centers for methylation and alkylation were both carbenium and alkene. Then all the species involved in the five reactions were further used as initial input species for prediction. Based on the obtained RTMR, we applied the Once-Reaction-Network method to the initial input species and can automatically generate the reaction network showing on the right side of Fig. 9 as output. As can be seen, the output reaction network can not only completely cover that the initial five reactions, but also generate some new reactions such as the alkylation reaction "$ZCH_3 + CH_3CH = CH_2 \rightarrow ZCH_2CH(CH_3)CH_3$ (Z means the active center of molecular sieve […Si-O-Al…])". Apparently, the RTMR can effectively store the reaction rules which in fact contains the essential characteristics and features of reactions of interests.

4. Results and discussion

4.1. Induction period in methanol-to-olefins process

Methanol-to-olefins (MTO), which is a complex reaction process, is considered as an important route for light olefins production from no-oil resources such as coal, natural gas and biomass. Typically, MTO process includes three stages: the induction, autocatalytic and deactivation stage. The induction stage refers to the initial adsorption of methanol to the first carbon–carbon bond generation, as well as the generation and accumulation of hydrocarbon pool species (Hadi and Farzi, 2022; Standl and Hinrichsen, 2018; Tian et al., 2015; Wu et al., 2018; Yarulina et al., 2018). The second process is the autocatalytic step. The methanol reacts with the formed hydrocarbon pool species to generate low carbon olefin. Finally, the polycyclic aromatic hydrocarbons such as anthracene, phenanthrene and pyrene begin to appear in the hydrocarbon pool species and eventually form coke, resulting in the deactivation of the catalyst (Chowdhury et al., 2018). Many researchers have studied the induction phase and the dual-cycle process of MTO, and speculated the relevant reaction mechanisms and routes. They found that MTO contains a variety of different elementary reaction types, such as alkylation, β-scission, etc., and there are thousands of possible species and elementary reactions. In this part, taking the induction period of methanol to olefin as an example, reaction networks based on RTMR

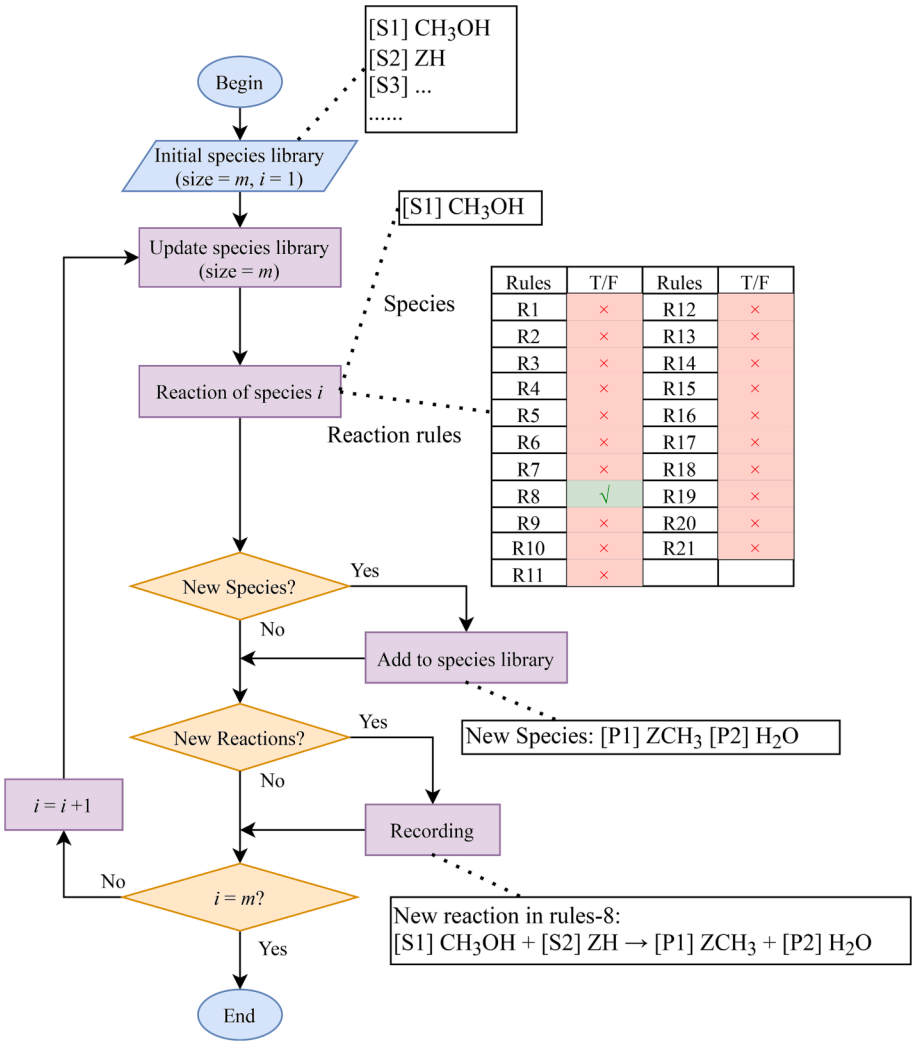


Fig. 11. Automatic generation of total reaction networks space of MTO induction period.

were established.

A total of 138 reactions were found in 19 literatures (Brogaard et al., 2014; Chowdhury et al., 2016; Chowdhury et al., 2018; Comas-Vives et al., 2015; Hutchings et al., 1987; Lesthaeghe et al., 2006, 2007; Li et al., 2014; Liu et al., 2016; Olsbye et al., 2015; Plessow and Studt, 2017; Stocker, 1999; Wang and Hunger, 2008; Wei et al., 2016; Weininger et al.; Wu et al., 2018; Wu et al., 2017; Yamazaki et al., 2011, 2012) available for MTO induction period, and amongst 111 were successfully extracted, with the extraction success rate reaching 80 % (see **Table S1-S4** in **Supplementary material**). According to the chronological order of publications, new reactions emerge as green in Fig. 10**a**. The blue line in Fig. 10**b** represents the cumulative value of new reactions, with a total of 54 non-repeated reactions. As mentioned earlier, the reactions were transformed into topological matrixes representation, and the transition state transformation process of each reaction equation was inverse deduced.

As can be seen from Fig. 10**c**, the number of reaction rules and corresponding RTMR of each reaction rule increase with the increase of the initial reaction number of inputs, and the total number of all reaction rules reaches 21. Note that the emergence of so many reaction rules is mainly due to the fact that we consider reactions in molecular sieves as well as that in the gas phase (see **Table S4** in **Supplementary material**). However, as can be seen from Fig. 10**d**, it can be found that the reaction rules were not necessarily increased by increasing the initial reaction number, but may also be a supplement to the existing reaction rules. So, the features of these new reactions will be stored in different RTMR in the existing reaction rules.

4.2. *Chemical reaction space of methanol-to-olefins process*

RTMR is first applied to generate chemical reaction space. Chemical reaction space, which refers to a large set of all possible species and reactions, has been increasingly used in machine learning to accurately predict some important molecular properties (Stocker et al., 2020). The formation of chemical reaction space is helpful for routes identification and mechanism discovery. For MTO induction period process, the Cyclic-Reaction-Network method can be used to obtain the reaction space under the condition of existing reaction rules from RTMR.

As mentioned above, the related algorithm was used to identify the characteristics of each group of reactions, and the reactions were divided into 21 groups according to the reaction center atoms, bond forming and bond breaking types. Then, 21 groups of RTMR and 21 groups of corresponding relations were established according to 21 types of reaction characteristics (here, the corresponding relations included the transformation rules of reactants to products and the definition domain of the transformation rule '$\widehat{F}(X)$').

As shown in the Fig. 11, the reaction network in the induction period of MTO was established, and reactions were carried out successively by importing species. Only when the species satisfy the domain of RTMR, the reaction can be carried out. For example, the acidic central reaction

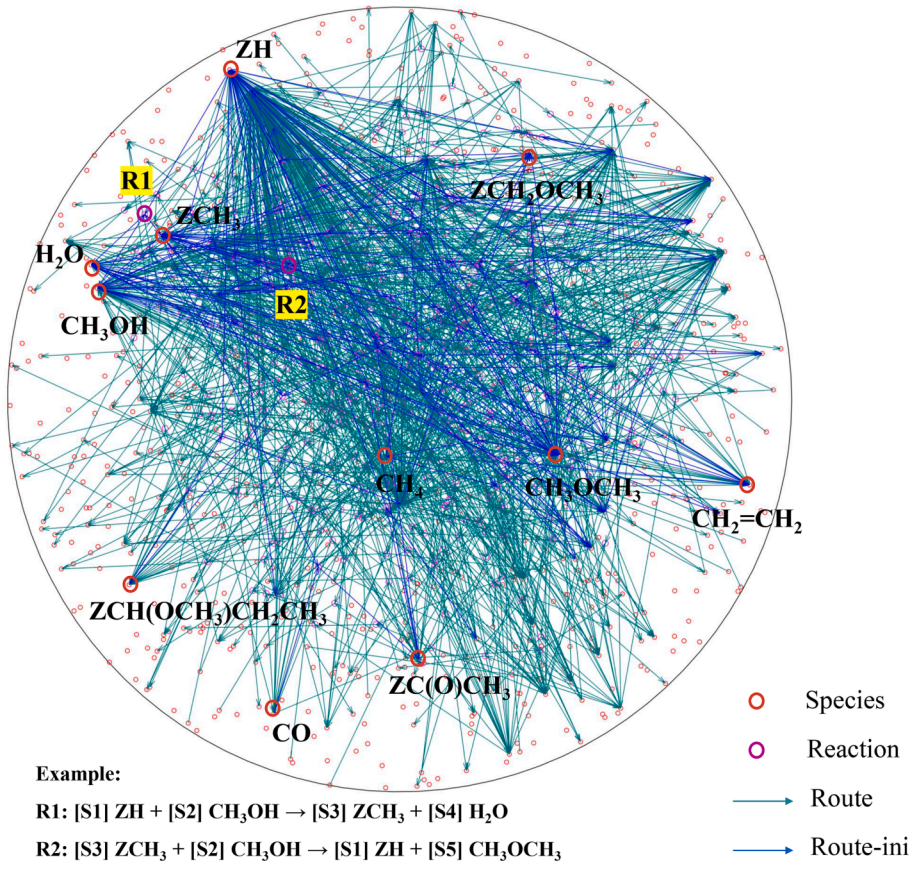


Fig. 12. Species and reactions of total reaction networks space (Some typical species and reactions are highlighted), small red circles represent species; purple circles represent reactions; green lines represent reactions by automatically generated; blue lines represent initial reactions).

between methanol and molecular sieve was described in rule-8. After standardization of the new species [P1] ZCH_3 [P2] H_2O was obtained, they were added to the end of the species library, and the reaction of rule-8 "[S1] CH_3OH + [S2] ZH → [P1] ZCH_3 + [P2] H_2O" was recorded. Under certain constraints, when all species participate in the reaction and no new species were generated in the species library, the whole cycle was completed, and finally all possible species and the reaction process in each RTMR of all rules were obtained.

Taking RTMR distance as 1, the reaction networks of MTO initial process based on 21 reaction rules can be obtained by the above algorithm. As shown in Fig. 12, unit circle reflected the topological space of reaction network; small red circles represented species; purple circles represented reactions; green lines represented reactions by automatically generated; blue lines represented initial reactions. The final reaction space contained a total of 845 species, 3748 reactions, a total of 21 categories.

4.3. *Prediction of reaction routes of methanol-to-olefins process*

RTMR can predict and connect chemical reaction routes. Finding simple and efficient reaction routes in complex chemical reaction space is usually a non-trivial task for chemical engineers (Kim et al., 2018). As

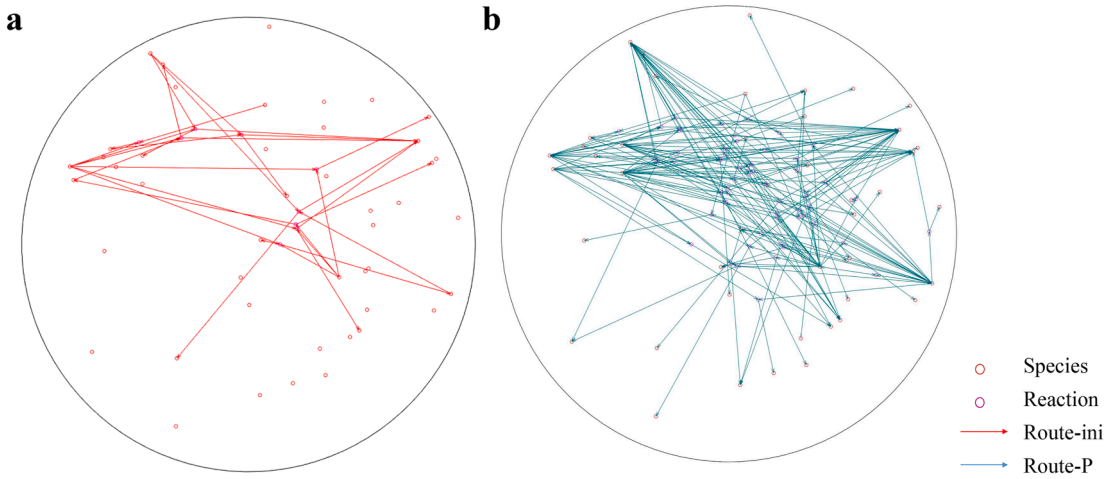


Fig. 13. Prediction of reaction routes of methanol-to-olefins process (a: Initial reaction networks in 19th articles; b: Predicted reaction networks).

the efficiency of RTMR depends on training samples, which is similar to machine learning and chemical engineers can easily implement it in the automatic reaction network generation.

As stated above, in MTO induction period process shown on Fig. 10**b**, there were 21 types of reactions based on the information accumulated in 12 articles, and the subsequent reactions supplemented were essentially the same type. Here, reactions following the reaction mechanism presented in the first 18 articles were now selected as training set for RTMR. By using the Once-Reaction-Network generation method and taking the reaction species and intermediates in 19th article as input files, we can generate a reaction network via reaction rules with the trained RTMR. In total 21 reaction rules from the first 18 articles were generated and stored in RTMR. Then all species in the 19th article were selected as input to automatically generate reaction network, including 2 reactants, 1 product and 7 intermediates (see **Table S6** in **Supplementary material** for details). Fig. 13**a** is the initial reaction networks in 19th articles (see **Table S2** in **Supplementary material**), and Fig. 13**b** shows the final predicted reaction network, which is consisting of 45 species and 63 reactions (see **Table S7** and **Table S8** in **Supplementary material**). By comparison, it can be found that the final predicted reactions completely cover the initial reactions, demonstrating the predictability of RTMR.

5. Conclusions

In this work, a strategy to automatically establish reaction networks directly from the reaction mechanisms using a *Reaction Rules Topological Matrix Representation* (RTMR) method was proposed. The reaction space of a system can be easily obtained according to the RTMR model after training. New reactions and routes can also be predicted by adding the necessary reactants and intermediates. Taking the induction period of methanol to olefins (MTO) process as an example, we showed how to use RTMR to extract the information of reaction mechanism presented in the literatures. It is expected that this strategy can extract reaction characteristics through a small amount of existing and representative reaction knowledge, which can be of interest for the automatic exploration of multiphase catalytic reaction mechanism via a knowledge-driven approach.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is part of a research program financially supported by the National Natural Science Foundation of China under Grant No. 21991093, 22293021, 22288101.

Appendix A. Supplementary data

The **Supplementary material** includes the details in Section 3.1: all initial reactions from literatures, reactions successfully extracted from initial reactions, the unique reactions, and the reactions after the classification. The **Supplementary material** includes the statistics of quantity of species and reactions of reaction networks of MTO induction period in Section 3.2. The **Supplementary material** includes the details in Section 3.3: all species from literature-19, and the species and reactions of reaction networks from automatic generation. Supplementary data to this article can be found online at https://doi.org/10.1016/j.ces.2023.119461.

References

Apostolakis, J., Sacher, O., Korner, R., Gasteiger, J., 2008. Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database. J. Chem. Inf. Model 48, 1190–1198.
Broadbelt, L.J., Stark, S.M., Klein, M.T., 1994. Computer-generated pyrolysis modeling - on-the-fly generation of species, reactions, and rates. Ind. Eng. Chem. Res. 33, 790–799.
Brogaard, R.Y., Henry, R., Schuurman, Y., Medford, A.J., Moses, P.G., Beato, P., Svelle, S., Norskov, J.K., Olsbye, U., 2014. Methanol-to-hydrocarbons conversion: the alkene methylation pathway. J. Catal. 314, 159–169.
Buda, F., Bounaceur, R., Warth, V., Glaude, P., Fournet, R., Battin-Leclerc, F., 2005. Progress toward a unified detailed kinetic model for the autoignition of alkanes from C-4 to C-10 between 600 and 1200 K. Combust. Flame 142, 170–186.
Chowdhury, A.D., Houben, K., Whiting, G.T., Mokhtar, M., Asiri, A.M., Al-Thabaiti, S.A., Basahel, S.N., Baldus, M., Weckhuysen, B.M., 2016. Initial carbon-carbon bond formation during the early stages of the methanol-to-olefin process proven by zeolite-trapped acetate and methyl acetate. Angew. Chem. Int. Ed. Engl. 55, 15840–15845.
Chowdhury, A.D., Paioni, A.L., Houben, K., Whiting, G.T., Baldus, M., Weckhuysen, B.M., 2018. Bridging the gap between the direct and hydrocarbon pool mechanisms of the methanol-to-hydrocarbons process. Angew. Chem. Int. Ed. Engl. 57, 8095–8099.
Clymans, P.J., Froment, G.F., 1984. Computer-generation of reaction paths and rate-equations in the thermal-cracking of normal and branched paraffins. Comput. Chem. Eng. 8, 137–142.
Comas-Vives, A., Valla, M., Coperet, C., Sautet, P., 2015. Cooperativity between Al sites promotes hydrogen transfer and carbon-carbon bond formation upon dimethyl ether activation on alumina. ACS Cent. Sci. 1, 313–319.
De Witt, M.J., Dooling, D.J., Broadbelt, L.J., 2000. Computer generation of reaction mechanisms using quantitative rate information: application to long-chain hydrocarbon pyrolysis. Ind. Eng. Chem. Res. 39, 2228–2237.
First, E.L., Gounaris, C.E., Floudas, C.A., 2012. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. J. Chem. Inf. Model. 52, 84–92.
Froment, G.F., 2013. Fundamental kinetic modeling of catalytic hydrocarbon conversion processes. Rev. Chem. Eng. 29, 385–412.
Gao, C.W., Allen, J.W., Green, W.H., West, R.H., 2016. Reaction mechanism generator: automatic construction of chemical kinetic mechanisms. Comput. Phys. Commun. 203, 212–225.
Golender, V.E., Drboglav, V.V., Rosenblit, A.B., 1981. Graph potentials method and its application for chemical information-processing. J. Chem. Inf. Comput. Sci. 21, 196–204.
Gupta, U., Vlachos, D.G., 2021. Learning chemistry of complex reaction systems via a python first-principles reaction rule stencil (pReSt) generator. J. Chem. Inf. Model. 61, 3431–3441.
Hadi, N., Farzi, A., 2022. A review on reaction mechanisms and catalysts of methanol to olefins process. Chem. Eng. Commun. 209, 1664–1710.
Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D., 2015. InChI, the IUPAC international chemical identifier. J Cheminform 7, 23.
Hillewaert, L.P., Dierickx, J.L., Froment, G.F., 1988. Computer-generation of reaction schemes and rate-equations for thermal-cracking. AIChE J 34, 17–24.
Hutchings, G.J., Gottschalk, F., Hall, M.V.M., Hunter, R., 1987. Hydrocarbon formation from methylating agents over the zeolite catalyst ZSM-5. Comments on the mechanism of carbon–carbon bond and methane formation. J. Chem. Soc. Faraday Trans. 1. Physical Chemistry in Condensed Phases 83, 571–583.
Karaba, A., Zamostny, P., Lederer, J., Belohlav, Z., 2013. Generalized model of hydrocarbons pyrolysis using automated reactions network generation. Ind. Eng. Chem. Res. 52, 15407–15416.
Kim, Y., Kim, J.W., Kim, Z., Kim, W.Y., 2018. Efficient prediction of reaction paths through molecular graph and reaction network analysis. Chem. Sci. 9, 825–835.
Korner, R., Apostolakis, J., 2008. Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. J. Chem. Inf. Model. 48, 1181–1189.
Lesthaeghe, D., Van Speybroeck, V., Marin, G.B., Waroquier, M., 2006. What role do oxonium ions and oxonium ylides play in the ZSM-5 catalysed methanol-to-olefin process? Chem. Phys. Lett. 417, 309–315.
Lesthaeghe, D., Van Speybroeck, V., Marin, G.B., Waroquier, M., 2007. The rise and fall of direct mechanisms in methanol-to-olefin catalysis: an overview of theoretical contributions. Ind. Eng. Chem. Res. 46, 8832–8838.
Li, J.F., Wei, Z.H., Chen, Y.Y., Jing, B.Q., He, Y., Dong, M., Jiao, H.J., Li, X.K., Qin, Z.F., Wang, J.G., Fan, W.B., 2014. A route to form initial hydrocarbon pool species in methanol conversion to olefins over zeolites. J. Catal. 317, 277–283.
Liu, Y., Muller, S., Berger, D., Jelic, J., Reuter, K., Tonigold, M., Sanchez-Sanchez, M., Lercher, J.A., 2016. Formation mechanism of the first carbon-carbon bond and the first olefin in the methanol conversion into hydrocarbons. Angew. Chem. Int. Ed. Engl. 55, 5723–5726.
Olsbye, U., Svelle, S., Lillerud, K.P., Wei, Z.H., Chen, Y.Y., Li, J.F., Wang, J.G., Fan, W.B., 2015. The formation and degradation of active species during methanol conversion over protonated zeotype catalysts. Chem. Soc. Rev. 44, 7155–7176.
Plessow, P.N., Studt, F., 2017. Unraveling the mechanism of the initiation reaction of the methanol to olefins process using ab initio and DFT calculations. ACS Catal. 7, 7987–7994.

Puliyanda, A., Srinivasan, K., Sivaramakrishnan, K., Prasad, V., 2022. A review of automated and data-driven approaches for pathway determination and reaction monitoring in complex chemical systems. Digital Chem. Eng. 2.
Rangarajan, S., Bhan, A., Daoutidis, P., 2010. Rule-based generation of thermochemical routes to biomass conversion. Ind. Eng. Chem. Res. 49, 10459–10470.
Raymond, J.W., Gardiner, E.J., Willett, P., 2002. RASCAL: calculation of graph similarity using maximum common edge subgraphs. Comput. J. 45, 631–644.
Standl, S., Hinrichsen, O., 2018. Kinetic modeling of catalytic olefin cracking and methanol-to-olefins (MTO) over zeolites: a review. Catalysts 8, 626.
Steiner, M., Reiher, M., 2022. Autonomous reaction network exploration in homogeneous and heterogeneous catalysis. Top. Catal. 65, 6–39.
Stocker, M., 1999. Methanol-to-hydrocarbons: catalytic materials and their behavior. Microporous Mesoporous Mater. 29, 3–48.
Stocker, S., Csanyi, G., Reuter, K., Margraf, J.T., 2020. Machine learning in chemical reaction space. Nat. Commun. 11, 5505.
Suleimanov, Y.V., Green, W.H., 2015. Automated discovery of elementary chemical reaction steps using freezing string and berny optimization methods. J. Chem. Theory Comput. 11, 4248–4259.
Thybaut, J.W., Marin, G.B., 2013. Single-Event MicroKinetics: catalyst design for complex reaction networks. J. Catal. 308, 352–362.
Tian, P., Wei, Y.X., Ye, M., Liu, Z.M., 2015. Methanol to olefins (MTO): from fundamentals to commercialization. ACS Catal. 5, 1922–1938.
Turtscher, P.L., Reiher, M., 2023. Pathfinder horizontal line navigating and analyzing chemical reaction networks with an efficient graph-based approach. J. Chem. Inf. Model. 63, 147–160.
Unsleber, J.P., 2023. Accelerating reaction network explorations with automated reaction template extraction and application. J. Chem. Inf. Model. 63, 3392–3403.
Unsleber, J.P., Reiher, M., 2020. The exploration of chemical reaction networks. Annu. Rev. Phys. Chem. 71, 121–142.
Vandewiele, N.M., Van Geem, K.M., Reyniers, M.F., Marin, G.B., 2012. Genesys: kinetic model construction using chemo-informatics. Chem. Eng. J. 207, 526–538.
Vernuccio, S., Broadbelt, L.J., 2019. Discerning complex reaction networks using automated generators. AIChE J 65, e16663.
Walters, W.P., Yalkowsky, S.H., 1996. ESCHER - a computer program for the determination of external rotational symmetry numbers from molecular topology. J. Chem. Inf. Comput. Sci. 36, 1015–1017.
Wang, W., Hunger, M., 2008. Reactivity of surface alkoxy species on acidic zeolite catalysts. Acc. Chem. Res. 41, 895–904.
Wei, Z.H., Chen, Y.Y., Li, J.F., Wang, P.F., Jing, B.Q., He, Y., Dong, M., Jiao, H.J., Qin, Z. F., Wang, J.G., Fan, W.B., 2016. Methane formation mechanism in the initial methanol-to-olefins process catalyzed by SAPO-34. Cat. Sci. Technol. 6, 5526–5533.
Weininger, D., 1988. Smiles, a chemical language and information-system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28, 31–36.
Weininger, D., Weininger, A., Weininger, J.L., 1989. Smiles. 2. algorithm for generation of unique smiles notation. J. Chem. Inf. Comput. Sci. 29, 97–101.
Wen, M., Spotte-Smith, E.W.C., Blau, S.M., McDermott, M.J., Krishnapriyan, A.S., Persson, K.A., 2023. Chemical reaction networks and opportunities for machine learning. Nature Computational Science 3, 12–24.
Wu, X., Xu, S., Zhang, W., Huang, J., Li, J., Yu, B., Wei, Y., Liu, Z., 2017. Direct mechanism of the first carbon-carbon bond formation in the methanol-to-hydrocarbons process. Angew. Chem. Int. Ed. Engl. 56, 9039–9043.
Wu, X., Xu, S., Wei, Y., Zhang, W., Huang, J., Xu, S., He, Y., Lin, S., Sun, T., Liu, Z., 2018. Evolution of C-C bond formation in the methanol-to-olefins process: from direct coupling to autocatalysis. ACS Catal. 8, 7356–7361.
Yamazaki, H., Shima, H., Imai, H., Yokoi, T., Tatsumi, T., Kondo, J.N., 2011. Evidence for a “carbene-like” intermediate during the reaction of methoxy species with light alkenes on H-ZSM-5. Angew. Chem. Int. Ed. Engl. 50, 1853–1856.
Yamazaki, H., Shima, H., Imai, H., Yokoi, T., Tatsumi, T., Kondo, J.N., 2012. Direct production of propene from methoxy species and dimethyl ether over H-ZSM-5. J. Phys. Chem. C 116, 24091–24097.
Yarulina, I., Chowdhury, A.D., Meirer, F., Weckhuysen, B.M., Gascon, J., 2018. Recent trends and fundamental insights in the methanol-to-hydrocarbons process. Nat. Catal. 1, 398–411.
Zámostný, P., Karaba, A., Olahová, N., Petrů, J., Patera, J., Hájeková, E., Bajus, M., Bělohlav, Z., 2014. Generalized model of n-heptane pyrolysis and steam cracking kinetics based on automated reaction network generation. J. Anal. Appl. Pyrol. 109, 159–167.