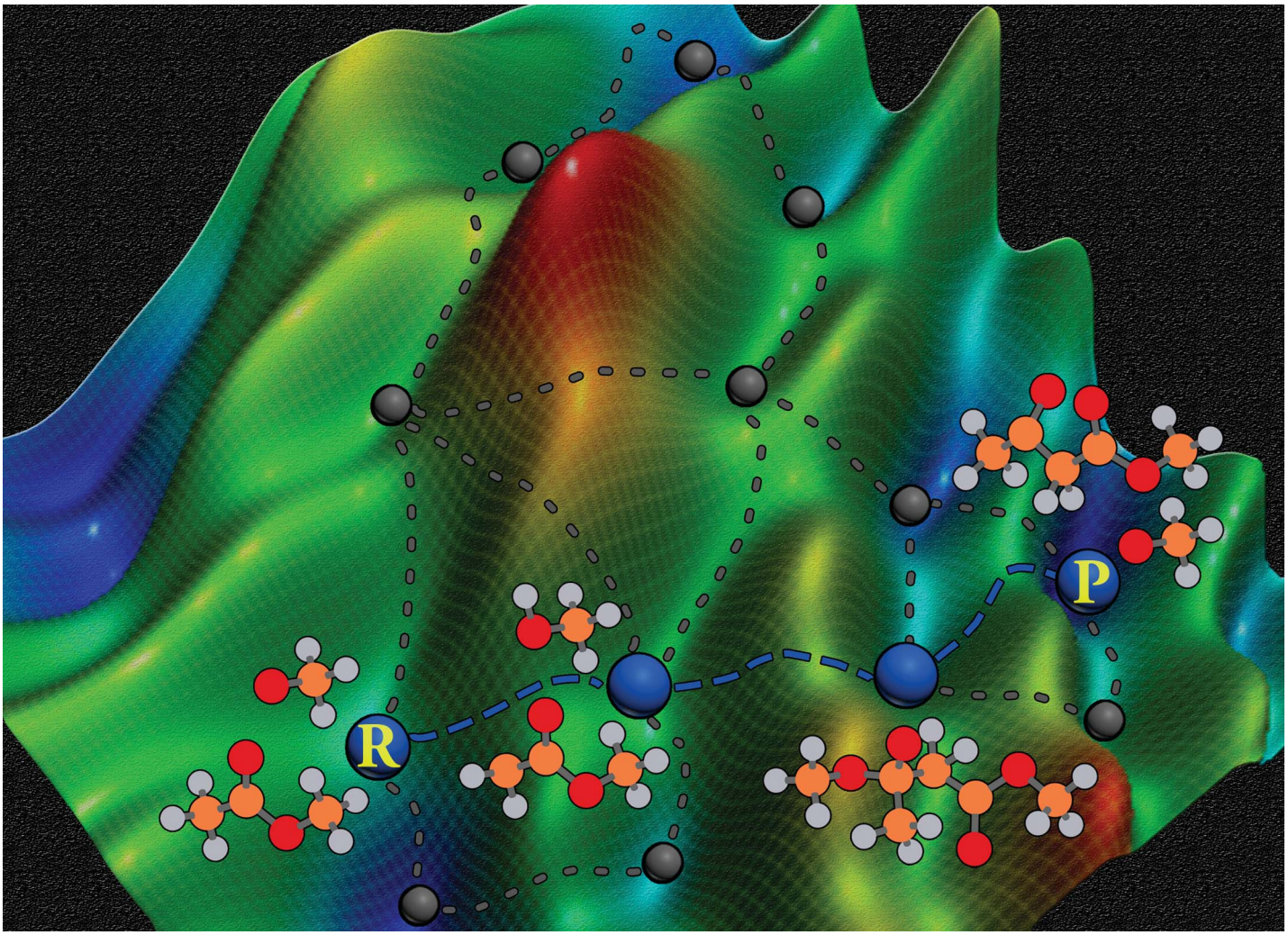


Showcasing research from Professor Woo Youn Kim's group, Department of Chemistry, KAIST, Daejeon, Korea.

Efficient prediction of reaction paths through molecular graph and reaction network analysis

A minimal subnetwork containing favourable reaction mechanisms is extracted very rapidly from a full reaction network by counting the number of dissociated and formed bonds in molecular graphs. This approach was successfully applied to two reaction examples.

As featured in:

See Woo Youn Kim *et al.*, *Chem. Sci.*, 2018, **9**, 825.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

Registered charity number: 207890

Chemical Science

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2018, **9**, 825

Efficient prediction of reaction paths through molecular graph and reaction network analysis†

Yeonjoon Kim, Jin Woo Kim, Zeehyo Kim and Woo Youn Kim *

Received 18th August 2017
Accepted 11th December 2017

DOI: 10.1039/c7sc03628k

rsc.li/chemical-science

Despite remarkable advances in computational chemistry, prediction of reaction mechanisms is still challenging, because investigating all possible reaction pathways is computationally prohibitive due to the high complexity of chemical space. A feasible strategy for efficient prediction is to utilize chemical heuristics. Here, we propose a novel approach to rapidly search reaction paths in a fully automated fashion by combining chemical theory and heuristics. A key idea of our method is to extract a minimal reaction network composed of only favorable reaction pathways from the complex chemical space through molecular graph and reaction network analysis. This can be done very efficiently by exploring the routes connecting reactants and products with minimum dissociation and formation of bonds. Finally, the resulting minimal network is subjected to quantum chemical calculations to determine kinetically the most favorable reaction path at the predictable accuracy. As example studies, our method was able to successfully find the accepted mechanisms of Claisen ester condensation and cobalt-catalyzed hydroformylation reactions.

Introduction

Computational chemistry is a powerful approach for the mechanistic study of chemical reactions, because it can offer deep insight on reaction mechanisms at the atomistic level.[1–3] Remarkable advances in quantum chemistry especially with the rise of density functional theory (DFT) have provided a handful tool to obtain energy profiles of reaction paths for verifying experimentally and/or intuitively proposed mechanisms. However, prediction of chemical reactions is still challenging because exploring all possible paths on potential energy surface (PES) is intractable due to its high complexity.[4–6]

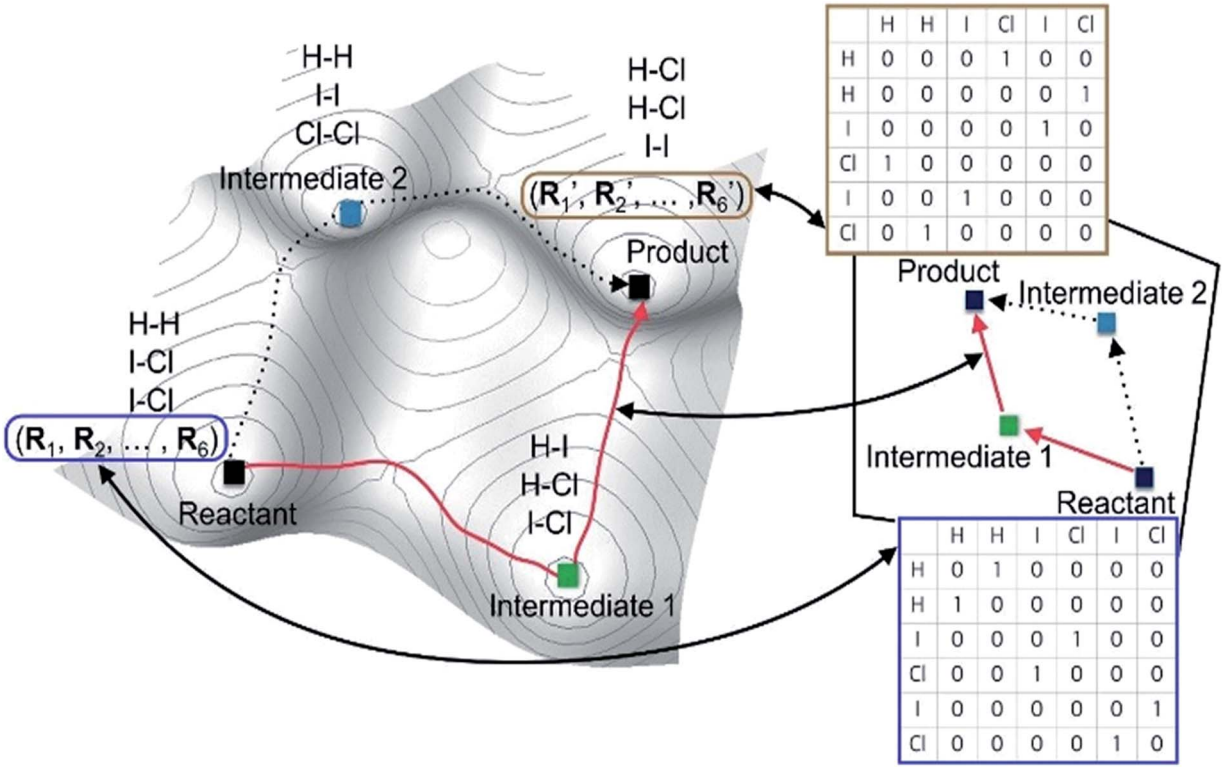
Substantial efforts have been devoted to developing automated exploration methods of reaction paths on the PES.[7–16] Maeda and coworkers developed so-called anharmonic downward distortion following method for global exploration of isomerization paths for a single molecule.[7,8] They also proposed the artificial force-induced reaction method that finds a reaction path through accelerated chemical reactions with an artificial force.[9–11] Local minima-sampling methods such as basin-hopping Monte-Carlo and minima hopping algorithms can be used to find appropriate reaction intermediates.[6,17–20] We also reported a graph-theoretic approach combined with the basin-hopping Monte-Carlo algorithm.[21] Heuristic rules combined with quantum chemistry were utilized for the efficient generation of intermediates. For instance, Bergler and coworkers discovered new intermediates through the structural relaxation of numerous reactive complexes prepared according to their reactivity.[22,23] This approach has been used for the automated exploration of reaction paths with the quantification of uncertainties in solving rate equations.[24] An efficient method combining a transition state search using accelerated chemical dynamics with kinetic Monte-Carlo simulations has also been devised for the kinetic study of organometallic catalysis.[25,26] However, those methods inevitably require large computational costs, since every movement of molecules on the PES entails quantum chemical calculations.

Alternatively, chemical reactions can be described by the successive change in the chemical bonds of reactive molecules. Stable molecular structures as local minima on the PES can be mapped to molecular graphs, as illustrated in Fig. 1. In this point of view, chemical reactions may be equivalently described by the successive conversion of a reactant molecular graph into isomeric graphs. In fact, this kind of graph-theoretic approaches has attracted great attention in the past for computer-assisted mechanistic study thanks to their efficient algorithm accelerated by heuristic rules.[27–45]

Most graph-theoretic methods adopt one of the following three steps or a combination of them. First, the combinatorial enumeration of molecular graphs generates a set of molecules that can be made from the reactants. For example, if a table of reaction graphs is available, Pólya's theorem offers an analytical enumeration technique considering permutations due to

Department of Chemistry, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Korea. E-mail: wooyoun@kaist.ac.kr

† Electronic supplementary information (ESI) available: Detailed information on reaction networks and pathways for two example reactions, Cartesian coordinates of molecules in the reaction networks obtained at the DFT level for the hydroformylation example, and conformers and isomers of the intermediates in the Heck–Breslow mechanism. See DOI: 10.1039/c7sc03628k



	H	H	I	Cl	I	Cl
H	0	0	0	1	0	0
H	0	0	0	0	0	1
I	0	0	0	0	1	0
Cl	1	0	0	0	0	0
I	0	0	1	0	0	0
Cl	0	1	0	0	0	0

	H	H	I	Cl	I	Cl
H	0	1	0	0	0	0
H	1	0	0	0	0	0
I	0	0	0	1	0	0
Cl	0	0	1	0	0	0
I	0	0	0	0	0	1
Cl	0	0	0	0	1	0

Fig. 1 Correspondence between a real potential energy surface and a hypothetical chemical space represented with molecular graphs for the reaction in eqn (1) as an example.

molecular symmetry.[46,47] A constructive enumeration algorithm can also be used to find a complete list of graphs.[48–50] Second, molecules obtained from the resulting graphs are linked with each other to make hypothetical elementary reactions, leading to a chemical reaction network. Finally, the network is analyzed to determine kinetically favorable reaction mechanisms. In each procedure, they adopt heuristic rules based on chemical concepts and databases, which limits their applicability and reliability. In 2017, Segler and Waller proposed a data-driven model to improve such rule-based methods by predicting the reactivity of molecules based on the complete published knowledge, but its use is limited to binary reactions.[51]

More recently, complementary approaches are being actively developed to exploit the advantages of both graph-theoretic and quantum chemical methods.[22,23,52–61] Molecular graphs are used to generate hypothetical reaction intermediates. To find kinetically feasible elementary reaction steps, the activation energy between two intermediates is explicitly calculated using conventional methods such as the nudged elastic band[52,53] and eigenvector-following with freezing[22,54] and growing[55–58] string methods. Single-ended algorithms such as Berny optimization and intrinsic reaction coordinate calculations[62–64] can be employed for further improvements. To consider multiple paths, Habershon devised a novel constrained molecular dynamics using a model Hamiltonian.[52,53] Hammond's postulate[65] was also used for efficiency.[59,60] These methods aim to automatically discover reaction mechanisms from a single input molecular structure with minimal human efforts and were successfully applied to several organic reactions. However, molecular graph enumeration results in a huge number of intermediates and hypothetical elementary reactions due to combinatorial explosion. As a result, calculating transition states for every elementary reaction is a computational bottleneck. Therefore, it is critical to remove chemically irrelevant hypothetical elementary reactions in an efficient way for the success of such automated approaches.

Here we propose a fast prediction method of reaction paths through molecular graph and reaction network analysis. Our method adopts some idea of the aforementioned methods such as molecular graph enumeration,[52–61] but also introduces new fascinating features; a key distinctive one is to efficiently extract the minimal subnetwork from a complex full network. This can be done by exploring multiple reaction paths connecting reactants and products with minimum dissociation and formation of chemical bonds using a graph-theoretic method. Another important feature of our method is its wide applicability. This is because the method is able to explore in principle all possible reaction routes by considering all combinations of chemical bond formation and dissociation, which is typically intractable. To make it computationally efficient, we devised *de novo* protocols to rule out many unimportant reaction routes and intermediates according to general chemical rules. As a result, fast searching for most plausible reaction paths is feasible within an hour on a single workstation. The resultant paths can be verified with further refinements using quantum chemical methods. This first-screening and then-verifying strategy

minimizes expensive computational parts, which is critical to achieve both predictive power and efficiency. In what follows, we first explain the details of the proposed method. Then, to demonstrate its reliability and efficiency, we provide two example studies: a well-known organic reaction and a simple organometallic reaction. Finally, we conclude with a summary and outlook for future works.

Methods

Most graph-theoretic approaches are based on the following concept; chemical reactions can be described by changing molecular graphs according to heuristic rules. This implicitly assumes that there exists a hypothetical chemical space, in which molecular structures are expressed with graphs. Fig. 1 schematically illustrates the relation between a real PES and the corresponding chemical space. The left figure shows the PES of the following reaction (eqn (1)) as an example.

$$2ICl + H_2 \rightarrow HI + HCl + ICl \rightarrow 2HCl + I_2. \quad (1)$$

Local minima on the PES include reactants, intermediates, and products. The chemical reaction occurs along the minimum energy path between the reactants and the products, as indicated by the red line in Fig. 1. The same chemistry can be described in the hypothetical chemical space (the right side of Fig. 1). The local minima can be mapped to molecular graphs in the chemical space. An elementary reaction between two intermediates corresponds to an edge linking two graphs. The length of each edge is related to the rate constant of the corresponding elementary reaction.

A key advantage of using the hypothetical chemical space is that all possible molecular graphs, each of which corresponds to a stable molecular structure, can efficiently be generated by the combinatorial enumeration of an input graph, resulting in an extensive set of reaction intermediates without resorting to quantum chemical calculations. This fascinating feature enables us to avoid huge computational costs required by direct searching methods on PES. However, the hypothetical chemical space may not be complete to encompass the entire PES; for instance, conformational isomers are mapped to an identical molecular graph. Some of such problems can be resolved by using additional quantum chemical methods. Hence, this approach can be applied to a wide range of chemical reactions.

Fig. 2 shows the flowchart of our method. As in previous works,[52–61] we use molecular graphs expressed specifically with an atom connectivity (AC) matrix to represent molecular structures. Namely the bond-electron matrix can also be used as an alternative to the AC matrix.[27,28,54] It is different from the AC matrix in its diagonal elements containing the number of valence electrons that do not participate in chemical bonds. Therefore, an electron-pushing model mimicking the language of organic chemistry can be utilized for matrix enumeration. However, those two models are equivalent with one another in a sense that one of them can be converted to the other by a graph-theoretic analysis.[66,67]

Since we aim to find a reaction path from given reactants ($\boldsymbol{R}$) to designated products ($\boldsymbol{P}$), both $\boldsymbol{R}$ and $\boldsymbol{P}$ structures are given as an input and converted to the corresponding AC matrices in Step 1. As is indicated by the shaded color, the reactant and product matrices may contain a few block matrices, each of which denotes a constituent molecule of $\boldsymbol{R}$ and $\boldsymbol{P}$, respectively. In Step 2, a number of AC matrices are generated by consecutively applying a set of conversion matrices ($\{\boldsymbol{C}\}$) to $\boldsymbol{R}$ until satisfying predefined termination criteria, which corresponds to the combinatorial enumeration of molecular graphs. In Step 3, a reaction network is constructed by calculating the length of edges between the AC matrices. Finally, the reaction network is analyzed to determine kinetically favorable reaction paths. Steps 1 and 2 are similar to other methods.[29] In Step 3, however,

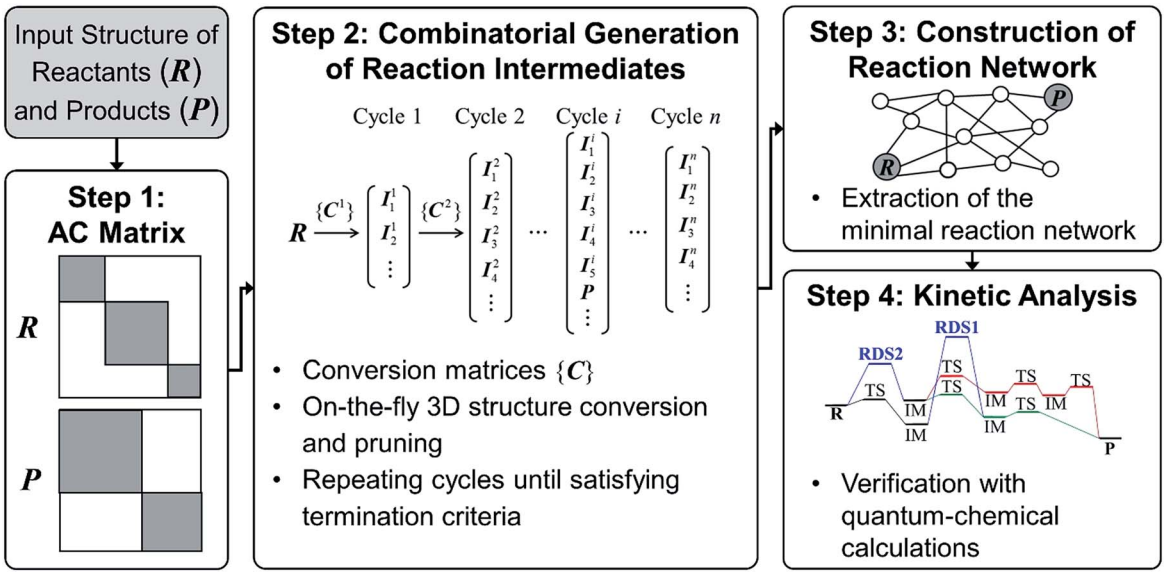


Fig. 2 Flowchart of our graph-based method for fast prediction of reaction paths.

we introduce a novel graph-theoretic method to extract a subnetwork including an essential part directly relevant to reaction mechanism from the full network. In Step 4, we perform transition state calculations for the extracted minimal reaction network to determine kinetically the most favorable reaction path. In what follows, we explain Steps 2–4 more in detail.

1. Step 2: combinatorial generation of reaction intermediates

Enumeration of AC matrices. For the combinatorial sampling of intermediates using molecular graphs, we start with the AC matrix of reactants as illustrated in Fig. 2. In the first cycle, a set of conversion matrices ($\{\boldsymbol{C}^1\}$) is applied to $\boldsymbol{R}$ to generate a set of intermediates ($\{\boldsymbol{I}^1\}$) through dissociation and formation of bonds; $\boldsymbol{R} + \boldsymbol{C}_j^1 = \boldsymbol{I}_j^1$. In the second cycle, a new set of conversion matrices ($\{\boldsymbol{C}^2\}$) is constructed for each of $\{\boldsymbol{I}^1\}$ and is applied to it to obtain new intermediates; $\boldsymbol{I}_j^1 + \boldsymbol{C}_j^2 = \boldsymbol{I}_j^2$. This process is repeated until predefined termination criteria are satisfied. The conversion matrix $\boldsymbol{C}$ for a given intermediate $\boldsymbol{I}$ is constructed as follows. Elements of $\boldsymbol{C}$ consist of 1, −1, and 0, which correspond to the formation, cleavage, and no change of chemical bonds, respectively. To take into account all possible combinations of chemically allowed matrix elements, we use the following rules:

$$I_{ii} = 0 \Rightarrow C_{ii} = 0$$
$$I_{ij} = \begin{cases} 0 & \Rightarrow C_{ij} = \begin{cases} 0 & \text{if } R_{ij} = 1 \\ 0 \text{ or } 1 & \text{otherwise,} \end{cases} \\ 1 & \Rightarrow C_{ij} = \begin{cases} -1 & \text{if } R_{ij} = 1 \\ 0 & \text{otherwise.} \end{cases} \end{cases} \quad (2)$$

All diagonal elements are zero. If an off-diagonal element of $\boldsymbol{I}$ is zero ($I_{ij} = 0$), the corresponding element of $\boldsymbol{C}$ can be either 0 or 1. If the same row-column element of $\boldsymbol{R}$ is 1 ($R_{ij} = 1$), $C_{ij} = 0$ because $I_{ij} = 0$ means that the chemical bond between atoms i and j in reactants has been broken in a previous cycle. This condition is necessary to prevent from generating AC matrices appeared in previous cycles once again. Otherwise, $C_{ij} = 0$ or 1. Similarly, for $I_{ij} = 1$, $C_{ij} = -1$, if $R_{ij} = 1$. Otherwise, $C_{ij} = 0$.

The above rules will produce all possible conversion matrices for each $\boldsymbol{I}$. However, it is inefficient for a large matrix. Various user-defined constraints may be helpful to reduce computational costs. At the same time, excessive constraints may provoke biased results. To compromise between efficiency and reliability, we apply only the following two constraints. The maximum number of bond formations and dissociations at each elementary reaction is limited to two, respectively, *i.e.*,

$$-2 \le \frac{1}{2}\sum_{ij} C_{ij} \le 2 \text{ and } \frac{1}{2}\sum_{ij} |C_{ij}| \le 4. \quad (3)$$

The same constraint was imposed in other works.[54,55] Only unimolecular and bimolecular reactions at each elementary reaction are allowed. This constraint is problematic for termolecular reactions, but it still encompasses most organic reactions. These two constraints can be controlled as input variables. It is also important to delete permutational isomers (or isomorphic copies) produced by the combinatorial enumeration. They can be discriminated by investigating the eigenvalues of the *alternative* Coulomb matrix of each AC matrix, which is modified from the original Coulomb matrix to detect graph isomorphism.[21] This conversion cycle is continued to find a number of intermediates including products and finally terminated if no new matrix is generated. Other termination conditions such as the maximum number of cycles can also be imposed.

On-the-fly 3D structure conversion and pruning. AC matrices at each cycle are subjected to on-the-fly 3D structure conversion and pruning. In our previous work, we proposed a reliable method that sequentially converts a given AC matrix to a bond order matrix, then to a SMILES code, and finally to a 3D geometry.[66,67] The reliability of this process has been proved by successfully applying it to 10 000 organic molecules randomly chosen from the PubChem database.[66] While we refer to ref. 66 for the technical details of the method, we here provide its overall procedure briefly. At each step, we screen out inappropriate molecular structures as follows. Information on the atomic valence and formal charge of each atom can be deduced by transforming AC matrices to bond order matrices. Molecules having atoms with inappropriate atomic valence or formal charge are discarded. In the SMILES conversion step, those having an inappropriate number of rings for a given reaction are removed. The remaining SMILES codes are then converted to 3D geometries, which are subsequently optimized by using conventional methods with desirable accuracy.

Our structure conversion method also yields all stereoisomers of organic molecules that can be constructed from a given AC matrix; *cis–trans* isomers and enantiomers can be specified explicitly by SMILES, and they are readily converted to 3D geometries, as explained in ref. 66. To find all possible conformers sharing an identical AC, we perform additional basin-hopping Monte-Carlo samplings with bond constraints to prevent from breaking the AC, as explained in ref. 21. However, conformers and stereoisomers of a metal complex cannot be specified by SMILES. Therefore, we developed a new method combining the bond constraint basin-hopping Monte-Carlo method with force field calculations. At each basin-hopping sampling cycle, the relative positions of ligands with respect to the metal center are distributed randomly, and this new structure is relaxed using force field calculations with fixed AC and bond orders. After finishing the combinatorial enumeration, molecules with energy higher than a threshold are screened out. The threshold energy is defined as the sum of reactant energy and a given tolerance value (E_{tol}).

Introduction to active atoms. Although the combinatorial sampling with the above rules is very comprehensive, it is demanding to deal with a large number of atoms due to combinatorial explosion. Fortunately, we note that, in most chemical reactions, only a few atoms regarded as reaction centers are directly involved in bond formation and dissociation, even for large molecules. Therefore, we designate these special atoms as 'active atoms' and build AC matrices on the

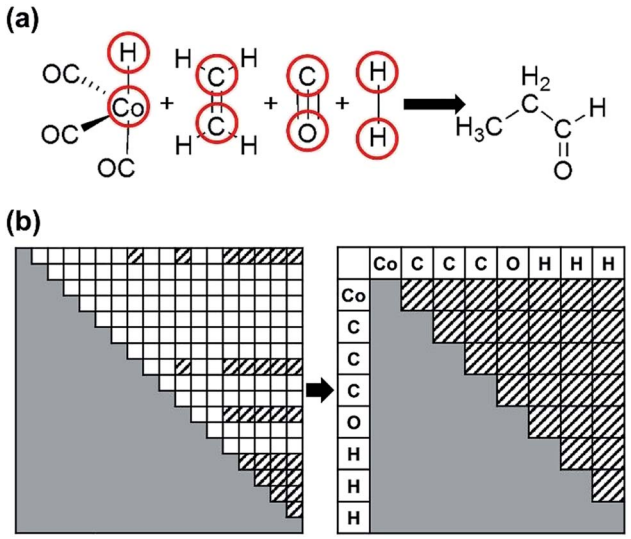


Fig. 3 (a) The reactant and product molecules for the cobalt-catalyzed hydroformylation reaction. The red circles indicate active atoms. (b) Construction of an active-atom connectivity matrix from an all-atom connectivity matrix. Matrix elements filled with slashes indicate bonds between active atoms.

basis of these active atoms. Fig. 3a shows an example of active atoms for the hydroformylation reaction, as denoted by red circles. Then, the initial AC matrix of the reactants and corresponding conversion matrices are reduced to the ones on the basis of only the eight active atoms as shown in Fig. 3b. Once the enumeration of AC matrices is completed, their basis transformation from active-atom to all-atom is followed to obtain all-atom AC matrices. Resulting matrices may contain a single molecule or several molecules. In the latter case, they are decomposed into several block matrices, each of which corresponds to a single molecule.

2. Step 3: construction of reaction network

Construction of reaction network. A reaction network can be made by connecting remaining intermediates after the structure conversion. Since we are interested in the most favorable reaction path starting from reactants to products, a subnetwork including both the reactants and products is crucial. A small subnetwork may not include important intermediates, whereas a large one entails computational costs. Thus, we need to determine an appropriate range of subnetwork to compromise between accuracy and efficiency prior to constructing the reaction network. We invoke the so-called principle of minimum structure change, which states that most chemical reactions proceed along a pathway with minimum dissociation and formation of bonds.[68,69] This heuristic rule, often regarded as a principle, has been applied to various chemical problems including the elucidation of reaction mechanism.[68,69] To implement this idea in our method, we devised a novel way of discarding intermediates that are placed too far from reactants and products in a reaction network. The concept of chemical distance (CD)[68,69] is used to perform such geometric analysis, which is defined as

$$\mathrm{CD}(\boldsymbol{A},\boldsymbol{B}) = \frac{1}{2}\sum_{ij}\left|A_{ij} - B_{ij}\right|, \tag{4}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ denote the AC matrices of two intermediates, respectively. The CD gives the minimum number of bond changes needed to transform $\boldsymbol{A}$ into $\boldsymbol{B}$. It can be overestimated due to the permutation between the two AC matrices, as illustrated in Fig. 4a. This problem can be resolved by calculating the minimum CD out of all possible combinations using the mixed-integer linear programming (MILP) scheme with appropriate variables and objective functions,[70] as shown in Fig. 4b.

Then, we collect all intermediates that satisfy the following criterion:

$$\mathrm{CD}(\boldsymbol{R},\boldsymbol{I}) + \mathrm{CD}(\boldsymbol{I},\boldsymbol{P}) \leq \mathrm{CD}(\boldsymbol{R},\boldsymbol{P}) + \Delta, \tag{5}$$

where Δ is called the 'digression factor'. This criterion determines intermediates that are located inside an ellipse whose focal points correspond to $\boldsymbol{R}$ and $\boldsymbol{P}$ in the reaction network, as illustrated in Fig. 5. We note that the screening criterion using eqn (5) can also be applied during the combinatorial generation of AC matrices to further accelerate the process for large systems. The factor Δ can be regarded as a convergence parameter which is determined in a way so that top ranked reaction paths in the final stage do not change.

The remaining intermediates are used to build a reaction network. These intermediates are regarded as vertices in the network. It should be noted that molecular conformers having an identical AC matrix share a single vertex. However, they can be treated independently as evaluating the activation energy of each conformer in the next step. The network with N intermediates can have $N(N-1)/2$ connections, which makes the process computationally demanding for a large value of N. To find kinetically appropriate elementary reactions, the same criteria used in Step 2 for the intermediate sampling are

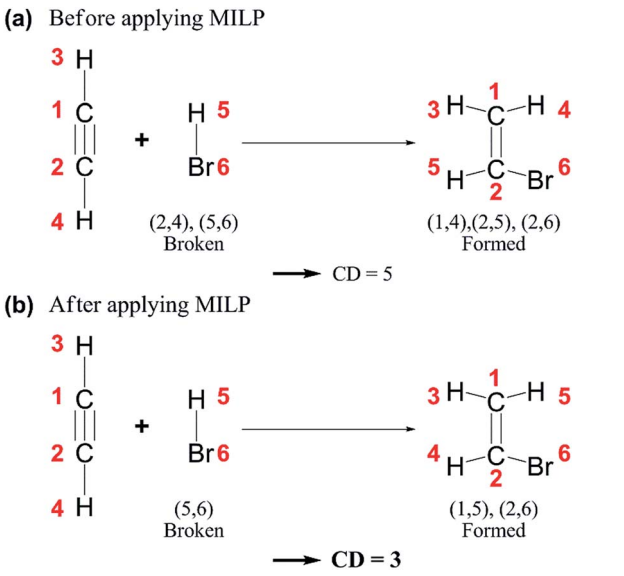


Fig. 4 Chemical distances (CDs) of an example reaction (a) before and (b) after applying mixed-integer linear programming (MILP).

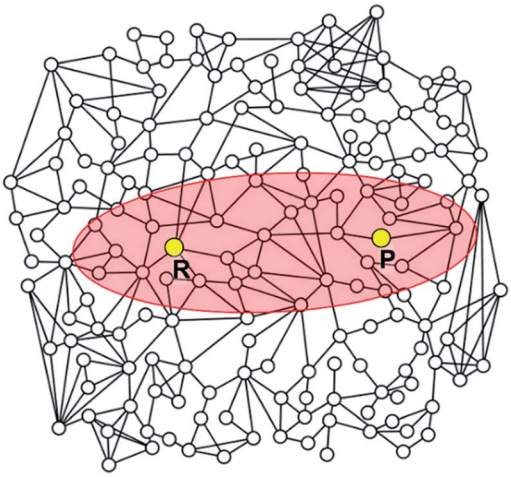


Fig. 5 Schematic illustration of a reaction network. The elliptic region shaded in red includes intermediates satisfying the criterion defined by eqn (5). Intermediates located outside the ellipse will not be included in the construction of the reaction network.

applied. That is, only unimolecular or bimolecular reactions are allowed, and the numbers of bond dissociation/formation should be smaller than or equal to predefined maximum values. In most cases, the maximum value is set to two as in eqn (3), but it is kept as small as possible to reduce computational costs. For catalytic reactions, elementary reactions not involving catalysts are not favored and hence are ignored in the network.

Extraction of the minimal reaction network. The distance between two vertices, *i.e.*, the length of an edge, should be related to the activation energy of the corresponding elementary reaction step. The activation energy can be calculated by using conventional quantum chemical methods. However, large computational costs are inevitable to deal with a number of elementary reactions. We note that it would be sufficient to use the CD given by eqn (4) as the distance between vertices for the purpose of first screening. According to the principle of minimum structure change,[68,69] the more the molecular structure changes, the higher the activation energy. Based on this idea, we first obtain the shortest reaction path passing through a specific reaction intermediate including all equidistant ones using the Dijkstra and Yen algorithms.[71,72] The same procedure is repeated for all intermediates in the reaction network, resulting in various reaction paths. Subsequently, all edges not belonging to the sampled paths are disconnected. If the network is decomposed into several subnetworks fully disconnected with each other, only the one containing both reactants and products is regarded as the minimal reaction network that can be determined without using quantum chemical calculations, while all the others are discarded. If the shortest paths are not sufficient, it is straightforward to extend to the second and the third shortest paths. However, this extension does not necessarily increase the network size because they share many vertices and edges with each other.

3. Step 4: kinetic analysis of reaction network

The minimal reaction network may have a few tens of reaction paths or more. To find the true minimum energy path out of them, we apply conventional transition state search algorithms to them. At this stage, we are able to take into account all possible molecular conformers or stereoisomers corresponding to a given vertex as described above. Each of them is subjected to the transition state search algorithm using DFT. To minimize computational load, we first consider most frequently appeared edges in the reaction paths. If the activation energy of an edge is above a given threshold, all the reaction paths including the edge are removed from the network. Also, the energy cutoff based on the DFT results can be applied to all intermediates. The remaining ones are considered as the most favorable reaction paths.

All the above procedures were implemented in our code, namely ACE-reaction, using Python 2.7,[73] with NumPy and SciPy,[74] and OpenOpt package[75] as the MILP solver. Our code can be combined with any structure-conversion and electronic structure calculation program. At present, we used the Pybel[76,77] for structure conversion, and DFTB+,[78] or GAUSSIAN 09 packages[79] for electronic structure calculations.

Computational details

We applied our method to two reaction examples: Claisen condensation and cobalt-catalyzed hydroformylation. For the 3D structure conversion of AC matrices sampled in the Claisen reaction, we employed the PM6 semiempirical method[80] with ethanol solvent described by the CPCM solvation model[81] as implemented in GAUSSIAN 09.[79] The density functional tight binding (DFTB) method[78] was used in the hydroformylation reaction with the trans3d-0-1,[82] and mio-1-1,[83] pairwise potential parameters. For DFTB calculations, the maximum numbers of cycles for self-consistent charge (SCC) and geometry relaxation were 500 and 10 000, respectively. The SCC tolerance was set to 10^{-5}, and the maximum force value for the geometry optimization was 10^{-3} Hartree Bohr^{-1}. The energy profiles of reaction paths for the hydroformylation were further investigated by employing M06 hybrid functional[84,85] with the 6-311++g(d,p) basis set, as implemented in GAUSSIAN 09.[79]

Results and discussion

1. Claisen ester condensation

Claisen ester condensation is a C–C coupling reaction between two ester molecules in the presence of strong bases. It has been widely utilized in total synthesis and biosynthesis.[86–90] We applied our method to this reaction to test whether or not it is able to find the accepted reaction mechanism. The input parameters and prediction results are summarized in Table 1. In Step 1, we assigned active atoms as shown in Fig. 6a. In Step 2, the combinatorial generation gave 113 intermediates, and 66 were left after screening with the energy tolerance of 20 kcal mol^{-1}. In Step 3, they were further screened out by the geometric analysis using eqn (5), resulting in only 32

Table 1 Input parameters and prediction results

	Step 2		Step 3		
Reaction	E_{tol} (kcal mol^{-1})	No. intermediates[a]	Δ	(N_V, N_E)[b]	Calculation time[c]
Claisen ester condensation	20.0	113 → 66	6	(32, 376) → (14, 35)	55 m 3 s (53 m 57 s + 1 m 6 s)[d]
Cobalt-catalyzed hydroformylation	20.0	239 → 224	6	(54, 403) → (39, 104)	56 m 2 s (53 m 35 s + 2 m 27 s)[d]

[a] Before and after screening with E_{tol}. [b] Number of vertices (N_V) and edges (N_E) after extraction of minimal reaction network. [c] Intel(R) Xeon(R) CPU E5-2690 v2@2.90 GHz (16 cores). [d] Time taken in Step 2 + Step 3.

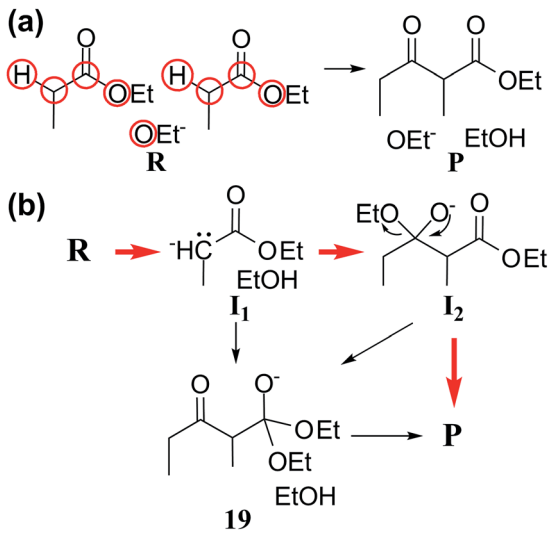


Fig. 6 (a) Reactants and products of Claisen ester condensation. The red circles indicate active atoms. (b) Three representative paths predicted by our method. Red arrows indicate the accepted mechanism.

intermediates. Subsequently, they were connected with each other according to the criterion in eqn (3), leading to a reaction network with 376 edges as shown in Fig. 7. At this stage, the number of intermediates is small enough to handle with quantum chemical methods, but the number of elementary reactions is relatively too large to perform accurate transition state calculations. Thus, we need to further rule out less favorable elementary reactions. We extracted the minimal reaction network composed of the paths within the top 50% in terms of CD using the Dijkstra and Yen algorithms.[71,72] As a result, 29 paths were obtained only with 14 vertices and 35 edges (circles and solid lines in Fig. 7).

Fig. 6b shows three representative paths; the other 26 paths are given in the ESI.† We note that the generally accepted mechanism in organic chemistry ($R \rightarrow I_1 \rightarrow I_2 \rightarrow P$) was included in the minimal reaction network. Surprisingly, this path also corresponds to the first shortest path in terms of CD. The overall process took about 55 minutes on a single workstation (Table 1).

2. Cobalt-catalyzed hydroformylation

As the second example, we chose the $HCo(CO)_3$-catalyzed hydroformylation whose mechanism has been proposed by Heck and Breslow,[91] because it is a relatively simple organometallic reaction and thus has been widely studied by other automated prediction methods.[10,26,52,53] In Step 1, we assigned active atoms as illustrated in Fig. 3a. Table 1 summarizes the input parameters and prediction results. Unlike the first example, we had 224 intermediates in Step 2. Here, screening by the energy cutoff was so ineffective that only 15 intermediates have been ruled out. In Step 3, however, filtering with eqn (5) drastically reduced the number of intermediates, leading to a reaction network with 54 vertices and 403 edges as shown in Fig. 8. This indicates that the geometric analysis illustrated in Fig. 5 is practically essential to extract a reaction network with a tractable size from complicated reactions. Then, we further reduced the size of the network so as to contain only the top 50% paths in terms of CD using the Dijkstra algorithm.[71] The resulting network was composed of 91 paths including 39 vertices and 104 edges (circles and solid lines in Fig. 8). The original Heck–Breslow mechanism[91] appeared within the top 33%. The other paths in the top 50% are given in the ESI.† It should be noted that we were able to arrive at this small network without quantum calculations at the first principle level, so that

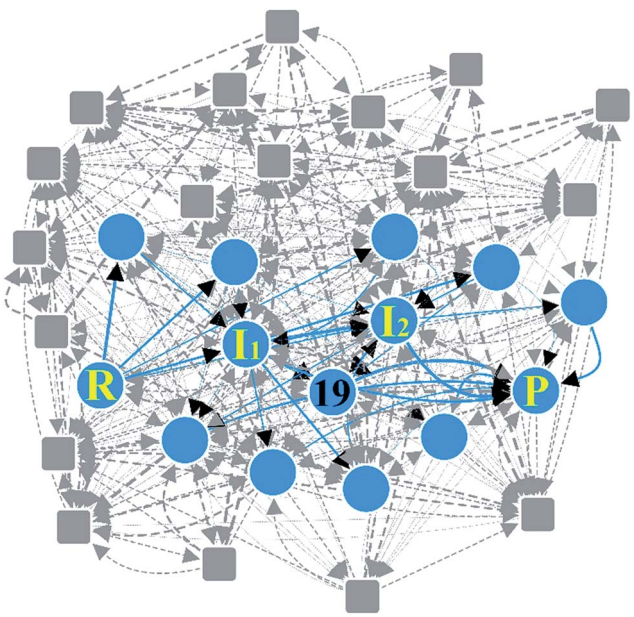


Fig. 7 Reaction network of Claisen ester condensation. The circles and solid lines indicate the vertices and edges in the minimal subnetwork obtained in Step 3, respectively. All molecular structures and chemical distance values in the network are available in the ESI.†

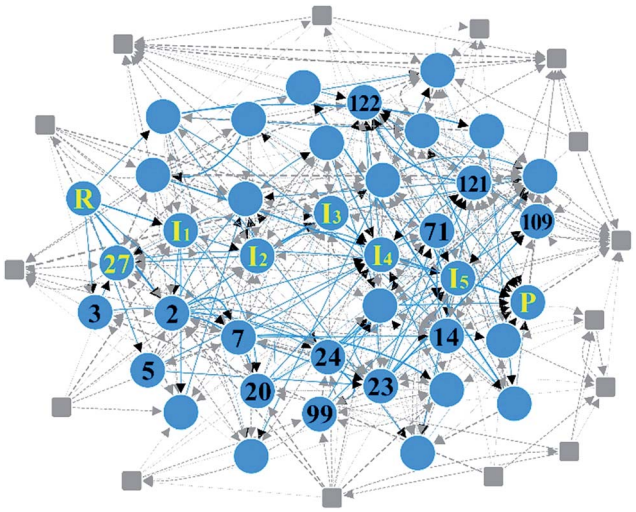


Fig. 8 Reaction network of cobalt-catalyzed hydroformylation. The circles and solid lines indicate the vertices and edges in the minimal subnetwork obtained in Step 3, respectively. All molecular structures and chemical distance values in the network are available in the ESI.†

the whole procedure took only around 56 minutes (Table 1) on a single workstation. Most of the time was used for the on-the-fly 3D geometry optimization of all intermediates using DFTB in Step 2. We note that the intermediates in the blue circle of Fig. 8 were also discovered by the automated prediction method in ref. 53. However, it obtained those after 45 simulations where each simulation took about 12 hours in the molecular dynamics sampling of reaction paths.

In Step 4, we performed DFT calculations for the 39 vertices in the minimal reaction network. Based on the DFT results, we removed vertices and edges using the following two criteria before transition state calculations: energy tolerance (E_{tol}) of 20 kcal mol^{-1} for intermediates and endothermic reactions with energy difference of 20 kcal mol^{-1} for vertices (those values can be changed according to reaction conditions). Isolated vertices in the network after applying the two criteria were also discarded. As a result, only 29 vertices and 74 edges were left. They were subjected to transition state calculations using DFT at the experimental temperature (403.15 K) and pressure (200 atm). At this step, we considered all possible conformers for each intermediate. Fig. 9 displays the final reaction network obtained from the DFT study. The numbers in the circles and the rectangles denote the relative energies of intermediates and transition states with respect to that of the reactants, respectively. The Cartesian coordinates of all the molecules and transition states optimized at the DFT level are available in the ESI.† Indeed, the Heck–Breslow mechanism (yellow circles

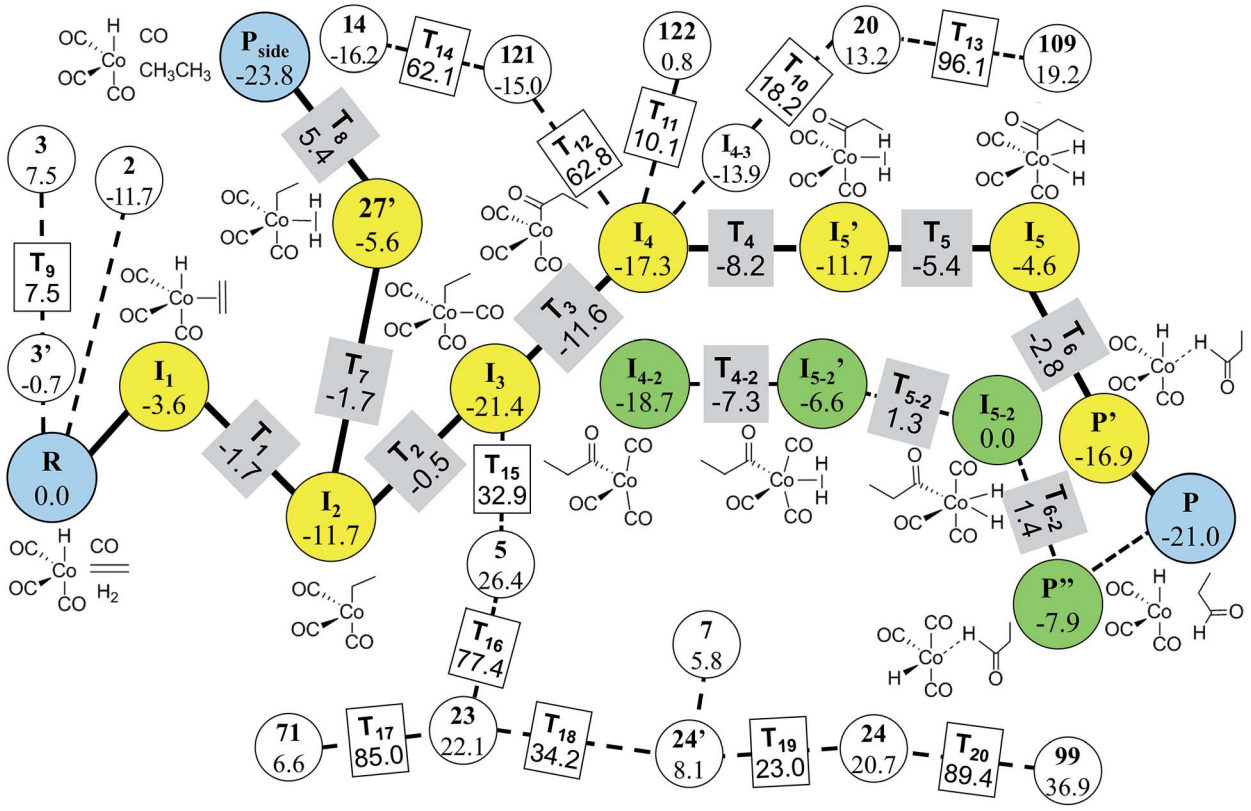


Fig. 9 Final reaction network for the hydroformylation reaction obtained at the DFT level. The circles indicate the reactants, products, and intermediates, while the rectangles indicate the transition states. The thick lines and the yellow circles denote the Heck–Breslow mechanism to give the product (*P*) and the hydrogenation mechanism to form the side product (P_{side}). The other paths are drawn with dashed lines. The green circles indicate the stereoisomers of the intermediates in the Heck–Breslow mechanism. The numbers in the circles and the rectangles denote the relative energies of intermediates and transition states with respect to that of the reactant (*R*), respectively. All the energy values are in kcal mol^{-1}.

connecting ***R*** and ***P*** in Fig. 9) was turned out to be kinetically the most favorable; the activation energy of the rate-determining step $I_2 \rightarrow I_3$ is 11.2 kcal mol^{-1}. We were able to find a path involving conformers or stereoisomers such as I_{4-2} and I_{5-2} (the green circles in Fig. 9), which was also reported in ref. 53. However, this path was not directly reachable from the reactants. In addition, the hydrogenation path as a well-known side reaction was also included there (the yellow circles connecting ***R*** and P_{side} in Fig. 9).[26,92] P_{side} was also sampled by our method (vertex **1**, see the ESI†), but was screened in Step 3. Nonetheless, it was readily derived from the intermediate **27′** in DFT calculations. The path linking to the intermediate **14** is also related to the hydrogenation mechanism, but it is kinetically unfavorable due to the very high barrier of over 60 kcal mol^{-1}.

It is emphasized that our method is much more efficient than the previous approaches. For the same hydroformylation reaction, we performed only 36 transition state searches including failed ones at the DFT level, whereas Maeda's approach performed 2266 Hessian calculations[10] and Varela's method dealt with 448 elementary reactions.[26] However, our approach has some limitations; intermediates such as I_5' and **27′** obtained from the DFT calculations did not appear in the combinatorial generation of intermediates because they have unusual chemical bonds such as H with two single bonds. It is possible to sample them by modifying the screening criteria, but then it will produce a lot of undesired molecules. Consequently, final reaction paths predicted from the graph-theoretic method should be refined through accurate quantum calculations.

Conclusions

We developed an efficient graph-theoretic method for the automated prediction of reaction mechanism. It is based on the fact that chemical reactions can be described by the successive changes in chemical bonds. Reactant molecules can be represented with atom connectivity (AC) matrices. Then, hypothetical intermediates can be sampled through the combinatorial enumeration of the matrix. Among them, chemically inappropriate AC matrices are discarded by on-the-fly 3D structure conversion and pruning criteria. In addition, the geometric analysis based on a chemical distance concept is used to further screen out intermediates whose structures are substantially different from reactants and products. The remaining molecules are regarded as vertices and connected to build a reaction network. The key feature of our method is to extract a minimal subnetwork from a very complex full network. To this end, we explore the reaction pathways connecting reactants and products with minimum dissociation and formation of chemical bonds for all intermediates in the network using the Dijkstra algorithm. Subsequently, they are subjected to accurate transition state calculations for refinements. It should be emphasized that though our method relies on chemical heuristics, the rules imposed for enumeration and screening are not reaction specific. Therefore, it can be applied to a wide range of chemical reactions.

The efficiency and reliability of our method have been assessed by applying it to two example reactions. It was able to successfully predict the accepted reaction mechanism of Claisen ester condensation. Also, it could find not only the original Heck–Breslow mechanism, but also the hydrogenation of ethylene as a side reaction for cobalt-catalyzed hydroformylation, showing its potential applicability to organometallic or inorganic reactions. It is remarkable that for both examples, our method completed the whole process, except for DFT calculations, within only an hour on a single workstation with 16 Intel Xeon cores.

The present work offers an efficient approach to predict reaction pathways. However, the following issues need to be addressed to further improve its reliability and applicability. First of all, molecular graphs have clear limitations to discriminate different electronic states of molecules with an identical AC. For instance, one can recall pre-reaction and post-reaction complexes, ion pairs, charge-transfer complexes, and many other types of spatial configurations of nuclei, which correspond to deep local minima in potential energy surfaces. Second, more rigorous catalytic effects need to be included throughout the process from combinatorial generation to distance evaluation. Chemical bonds between organic substrates and metal catalysts are often ill-defined. As a result, it is difficult to apply simple chemical rules such as atomic valences and formal charges to the combinatorial enumeration step. As future works, we expect that a novel combination of heuristic rules, first principles theory, and machine learning techniques will be a key to resolve the aforementioned problems.

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

This work was supported by Basic Science Research Programs (NRF-2015R1A1A1A05001480) funded by the Korea government [MSIP] and by KISTI supercomputing center through the strategic support program (No. KSC-2016-C2-0047).

Notes and references

1 S. Niu and M. B. Hall, *Chem. Rev.*, 2000, **100**, 353.
2 K. N. Houk and P. H. Cheong, *Nature*, 2008, **455**, 309.
3 P. H. Cheong, C. Y. Legault, J. M. Um, N. Çelebi-Ölçüm and K. N. Houk, *Chem. Rev.*, 2011, **111**, 5042.
4 C. Levinthal, *J. Chim. Phys. Phys.-Chim. Biol.*, 1968, **65**, 44.
5 C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House*, ed. J. T. P. DeBrunner and E. Munck, University of Illinois Press, Monticello, IL, 1969, p. 22.
6 S. Goedecker, *J. Chem. Phys.*, 2004, **120**, 9911.
7 S. Maeda and K. Ohno, *J. Phys. Chem. A*, 2005, **109**, 5742.
8 K. Ohno, N. Kishimoto, T. Iwamoto and H. Satoh, *J. Comput. Chem.*, 2017, **38**, 669.

9 S. Maeda and K. Morokuma, *J. Chem. Theory Comput.*, 2011, **7**, 2335.
10 S. Maeda and K. Morokuma, *J. Chem. Theory Comput.*, 2012, **8**, 380.
11 W. M. C. Sameera, S. Maeda and K. Morokuma, *Acc. Chem. Res.*, 2016, **49**, 763.
12 L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande and T. J. Martínez, *Nat. Chem.*, 2014, **6**, 1044.
13 I. Berente and G. Náray-Szabó, *J. Phys. Chem. A*, 2006, **110**, 772.
14 C. Shang and Z.-P. Liu, *J. Chem. Theory Comput.*, 2012, **8**, 2215.
15 C. Shang and Z.-P. Liu, *J. Chem. Theory Comput.*, 2013, **9**, 1838.
16 T. Lankau and C.-H. Yu, *J. Chem. Phys.*, 2013, **138**, 214102.
17 D. J. Wales, *Science*, 1999, **285**, 1368.
18 D. J. Wales, *Phys. Biol.*, 2005, **2**, S86.
19 D. J. Wales and T. V. Bogdan, *J. Phys. Chem. B*, 2006, **110**, 20765.
20 S. M. Woodley and R. Catlow, *Nat. Mater.*, 2008, **7**, 937.
21 Y. Kim, S. Choi and W. Y. Kim, *J. Chem. Theory Comput.*, 2014, **10**, 2419.
22 M. Bergeler, G. N. Simm, J. Proppe and M. Reiher, *J. Chem. Theory Comput.*, 2015, **11**, 5712.
23 G. N. Simm and M. Reiher, *J. Chem. Theory Comput.*, 2017, **13**, 6108.
24 J. Proppe, T. Husch, G. N. Simm and M. Reiher, *Faraday Discuss.*, 2016, **195**, 497.
25 E. Martínez-Núñez, *J. Comput. Chem.*, 2015, **36**, 222.
26 J. A. Varela, S. A. Vázquez and E. Martínez-Núñez, *Chem. Sci.*, 2017, **8**, 3843.
27 J. Dugundji and I. Ugi, *Top. Curr. Chem.*, 1973, **39**, 19.
28 I. Ugi, J. Bauer, K. Bley, A. Dengler, A. Dietz, E. Fontain, B. Gruber, R. Herges, M. Knauer, K. Reitsam and N. Stein, *Angew. Chem., Int. Ed. Engl.*, 1993, **32**, 201.
29 O. N. Temkin, A. V. Zeigarnik and D. Bonchev, *Chemical Reaction Networks*, CRC Press, New York, 1996.
30 E. J. Corey and W. L. Jorgensen, *J. Am. Chem. Soc.*, 1976, **98**, 189.
31 D. A. Pensak and E. J. Corey, in *Computer-Assisted Organic Synthesis*, ACS, Washington, US, 1977, vol. 61, ch. 1, pp. 1–32.
32 J. Gasteiger and W. D. Ihlenfeldt, in *Software Development in Chemistry 4*, ed. P. D. J. Gasteiger, Springer, Berlin, Heidelberg, 1990, pp. 57–65.
33 C. Rücker, G. Rücker and S. H. Bertz, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 378.
34 M. H. Todd, *Chem. Soc. Rev.*, 2005, **34**, 247.
35 B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk and C. E. Wilmer, *Nat. Chem.*, 2009, **1**, 31.
36 M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski and K. J. M. Bishop, *Angew. Chem., Int. Ed.*, 2012, **51**, 7928.
37 J. H. Chen and P. Baldi, *J. Chem. Inf. Model.*, 2009, **49**, 2034.
38 P. E. Fuller, C. M. Gothard, N. A. Gothard, A. Weckiewicz and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2012, **51**, 7933.
39 J. H. Chen and P. Baldi, *J. Chem. Educ.*, 2008, **85**, 1699.
40 N. Graulich, H. Hopf and P. R. Schreiner, *Chem. Soc. Rev.*, 2010, **39**, 1503.
41 M. A. Kayala, C.-A. Azencott, J. H. Chen and P. Baldi, *J. Chem. Inf. Model.*, 2011, **51**, 2209.
42 M. A. Kayala and P. Baldi, *J. Chem. Inf. Model.*, 2012, **52**, 2526.
43 A. E. Clark, in *Annual Reports in Computational Chemistry*, Elsevier, New York, 2015, vol. 11, ch. 6, pp. 326–359.
44 R. García-Domenech, J. Galvez, J. V. de Julian-Ortiz and L. Pogliani, *Chem. Rev.*, 2008, **108**, 1127.
45 A. T. Balaban, *J. Chem. Inf. Model.*, 1985, **25**, 334.
46 F. Harary, in *Graph Theory*, Addison-Wesley, Reading, MA, 1969, pp. 185–187.
47 G. Pólya and R. C. Reed, *Combinatorial Enumeration of Groups, Graphs and Chemical Compounds*, Springer-Verlag, New York, 1987.
48 N. S. Zefirov and S. S. Tratch, *Anal. Chim. Acta*, 1990, **235**, 115.
49 R. Herges, *J. Chem. Inf. Model.*, 1990, **30**, 377.
50 N. S. Zefirov, I. I. Baskin and V. A. Palyulin, *J. Chem. Inf. Model.*, 1994, **34**, 994.
51 M. H. S. Segler and M. P. Waller, *Chem.–Eur. J.*, 2017, **23**, 6118.
52 S. Habershon, *J. Chem. Phys.*, 2015, **143**, 094106.
53 S. Habershon, *J. Chem. Theory Comput.*, 2016, **12**, 1786.
54 Y. V. Suleimanov and W. H. Green, *J. Chem. Theory Comput.*, 2015, **11**, 4248.
55 P. M. Zimmerman, *J. Comput. Chem.*, 2013, **34**, 1385.
56 P. Zimmerman, *J. Chem. Theory Comput.*, 2013, **9**, 3043.
57 P. M. Zimmerman, *Mol. Simul.*, 2015, **41**, 43.
58 A. J. Nett, W. Zhao, P. M. Zimmerman and J. Montgomery, *J. Am. Chem. Soc.*, 2015, **137**, 7636.
59 D. Rappoport, C. J. Galvin, D. Y. Zubarev and A. Aspuru-Guzik, *J. Chem. Theory Comput.*, 2014, **10**, 897.
60 D. Y. Zubarev, D. Rappoport and A. Aspuru-Guzik, *Sci. Rep.*, 2015, **5**, 8009.
61 C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Comput. Phys. Commun.*, 2016, **203**, 212.
62 H. B. Schlegel, *J. Comput. Chem.*, 1982, **3**, 214.
63 H. B. Schlegel, *Theor. Chim. Acta*, 1984, **66**, 333.
64 C. Peng, P. Y. Ayala, H. B. Schlegel and M. J. Frisch, *J. Comput. Chem.*, 1996, **17**, 49.
65 G. S. Hammond, *J. Am. Chem. Soc.*, 1955, **77**, 334.
66 Y. Kim and W. Y. Kim, *Bull. Korean Chem. Soc.*, 2015, **36**, 1769.
67 H. Kim, Y. Kim, J. Kim and W. Y. Kim, *Carbon*, 2016, **98**, 404.
68 C. Jochum, J. Gasteiger and I. Ugi, *Angew. Chem., Int. Ed. Engl.*, 1980, **19**, 495.
69 C. Jochum, J. Gasteiger, I. Ugi and J. Dugundji, *Z. Naturforsch., A: Phys. Sci.*, 1982, **37b**, 1205.
70 E. L. First, C. E. Gounaris and C. A. Floudas, *J. Chem. Inf. Model.*, 2012, **52**, 84.
71 E. W. Dijkstra, *Numer. Math.*, 1959, **1**, 269.
72 J. Y. Yen, *Manage. Sci.*, 1971, **17**, 712.
73 G. Rossum, *Python reference manual*, CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands, 1995.
74 T. E. Oliphant, *Comput. Sci. Eng.*, 2007, **9**, 10.

75 D. Kroshko, *OpenOpt: Free scientific-engineering software for mathematical modeling and optimization, version 0.5610*, 2015, accessed July 2015, http://www.openopt.org.
76 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.
77 N. M. O'Boyle, C. Morley and G. R. Hutchison, *Chem. Cent. J.*, 2008, **2**, 5.
78 B. Aradi, B. Hourahine and T. Frauenheim, *J. Phys. Chem. A*, 2007, **111**, 5678.
79 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, *et al.*, *Gaussian 09 revision D.01*, Gaussian Inc., Wallingford, CT, 2009.
80 J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173.
81 V. Barone and M. Cossi, *J. Phys. Chem. A*, 1998, **102**, 1995.
82 G. Zheng, H. a. Witek, P. Bobadova-Parvanova, S. Irle, D. G. Musaev, R. Prabhakar, K. Morokuma, M. Lundberg, M. Elstner, C. Köhler and T. Frauenheim, *J. Chem. Theory Comput.*, 2007, **3**, 1349.
83 M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert, *Phys. Rev. B: Condens. Matter*, 1998, **58**, 7260.
84 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215.
85 Y. Zhao and D. G. Truhlar, *Acc. Chem. Res.*, 2008, **41**, 157.
86 T. Laue and A. Plagens, *Named Organic Reactions*, John Wiley & Sons, New York, 2nd edn, 2005.
87 R. J. Heath and C. O. Rock, *Nat. Prod. Rep.*, 2002, **19**, 581.
88 C. Gregg and M. V. Perkins, *Org. Biomol. Chem.*, 2012, **10**, 6547.
89 D. E. Ward, D. Kundu, M. Biniaz and S. Jana, *J. Org. Chem.*, 2014, **79**, 6868.
90 S. J. Leiris, O. M. Khdour, Z. J. Segerman, K. S. Tsosie, J.-C. Chapuis and S. M. Hecht, *Bioorg. Med. Chem.*, 2010, **18**, 3481.
91 R. F. Heck and D. S. Breslow, *J. Am. Chem. Soc.*, 1961, **83**, 4023.
92 L. E. Rush, P. G. Pringle and J. N. Harvey, *Angew. Chem., Int. Ed.*, 2014, **53**, 8672.