Supplementary Information

Leveraging Algorithmic Search in Quantum Chemical Reaction Path Finding

Atsuyuki Nakao,[*a] Yu Harabuchi,[bcd] Satoshi Maeda,[bcd] and Koji Tsuda[aef]

[a] Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 2778561, Japan. E-mail: tsuda@k.u-tokyo.ac.jp

[b] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Sapporo 001-0021, Japan.

[c] JST ERATO Maeda Artificial Intelligence for Chemical Reaction Design and Discovery Project, Sapporo 060-0810, Japan

[d] Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo 060- 0810, Japan

[e] RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan

[f] Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba 305-0047, Japan

1. Details of similarity measure:

In the SC-AFIR search, the present RRT selects a node for the expansion based on the structure-similarity defined between a node and product (goal) geometries. The actual computation of structure-similarity requires three conditions, as below.

i) Structure-similarity can be defined between a pair of structures consisting of different numbers of atoms.

ii) Structure-similarity does not depend on the rotation, translation, and atom permutation of the system.

iii) When there are multiple molecules within the structure. Structure-similarity does not depend on the relative position and molecular rotation of the molecules.

The first condition is important when there are additional molecules, such as H_2O, as shown in **Figure SI1**. In this case, users can specify the product molecules without considering geometry deformations of the solvent molecules, i.e., two H_2O molecules. The third condition is important when the reaction leads to multiple products such as X→Y+Z. In this case, structure-similarity is measured not depending on the relative position and molecular rotation of the molecular fragments, Y and Z. To achieve these conditions, we defined the structure-similarity by matching atom pairs within a pair of structures. **Figure SI1a** and **SI1b** indicate two cases for an EQ close to the goal and an EQ far from the goal, respectively. The structure-similarity is computed as the sum of atom-similarity over all the greedily matched atom pairs (indicated by red numbers in **Figure SI1a** and **SI1b**). The atom-similarity

for an atom pair is defined by using the interatomic distance, which does not depend on the translation and rotation of a structure. The definition of atom-similarity and structure-similarity is described below.

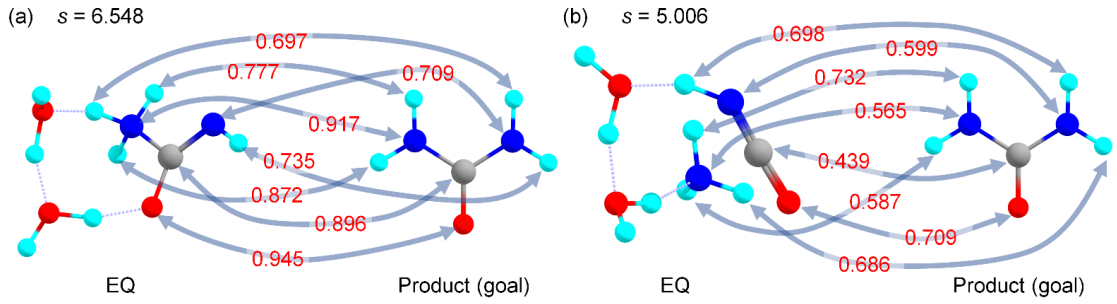

Figure SI1. Examples of greedy atom matching procedure. (a) is for an EQ close to the goal and (b) is for an EQ far from the goal. Lines indicate the matched atom pairs, and atom-similarity of the corresponding atom pairs values are indicated by red. Structure-similarity, *s*, is also indicated.

At first, a descriptor of an atomic environment is defined to achieve the matching of atom pairs, shown by blue allows in **Figure SI1a** and **SI1b**. The descriptor of atom a is the following double list constructed by interatomic distances:

$$a = [\boldsymbol{r}_x, \boldsymbol{r}_y, \dots,], \tag{1}$$

$$\boldsymbol{r}_x = [r_{ax^1}, \dots, r_{ax^n}], \tag{2}$$

where $\boldsymbol{r}_x$ is a list of distances from atom a to atoms of a chemical element x, r_{ax^i} is the distance to ith x, and n is the number of x.

The procedure calculating the atom-similarity between a pair of atomic environments is as follows.

1. A comparison matrix C_x is constructed for each x. Each matrix element is calculated as the following equation:
$$C_{x,ij}(a, a') = |r_{ax^i} - r_{a'x^j}|. \tag{3}$$
2. The minimum element in C_x is appended to a matched value list $\boldsymbol{v}_x$.
3. A row and a column corresponding to the selected element are removed from C_x.
4. 2 and 3 are repeated until all elements are removed from C_x.
5. Matched value lists $\boldsymbol{v}_x$ of all-atom kinds are created by repeating 1-4.
6. Atom-similarity between atomic environments $s_a(a, a')$ is calculated based on the matched value lists $\boldsymbol{v}_x$ as the following equation:
$$s_a(a, a') = \frac{1}{n} \sum_{x \in \boldsymbol{\pi}} \sum_{v_{x,i} \in \boldsymbol{v}_x} \exp(-\theta v_{x,i}), \tag{4}$$

where $\boldsymbol{\pi}$ is a list of chemical elements included in structures, n is the minimum number of atoms among comparing structures, and θ is a hyperparameter adjusting rigorousness of substructure

matching. It is noted that this atom-similarity becomes 1 when the geometries around an atom are the same completely.

The structure-similarity between a pair of structures is computed by summing up the atom-similarity. The atom pairs are defined by a greedy matching procedure based on atom-similarity. Examples of atom matchings are shown in **Figure SI1**. The procedure is as follows:

1. A similarity matrix S_x is constructed for each x. A matrix element is
$$S_{x,ij} = s_a(a_{x^i}, a'_{x^j}), \quad (5)$$
where a_{x^i}, a'_{x^j} are atomic environment descriptors of ith and jth atom x in each structure. If there are multiple target structures, all atoms of all target structures are compared at once in S_x.
2. The maximum element is selected and corresponding atoms are paired. The atom-similarity is appended to a matched similarity list $\boldsymbol{u}_x$.
3. A row and a column corresponding to the selected element are removed from S_x.
4. 2 and 3 are repeated until all elements are removed from S_x.
5. Matched similarity lists for all x are created by repeating 1-4.
6. A similarity between structure m and m' is calculated as the following equation:
$$s(m, m') = \sum_{x \in \boldsymbol{\pi}} \sum_{u_{x,i} \in \boldsymbol{u}_x} u_i. \quad (6)$$

A range of the structure-similarity is from 0 to the minimum number of atoms in m and m'.

2. The AFIR method and the SC-AFIR search:

The artificial force induced reaction (AFIR) method induces a molecular geometry deformation by adding an artificial force corresponding to pushing or pulling apart fragments consisting of atoms. **Figure SI2a** shows the schematic picture of the AFIR function, F^{AFIR}, and the original potential energy surface (PES), $E(\mathbf{Q})$. Here, $\mathbf{Q}$ indicates the molecular geometry parameters, and equilibrium structure (EQ) corresponds to the energy minima on $E(\mathbf{Q})$. By adding the artificial force corresponding to the term $V(\mathbf{Q})$, an energy peak between EQ1 and EQ2 disappears, and minimization of F^{AFIR} induces the molecular geometry deformation from EQ 1 to EQ2. **Figure SI2b** depicts an example of the artificial-force direction between two fragments. In this case, NH_3 and HNCO molecules are connected along the path, and a different single molecule is generated. In the graph network search, this procedure to search for EQ2 corresponds to the expansion of a new node (EQ2) from the selected node (EQ1). The SC algorism generates many fragment pairs by focusing on two atoms within a molecular structure for each EQ. The combinations of these fragment pairs and the directions of force (pushing or pulling apart) are listed as a task list for each EQ. When an EQ is selected during the search, the expansion of the reaction path network is done by applying AFIR minimizations based on the task list of the corresponding EQ.

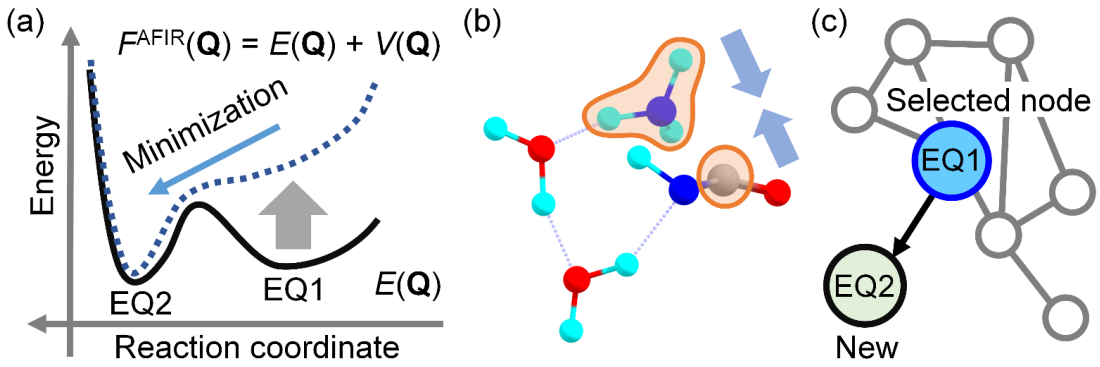

Figure SI2. Schematic pictures of the AFIR search. (a) is for the AFIR function, (b) is for the direction of artificial force between the fragment pairs, and (c) is for the expansion of the reaction path network by a single AFIR calculation in the SC-AFIR search.