# Theoretical chemical reaction database construction based on quantum chemistry-aided retrosynthetic analysis

Yu Harabuchi[a,b,c] and Satoshi Maeda[a,b,c,d,*]

[a]*Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0021, Japan.*

[b]*JST, ERATO Maeda Artificial Intelligence in Chemical Reaction Design and Discovery Project, Kita 10, Nishi 8, Kita-ku, Sapporo, Hokkaido 060-0810, Japan.*

[c]*Department of Chemistry, Faculty of Science, Hokkaido University, Kita 10, Nishi 8, Kita-ku, Sapporo, Hokkaido 060-0810, Japan.*

[d]*Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Tsukuba, Ibaraki 305-0044, Japan.*

*Corresponding author: E-mail: smaeda@eis.hokudai.ac.jp

**Abstract:** A theoretical database comprising experimentally accessible and inaccessible chemical reactions could complement the existing experimental databases and contribute significantly to data-driven chemical reaction discovery. Quantum chemistry-aided retrosynthetic analysis (QCaRA) can generate a network of elementary steps called a reaction-path network and predict hundreds or more of chemical reactions along with their theoretical yields. In contrast to ordinary simulations, QCaRA traces back the reaction paths from the target product to various reactant candidates while solving the kinetic equations. In this study, we propose theoretical reaction database construction based on QCaRA. Seven reaction-path networks containing 13,190 reactions, 108,754 reaction paths, and

2,552,652 geometries have been identified and discussed as examples. In addition to well-known reactions (i.e., synthesis of fluoroglycine, Wöhler's urea synthesis, base-catalysed aldol reaction, Lewis-acid-catalysed ene reaction, cobalt-catalysed hydroformylation, Strecker reaction, and Passerini reaction), numerous unexplored reactions with high, medium, low, near-zero, or zero yields have been identified. We anticipate that such a QCaRA-based theoretical reaction database will provide information on hitherto unexplored reactivities, especially those that are experimentally inaccessible.

## Introduction

Data-driven reaction discovery is a prominent field in modern chemistry. Experimental reaction databases have been widely used to develop efficient synthetic routes.[1–9] However, experimentally unexplored reactions and chemical transformations that have limitations such as reagent inaccessibility are not included in these databases. Supplying negative data also requires considerable experimental resources and effort. A theoretical reaction database is therefore required to complement these experimental databases and promote data-driven chemical reaction discovery.

The exploration of quantum chemical potential energy surfaces[10–15] can elucidate reaction mechanisms and thus enables the prediction of reactions. Quantum chemical calculations can reveal the molecular interactions between reagents and probable intermediates and elementary steps in a reaction.[16–25] Various automatic elementary step searching tools such as the freezing string method coupled with the Berny algorithm,[26] single/double-ended growing string methods,[27] nanoreactor,[28] artificial force induced reaction (AFIR) method,[29–31] reaction mechanism generator,[32,33] and Kinbot[34] have been reported in the literature. Moreover, theoretical databases of elementary steps have been created using these or other reaction-path searching tools.[35–39]

In general, a vast network of elementary steps is required to represent a reaction, and

elucidating an entire network from the reactants to the probable major and minor products, along with numerous intermediates and elementary steps, is complex and time-consuming. Following the terminology for experimental databases, we have termed this entire network as a reaction in this article and an elementary step as a reaction path. Correspondingly, the network of reaction paths is called a reaction-path network. Generally, a reaction-path network is constructed first, and the corresponding reaction is then studied by kinetic simulations.[40–44] However, examining each reaction individually through this two-step procedure requires enormous effort. Therefore, creating a theoretical database of reactions rather than elementary steps remains a major challenge in chemistry.

In this study, we have used quantum chemistry-aided retrosynthetic analysis (QCaRA) which is a reaction discovery concept proposed nearly a decade ago to identify reactants for a reaction-path network starting from the target product.[17] However, until very recently, its application had been limited to elementary reactions[17,45,46] consisting only of a single step owing to the combinatorial explosion of reactant candidates. Recently, its applicability has been considerably expanded by combining a reaction-path network exploration engine with a kinetics-based navigation algorithm.[47] Although one could adopt any automatic reaction-path searching tool in the exploration engine, we employed the AFIR method. We also used the rate constant matrix contraction (RCMC) method as the kinetics-based navigation algorithm, which narrows down the reactant candidates by identifying and excluding those kinetically inaccessible for the product. The use of RCMC suppresses combinatorial explosion and enables the application of QCaRA to various reactions.

When we applied QCaRA to known chemical reactions, we observed that QCaRA in combination with AFIR and RCMC can identify not only known reactions but also hundreds of unexplored reactions that can afford the target product with finite theoretical yields.[47] Based on these results, we propose theoretical reaction database construction using QCaRA. As initial examples, seven reaction-path networks containing 13,190 reactions, 108,754 reaction paths, and 2,552,652

geometries have been prepared and discussed. The inputs for these QCaRA calculations are products of seven known reactions in **Figure 1**: synthesis of fluoroglycine,[45,46] Wöhler's urea synthesis,[48] base-catalysed aldol reaction,[49–51] Lewis acid-catalysed ene reaction,[52,53] cobalt-catalysed hydroformylation,[54–56] Strecker reaction,[57–59] and Passerini reaction.[58–60]

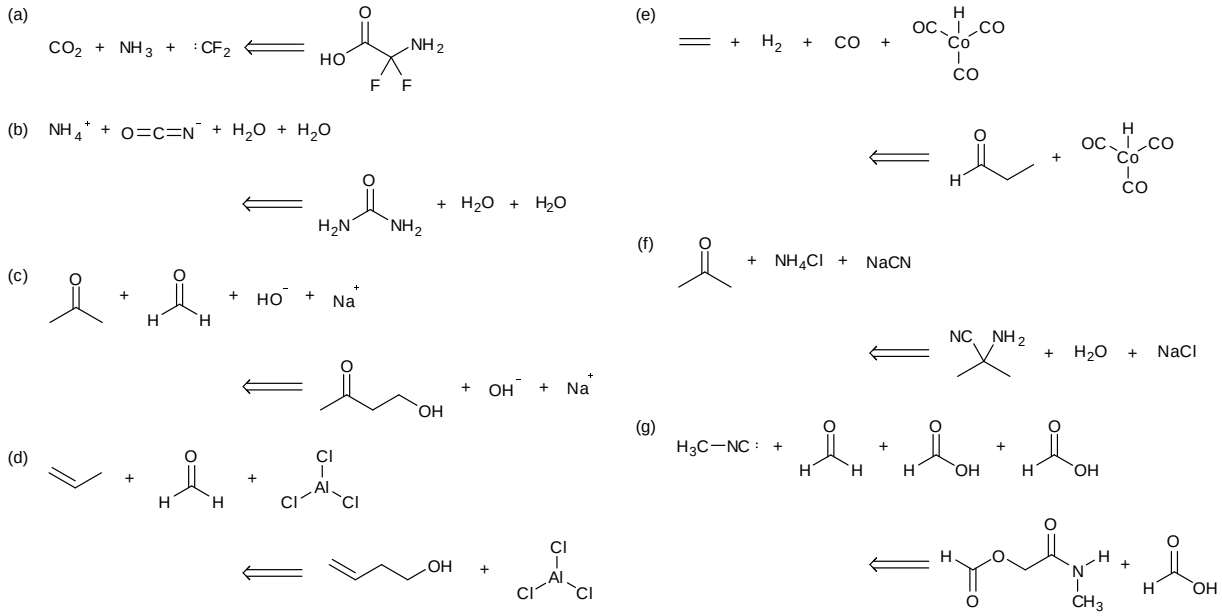


**Figure 1.** Input products of the seven reaction-path network calculations and the retrosynthesis arrows for transforming them into the respective known reactions: (a) synthesis of fluoroglycine, (b) Wöhler's urea synthesis, (c) base-catalysed aldol reaction, (d) Lewis-acid-catalysed ene reaction, (e) cobalt-catalysed hydroformylation, (f) Strecker reaction, and (g) Passerini reaction.

## Results and Discussion

### Reaction-path network data

The QCaRA data consist of nodes (equilibrium structures, EQs) and edges (a reaction path corresponding to an elementary step), as shown in **Figure 2**. QCaRA produced a significant amount of raw data, which contained three-dimensional molecular geometries, energies, molecular properties

of EQs and transition states (TSs) or path tops (PTs, structures with the maximum energy), all the discrete points along the reaction pathway, and meta-data of calculations such as creation dates.

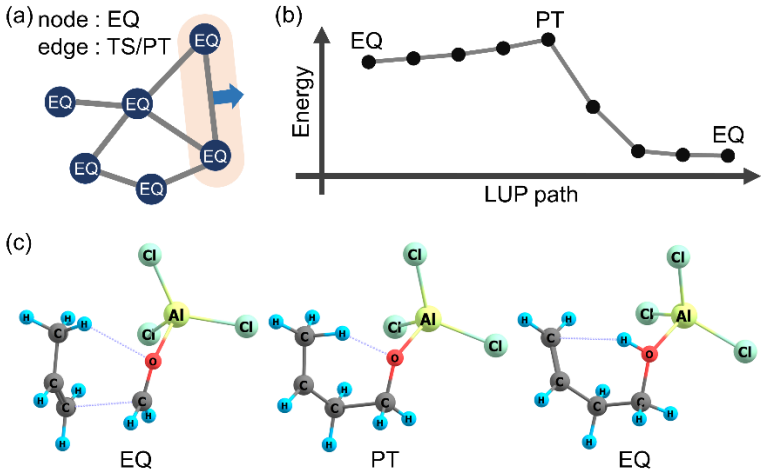


**Figure 2.** QCaRA data of a reaction-path network. (a) Graph network where nodes represent EQs and edges represent the corresponding TS/PT in the reaction. (b) Sequence of geometries and their energies along a reaction path. Each discrete point is a three-dimensional molecular geometry and its energy. (c) Example describing an LUP reaction path in the network for the Lewis-acid-catalysed ene reaction. Abbreviations: PT, path top; EQ, equilibrium structure; TS, transition state; LUP, locally updated planes.

These networks were constructed using a two-step procedure: (a) network exploration by AFIR and RCMC based on density functional theory (DFT) calculations with a small basis set and (b) network refinement by energy and gradient evaluations at all the discrete points along all paths obtained in Step (a) based on DFT calculations with a medium-sized basis set. For network refinement, Hessian calculations were also done at the path terminals (EQs) and PTs. Additionally, the structure geometries along all paths and all the energy, gradient, and Hessian data computed in Step (b) have been added to our database. **Table 1** lists the number of EQs and PTs, energy data, gradient data,

Hessian data, and the geometries available for each network. These data would also be useful in the training of machine learning potentials,[61–63] which could help accelerate molecular simulations and future QCaRA calculations.

**Table 1.** Data included in the seven reaction-path networks presented in this article.

| | entry 1[a] | entry 2[b] | entry 3[c] | entry 4[d] | entry 5[e] | entry 6[f] | entry 7[g] |
|---|---|---|---|---|---|---|---|
| Atom | 10 | 14 | 17 | 17 | 18 | 19 | 20 |
| Reaction[h] | 446 | 262 | 1087 | 2680 | 4629 | 1679 | 2407 |
| 0 – 1%[i] | 386 | 228 | 939 | 2281 | 3754 | 1512 | 1933 |
| 1 – 50%[i] | 6 | 5 | 33 | 43 | 20 | 28 | 57 |
| 50 – 100%[i] | 54 | 29 | 115 | 356 | 855 | 139 | 417 |
| EQ | 1765 | 1776 | 6199 | 8394 | 10810 | 9203 | 12215 |
| PT | 6526 | 11369 | 16669 | 17655 | 20177 | 18156 | 18202 |
| Gradient[j] | 100751 | 225522 | 330307 | 368421 | 402234 | 396148 | 396221 |
| Hessian[k] | 20136 | 36462 | 51594 | 53973 | 61311 | 54801 | 54771 |
| Geometry | 120887 | 261984 | 381901 | 422394 | 463545 | 450949 | 450992 |

[a]Network including synthesis of fluoroglycine. [b]Network including Wöhler's urea synthesis. [c]Network including base-catalysed aldol reaction. [d]Network including Lewis-acid-catalysed ene reaction. [e]Network including cobalt-catalysed hydroformylation, [f]Network including Strecker reaction, [g]Network including Passerini reaction, [h]Number of groups consisting of geometries with the same bonding pattern. [i]Number of groups that have the reaction yields of 0 – 1%, 1 – 50%, and 50 – 100% during simulation at 300 K. [j]Number of geometries on which a gradient vector was computed. [k]Number of geometries on which both a gradient vector and a Hessian matrix were computed.

**Seven reaction-path networks**

**Figures 3a−g** list the EQs of input products (left) and output reactants (right) for seven known reactions; the EQ ID has no significant meaning since it changes during network refinement. The electronic energies relative to the EQs of the input products are also shown. The theoretical product yields for a reaction starting from each EQ at 200 K, 300 K, and 400 K are presented below each EQ. In some cases, the theoretical yields for a reaction starting even from the input products are less than 100% owing to further transformation into EQs that are more stable at the corresponding

temperatures. QCaRA has been successful in predicting known reactions in all seven cases. The seven networks are described briefly below.

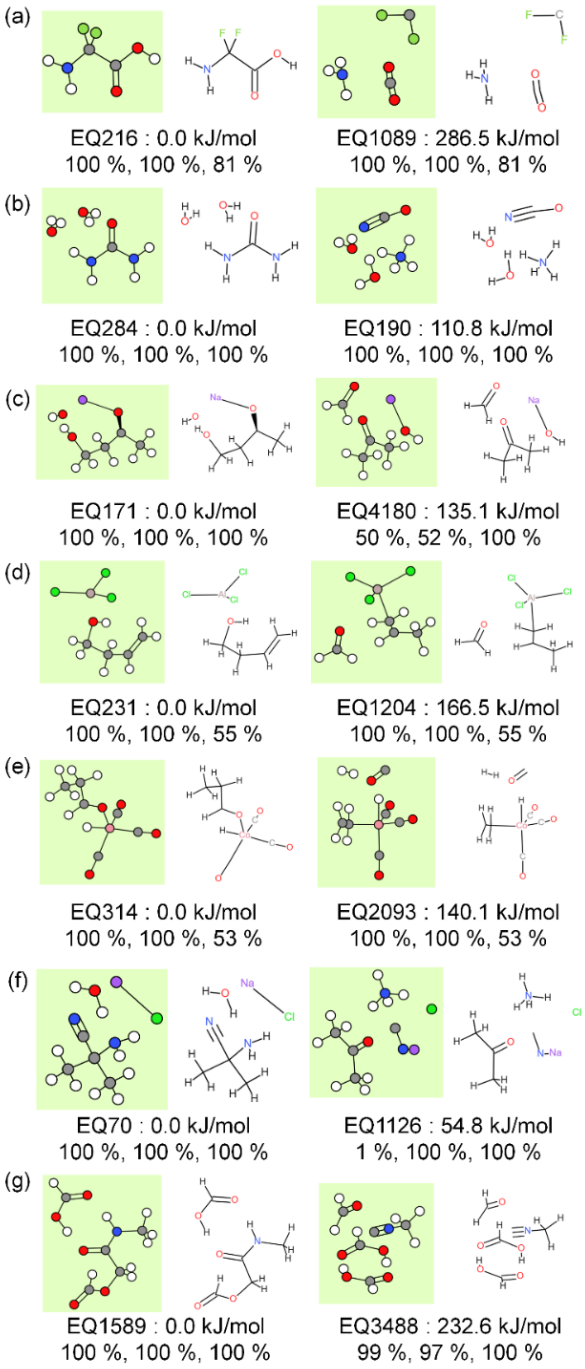


**Figure 3.** EQs of reactants and products of the reactions shown in **Figure 1**. (a) synthesis of fluoroglycine, (b) Wöhler's urea synthesis, (c) base-catalysed aldol reaction, (d) Lewis-acid-catalysed

ene reaction, (e) cobalt-catalysed hydroformylation, (f) Strecker reaction, and (g) Passerini reaction. The three reaction yields for each reaction are obtained by kinetic simulations at 200 K, 300 K, and 400 K. Electronic energies of EQs (kJ $mol^{-1}$) relative to the product EQs are also shown. Colour code: white, H; grey, C; blue, N; red, O; green, F; purple, Na; pink, Co.

**Synthesis of fluoroglycine.** The reaction-path network predicted the reactant EQ1089 consisting of $CO_2$, $NH_3$, and $CF_2$ molecules with a 100% reaction yield at 300 K (**Figure 3a**), which is consistent with a previous report.[45] In this network, one of the product EQs and EQ1089 were directly connected via a single-edge, and three components ($CO_2$, $NH_3$, and $CF_2$) were combined in a one-step reaction. As shown in **Table 1**, 446 species were identified in the reaction-path network, of which 54 species were predicted to afford the input products with theoretical yields greater than 50% at 300 K.

**Wöhler's urea synthesis.** An experimentally well-known reactant consisting of $NH_4^+$ and $OCN^-$ was obtained as EQ190 with a 100% yield (**Figure 3b**). The reaction-path network also includes an important intermediate, a zwitterionic adduct between HNCO and $NH_3$ ($HN{=}C(NH_3)^+{-}O^-$), which has been reported in a previous study.[64] As shown in **Table 1**, 262 species were identified in this reaction-path network, of which 29 species were predicted to afford the input products with theoretical yields greater than 50% at 300 K.

**Base-catalysed aldol reaction.** A known reactant consisting of acetone and formaldehyde molecules was obtained as EQ4180 with a 52% yield (**Figure 3c**). This calculation uses $Na^+$ and $OH^-$ as the base catalyst, which decreased the reaction barrier height. The $Na^+$ ion increased the possibility of conformationally different geometries, with more than 100 geometries identified as product EQs. As shown in **Table 1**, 1,087 species were identified in this reaction-path network, of which 115 species were predicted to afford the input products with theoretical yields greater than 50% at 300 K.

**Lewis-acid-catalysed ene reaction.** EQ1204 containing the known reactants (propylene and formaldehyde) was obtained with a 100% yield (**Figure 3d**). This calculation included the Lewis acid $AlCl_3$ for activation of the carbonyl group. Following the activation, C–C bond formation and proton transfer occur, as seen in the path shown in **Figure 2**. The reaction yield of the target product at 400 K was low (55%), which was attributed to the presence of several structures that were more stable. Therefore, this strategy also predicted the kinetic stability of the product at a given temperature. As shown in **Table 1**, 2,680 species were identified in this reaction-path network, of which 356 species were predicted to afford the input products with theoretical yields greater than 50% at 300 K.

**Cobalt-catalysed hydroformylation.** The node EQ2093 showed a 100% reaction yield (**Figure 3e**). EQ2093 consisted of CO, $H_2$, and a complex of $C_2H_4$ and the active catalyst $HCo(CO)_3$. This geometry is slightly different from that of the four component reactants shown in **Figure 1e**[65] owing to the coordination of ethylene to the Co centre, which occurred in a barrier-less manner on the potential energy surface during our calculations. Therefore, EQ2093 is equivalent to the four component reactants. As shown in **Table 1**, 4,629 species were identified in this reaction-path network, of which 855 species were predicted to afford the input products with theoretical yields greater than 50% at 300 K.

**Strecker reaction.** The network including the Strecker reaction (**Figure 3f**) has been discussed in our previous report.[47] The calculation resulted in a reaction-path network consisting of 9,203 EQs and 18,156 reaction paths. The 9,203 EQs were classified into 1,679 groups based on their bonding pattern. The well-known Strecker reaction[63] has been predicted correctly. Notably, when we performed network refinement, the additional procedure changed the number of EQs and PTs.[47] As shown in **Table 1**, 1,679 species were identified in this reaction-path network, of which 139 species were predicted to afford the input products with theoretical yields greater than 50% at 300 K.

**Passerini reaction.** The network including the Passerini reaction (**Figure 3g**) has also been

discussed in our previous report.[47] The calculation resulted in a reaction-path network consisting of 12,215 EQs and 18,202 reaction paths. The 12,215 EQs were classified into 2,407 groups based on their bonding pattern. The well-known Passerini reaction has been predicted via a known mechanism.[66,67] When we performed network refinement, the number of EQs and PTs changed.[47] As shown in **Table 1**, 2,407 species were identified in this reaction-path network, of which 417 species were predicted to afford the input products with theoretical yields greater than 50% at 300 K.

**Features of QCaRA-based networks**

To investigate more features of the QCaRA-based reaction-path networks, the network including base-catalysed aldol reaction was further investigated. The obtained network is shown in **Figure 4**, where nodes and edges represent EQs and reaction paths, respectively, and the nodes are coloured based on the theoretical yields of the target product of the reaction starting from each node. The reactions starting from various nodes afforded the target products in good yields. **Figure 5** shows the reactant candidates that possessed four or more molecular fragments and afforded theoretical yields larger than 50% at 300 K. Each geometry corresponds to the lowest energy EQ with the same bonding patterns. EQ4180 and EQ975 have been identified as the reactant and intermediate, respectively, for the well-known aldol reaction. In addition, QCaRA enabled the visualisation of numerous unknown reactions, providing potential target products. **Figure 5** also shows reactant candidates containing various molecules, namely, cyclopropene (EQ1132, EQ1084, and EQ1676), allene (EQ329 and EQ69), epoxide (EQ1932), and a four-membered ring (EQ1321). Highly reactive organosodium compounds, whose reactivities cannot be readily experimentally investigated, were also identified. Furthermore, many reactions with near-zero or zero yields were also predicted as blue nodes in the network, as shown in **Figure 4**. This database provides not only experimentally accessible reactions but also inaccessible reactions and near-zero- or zero-yield reactions, which would be valuable for future

informatics studies.

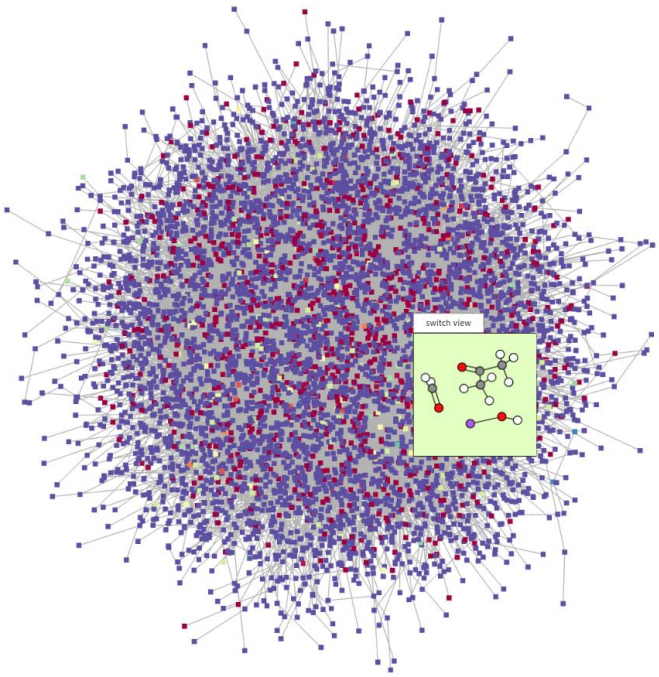


**Figure 4.** Network including base-catalysed aldol reaction. Nodes and edges represent EQs and reaction paths, respectively, and nodes are coloured based on theoretical yields of the target product of the reaction starting from each node at 300 K. The inset shows an EQ geometry corresponding to a well-known reactant of the base-catalysed aldol reaction. The network was visualized using the Searching Chemical Action and Network (SCAN) platform.[68]

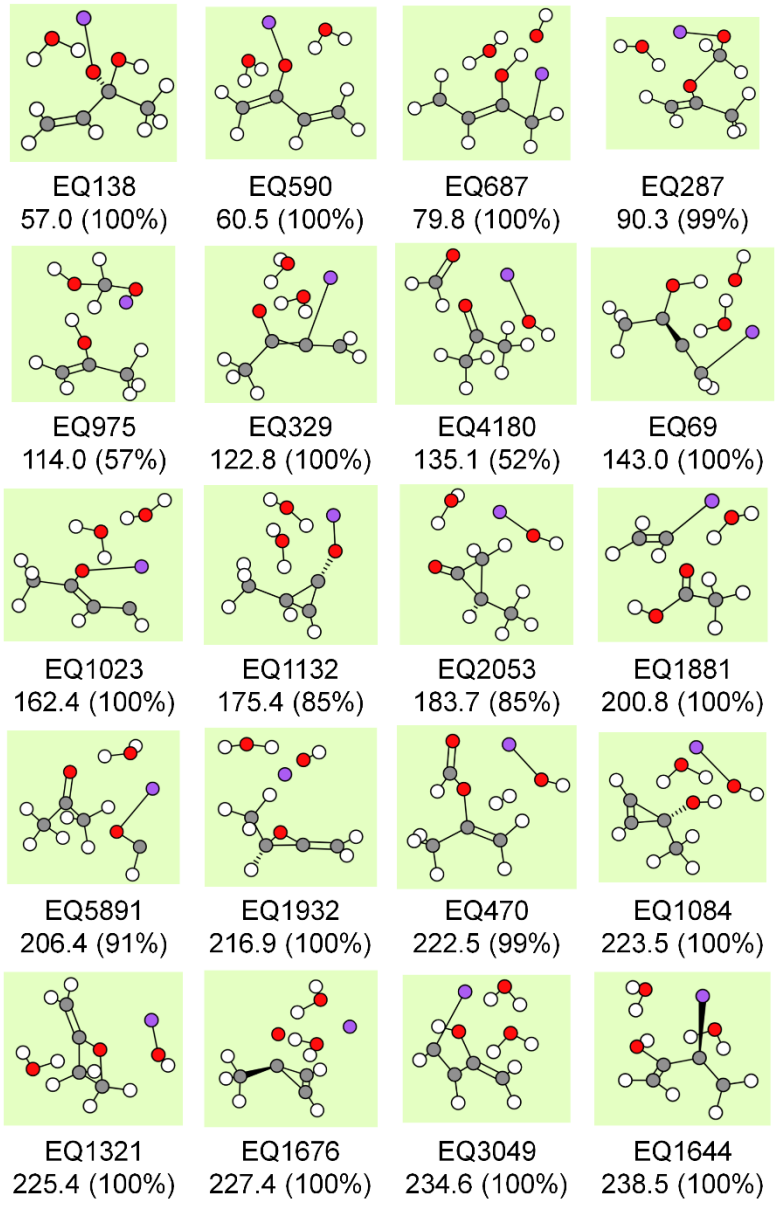


**Figure 5.** Reactant candidates for affording the product of base-catalysed aldol reaction. Sixteen reactant candidates with reaction yields larger than 50% at 300 K containing four or more molecular fragments are shown. Electronic energies of EQs (kJ $mol^{-1}$) relative to the product EQ are shown. The reaction yields correspond to the contribution ratios to the product state and are obtained by kinetic simulations at 300 K in a one-day time-period. Colour code: white, H; grey, C; red, O; purple, Na.

## Conclusions

In this article, we proposed theoretical chemical reaction database construction by presenting seven reaction-path networks obtained by combining QCaRA with AFIR and RCMC. Known reactions in these seven networks are: synthesis of fluoroglycine, Wöhler's urea synthesis, base-catalysed aldol reaction, Lewis-acid-catalysed ene reaction, cobalt-catalysed hydroformylation, Strecker reaction, and Passerini reaction. AFIR explored the reaction-path networks starting from a target product, while RCMC solved the kinetic equations inversely during the reaction-path exploration. Consequently, a network accessible to the target products under the given reaction conditions was obtained. This network also provided the theoretical yields of the target products for all reactions starting from the species predicted on the network. Both experimentally accessible and inaccessible chemical reactions have been identified, which is valuable data from an informatics perspective. In addition to known reactions, numerous high-, medium-, low-, near-zero-, and zero-yield reactions have been identified. Some of these data may have discrepancies with experimental data owing to the microscopic model and computational levels adopted. Nevertheless, we believe that this strategy provides information on experimentally unexplored or neglected reactivities. Furthermore, we will expand the database continuously by adding reaction-path networks for many other systems. For that, further high-throughput network data generation using the Fugaku supercomputer is currently underway.

Notably, the seven reaction-path network data will be available through the Searching Chemical Action and Network (SCAN) platform, where **Figures 3, 4,** and **5** were created using SCAN. The implementation, architecture, usage, and link of SCAN are described in another report.[68]

**Computational Details**

The single-component AFIR (SC-AFIR) method was implemented in the major version of the GRRM20 program package.[69] The SC-AFIR method was combined with Gaussian16[70] to achieve

an in-house-modified locally updated planes (LUP) method to relax and optimise the path.[71] The PT (edge) denotes the maximum energy of a geometry along an LUP path (**Figure 2b**). The three-dimensional structures and the corresponding energies of each EQ (node), PT (edge), and discrete geometry along the LUP paths were acquired. Inverse kinetic simulations[47] were used to estimate the reaction yield for each EQ. These calculations provided the final (refined) network containing the gradient, $< S^2 >$ values, and dipole moments in addition to the geometry and energy information for all geometries on the network. These calculations also provided the reaction yields from each EQ at 200 K, 300 K, and 400 K.

The SC-AFIR search with the QCaRA mode is performed to construct the reaction-path networks.[47] In the DFT calculations for reaction-path exploration, all gradient and Hessian calculations were performed using the $\omega$B97X-D functional with the Grid=FineGrid option in the vacuum. The Def2-SVP basis set was used for Na, Al, Cl, and Co atoms, whereas the SV basis set was used for the remaining atoms. The SC-AFIR search was conducted setting the $\gamma$ value to 500.0 kJ mol$^{-1}$, where $\gamma$ is the model collision energy parameter representing the strength of the artificial force. All atoms included in the system were set to the target of the SC algorithm.[29] A weak artificial force of $\gamma$ = 100.0/[$N$ ($N$ − 1) / 2] kJ mol$^{-1}$ was added between all atoms to prevent a substructure from being separated too far ($N$ is the number of atoms in the system). The force-induced reaction pathways were relaxed using the LUP method (denoted by LUP paths). AFIR further expanded the network based on the reaction yields of the target products obtained by the inverse kinetic simulations[47] via RCMC.[72] Gibbs energy values were evaluated assuming ideal gas, rigid-rotor, and harmonic vibrational models, where all harmonic frequencies smaller than 50 cm$^{-1}$ were set to 50 cm$^{-1}$.[73] The target products (input of the network explorations) are shown in **Figure 1**. Three reaction temperatures (200 K, 300 K, and 400 K) were considered, and the highest reaction yield at each temperature was adopted as the reaction yield of each EQ during the search. The reaction time was set to one day. The search was terminated

when one of the following conditions was met: (i) when a list of EQs with the 30$N$ largest $\Gamma_i$ values was not updated in the last 30$N$ path calculations, where $N$ is the number of atoms in the system and $\Gamma_i$ is the yield of the target product in a reaction starting from $EQ_i$ multiplied by the Boltzmann distribution of $EQ_i$; (ii) when the EQ with $\Gamma_i$ larger than $10^{-4}$ was not identified in the last 30$N$ paths; or (iii) when the SC-AFIR search computed 10,000 paths.

At all the discrete path points along all LUP paths within the reaction-path network, energy and gradient calculations were performed with the $\omega$B97X-D functional and Def2-SVP basis set. The Hessian is also computed at path terminals (EQs) and PTs. The Grid=FineGrid option was also employed in the calculations, and solvent effects were considered using the solvation model based on density method.[74] THF was adopted as the solvent in the synthesis of fluoroglycine, Passerini reaction, and Lewis-acid-catalysed ene reaction; water was adopted for the Strecker reaction, Wöhler's urea synthesis, and base-catalysed aldol reaction; and cobalt-catalysed hydroformylation was computed in vacuum. The results are presented in **Table 1**. By specifying the SaveDataAtAllPoints = 2 option of GRRM20, the gradients, energies, and dipole moments were stored in the additional output files. After the network refinement, kinetic analysis using the RCMC method was performed again at 200 K, 300 K, and 400 K based on the refined energetics, which provided the reaction yields of the target product starting from each EQ.

## Conflicts of interest

There are no conflicts to declare.


## Acknowledgements

The authors are grateful to Dr. Jun Fujima of the National Institute for Materials Science, Mr. Mikael Kuwahara, Dr. Lauren Takahashi, and Prof. Keisuke Takahashi of Hokkaido University

for the design and development of the SCAN data platform. This study was funded by the Japan Science and Technology Agency (JST) (ERATO Grant Numbers JPMJER1903 and JSPS-WPI). The computations of this study were performed using the Fugaku supercomputer at the RIKEN Center for Computational Science and the supercomputer system at the Hokkaido University Information Initiative Center.

**Notes and references**

1. E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.
2. H. Satoh and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 34–44.
3. A. Cook, A. P. Johnson, J. Law, M. Mirzazadeh, O. Ravitz and A. Simon, *WIREs Comput. Mol. Sci.*, 2012, **2**, 79–107.
4. S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem. Int Ed Engl*, 2016, **55**, 5904–5937.
5. M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
6. C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281–1289.
7. S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, *J. Cheminform.*, 2020, **12**, 70.
8. K. Lin, Y. Xu, J. Pei and L. Lai, *Chem. Sci.*, 2020, **11**, 3355–3364.
9. W. Bort, I. I. Baskin, T. Gimadiev, A. Mukanov, R. Nugmanov, P. Sidorov, G. Marcou, D. Horvath, O. Klimchuk, T. Madzhidov and A. Varnek, *Sci. Rep.*, 2021, **11**, 3178.
10. H. B. Schlegel, *J. Comput. Chem.*, 2003, **24**, 1514–1527.
11. K. N. Houk and P. H.-Y. Cheong, *Nature*, 2008, **455**, 309–313.
12. W. Thiel, *Angew. Chem. Int Ed Engl*, 2014, **53**, 8605–8613.
13. W. M. C. Sameera, S. Maeda and K. Morokuma, *Acc. Chem. Res.*, 2016, **49**, 763–773.

14. K. N. Houk and F. Liu, *Acc. Chem. Res.*, 2017, **50**, 539–543.
15. S. Ahn, M. Hong, M. Sundararajan, D. H. Ess and M. H. Baik, *Chem. Rev.*, 2019, **119**, 6509–6560.
16. D. J. Wales and T. V. Bogdan, *J. Phys. Chem. B*, 2006, **110**, 20765–20776.
17. S. Maeda, K. Ohno and K. Morokuma, *Phys. Chem. Chem. Phys.*, 2013, **15**, 3683–3701.
18. L. P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande and T. J. Martínez, *Nat. Chem.*, 2014, **6**, 1044–1048.
19. S. Maeda, Y. Harabuchi, Y. Ono, T. Taketsugu and K. Morokuma, *Int. J. Quantum Chem.*, 2015, **115**, 258–269.
20. Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 14621.
21. A. L. Dewyer, A. J. Argüelles and P. M. Zimmerman, *WIREs Comput. Mol. Sci.*, 2018, **8**, e1354.
22. C. A. Grambow, A. Jamal, Y. P. Li, W. H. Green, J. Zádor and Y. V. Suleimanov, *J. Am. Chem. Soc.*, 2018, **140**, 1035–1048.
23. G. N. Simm, A. C. Vaucher and M. Reiher, *J. Phys. Chem. A*, 2019, **123**, 385–399.
24. J. P. Unsleber and M. Reiher, *Annu. Rev. Phys. Chem.*, 2020, **71**, 121–142.
25. L. Takahashi, J. Ohyama, S. Nishimura and K. Takahashi, *J. Phys. Chem. Lett.*, 2021, **12**, 558–568.
26. Y. V. Suleimanov and W. H. Green, *J. Chem. Theor. Comput.*, 2015, **11**, 4248–4259.
27. P. M. Zimmerman, *J. Comput. Chem.*, 2015, **36**, 601–611.
28. L. P. Wang, R. T. McGibbon, V. S. Pande and T. J. Martinez, *J. Chem. Theor. Comput.*, 2016, **12**, 638–649.
29. S. Maeda, Y. Harabuchi, M. Takagi, T. Taketsugu and K. Morokuma, *Chem. Rec.*, 2016, **16**, 2232–2248.
30. S. Maeda, Y. Harabuchi, M. Takagi, K. Saita, K. Suzuki, T. Ichino, Y. Sumiya, K. Sugiyama and

Y. Ono, *J. Comput. Chem.*, 2018, **39**, 233–251.

31. S. Maeda and Y. Harabuchi, *WIREs Comput. Mol. Sci.*, 2021, **11**, e1538.
32. C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Comput. Phys. Commun.*, 2016, **203**, 212–225.
33. M. Liu, A. Grinberg Dana, M. S. Johnson, M. J. Goldman, A. Jocher, A. M. Payne, C. A. Grambow, K. Han, N. W. Yee, E. J. Mazeau, K. Blondal, R. H. West, C. F. Goldsmith and W. H. Green, *J. Chem. Inf. Model.*, 2021, **61**, 2686–2696.
34. R. Van de Vijver and J. Zádor, *Comput. Phys. Commun.*, 2020, **248**, 106947.
35. K. Hori, A. Hasegawa, N. Okimoto and S. Yamazaki, *J. Comput. Aided Chem.*, 2019, **20**, 50–55.
36. C. A. Grambow, L. Pattanaik and W. H. Green, *Sci. Data*, 2020, **7**, 137.
37. H. Okada and S. Maeda, *Mol. Inform.*, 2022, **41**, e2100216.
38. S. Chen, T. Nielson, E. Zalit, B. B. Skjelstad, B. Borough, W. J. Hirschi, S. Yu, D. Balcells and D. H. Ess, *Top. Catal.*, 2022, **65**, 312–324.
39. Q. Zhao, H. H. Hsu and B. M. Savoie, *J. Chem. Theor. Comput.*, 2022, **18**, 3006–3016.
40. A. Fernández-Ramos, J. A. Miller, S. J. Klippenstein and D. G. Truhlar, *Chem. Rev.*, 2006, **106**, 4518–4584.
41. S. Kozuch and S. Shaik, *Acc. Chem. Res.*, 2011, **44**, 101–110.
42. Y. Sumiya, T. Taketsugu and S. Maeda, *J. Comput. Chem.*, 2017, **38**, 101–109.
43. J. Ford, S. Seritan, X. Zhu, M. N. Sakano, M. M. Islam, A. Strachan and T. J. Martínez, *J. Phys. Chem. A*, 2021, **125**, 1447–1460.
44. D. Garay-Ruiz, M. Álvarez-Moreno, C. Bo and E. Martínez-Núñez, *ACS Phys. Chem. Au*, 2022, **2**, 225–236.
45. T. Mita, Y. Harabuchi and S. Maeda, *Chem. Sci.*, 2020, **11**, 7569–7577.
46. H. Hayashi, H. Takano, H. Katsuyama, Y. Harabuchi, S. Maeda and T. Mita, *Chem. Eur. J.*, 2021,

**27**, 10040–10047.

47. Y. Sumiya, Y. Harabuchi, Y. Nagata and S. Maeda, *JACS Au*, 2022, **2**, 1181–1188.
48. F. Wöhler, *Ann. Phys. Chem.*, 1828, **87**, 253–256.
49. C. A. Wurtz, *Bull. Soc. Chim. Fr.*, 1872, **17**, 436–442.
50. A. T. Nielsen and W. J. Houlihan, in Org. React., John Wiley & Sons, Inc., Hoboken, NJ, 2011, 1–438.
51. T. Mukaiyama, in Org. React., John Wiley & Sons, Inc., Hoboken, NJ, 1982, 203–331.
52. H. J. Prins, *J. Chem. Weekbl.*, 1919, **16**, 1072.
53. H. J. Prins, *J. Chem. Weekbl.*, 1919, **16**, 1510.
54. R. F. Heck and D. S. Breslow, *J. Am. Chem. Soc.*, 1961, **83**, 1097–1102.
55. L. E. Rush, P. G. Pringle and J. N. Harvey, *Angew. Chem. Int Ed Engl*, 2014, **53**, 8672–8676.
56. F. Hebrard and P. Kalck, *Chem. Rev.*, 2009, **109**, 4272–4282.
57. A. Strecker, *Ann. Chem. Pharm.*, 1850, **75**, 27–45.
58. T. W. G. Solomons, Organic chemistry, Wiley N. Y., 1996, 6. ed.
59. L. Kürti and B. Czakó, Strategic Applications of Named Reactions in Organic Synthesis: Background and Detailed Mechanisms, Elsevier Academic Press, Amsterdam; Boston, 2005.
60. M. Passerini and L. Simone, *Gazz. Chim. Ital.*, 1921, **51**, 126–129.
61. J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev and A. E. Roitberg, *Nat. Commun.*, 2019, **10**, 2903.
62. J. Li, K. Song and J. Behler, *Phys. Chem. Chem. Phys.*, 2019, **21**, 9672–9682.
63. A. Nandi, C. Qu, P. L. Houston, R. Conte and J. M. Bowman, *J. Chem. Phys.*, 2021, **154**, 051102.
64. Y. Sumiya and S. Maeda, *Chem. Lett.*, 2019, **48**, 47–50.
65. K. Takahashi and M. Satoshi, *RSC Adv.*, 2021, **11**, 23235–23240.
66. S. Maeda, S. Komagawa, M. Uchiyama and K. Morokuma, *Angew. Chem. Int. Ed.*, 2011, **50**,

644–649.

67. R. Ramozzi and K. Morokuma, *J. Org. Chem.*, 2015, **80**, 5652–5657.
68. J. Fujima, Y. Harabuchi, M. Kuwahara, L. Takahashi, S. Maeda and K. Takahashi, submitted.
69. GRRM20, https://www.hpc.co.jp/chem/software/grrm20_e.
70. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, F. D. Williams, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, 2016.
71. C. Choi and R. Elber, *J. Chem. Phys.*, 1991, **94**, 751–760.
72. Y. Sumiya and S. Maeda, *Chem. Lett.*, 2020, **49**, 553–564.
73. R. F. Ribeiro, A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2011, **115**, 14556–14562.
74. A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.