Exact solution for minimization of root mean square deviation with G-RMSD to determine molecular similarity

Nabika, Tomohiro ; Iwata, Satoru ; Satoh, Hiroko

DOI: https://doi.org/10.1093/bulcsj/uoae037

Bulletin of the Chemical Society of Japan, 2024, **97**, uoae037
https://doi.org/10.1093/bulcsj/uoae037
Advance access publication 18 April 2024

Articles

Exact solution for minimization of root mean square deviation with G-RMSD to determine molecular similarity

Tomohiro Nabika[1], Satoru Iwata[2,3,*], Hiroko Satoh[4,5]

[1]Department of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha Kashiwa, Kashiwa, Chiba 277-8561, Japan

[2]Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

[3]Institute for Chemical Reaction Design and Discovery, Hokkaido University, Kita 21 Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0021, Japan

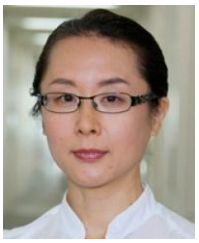
[4]Department of Chemistry, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[5]Joint Support Center for Data Science Research, Research Organization of Information and Systems (ROIS), 10-3 Midori-cho, Tachikawa, Tokyo 190-0014, Japan

*Corresponding author: Department of Mathematical Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan; Institute for Chemical Reaction Design and Discovery, Hokkaido University, Sapporo, Hokkaido 001-0021, Japan. Email: iwata@mist.i.u-tokyo.ac.jp

Tomohiro Nabika
Tomohiro Nabika graduated from the Department of Mathematical Engineering and Information Physics, University of Tokyo in 2022. He is currently an M.S. student at the Graduate School of Frontier Science at The University of Tokyo. His current research interest is in data-driven science with Bayesian estimation.

Satoru Iwata
Satoru Iwata received his M. Eng. degree from the University of Tokyo in 1993 and Ph.D. degree from Kyoto University in 1996. He was appointed as a research associate at the Research Institute for Mathematical Sciences, Kyoto University in 1994. After working at Osaka University, University of Tokyo, and Kyoto University, he joined the University of Tokyo as a professor in the Department of Mathematical Informatics in 2013. His current research interests include combinatorial optimization and its applications to computational chemistry. He shared the Delbert Ray Fulkerson Prize in 2003 for his joint work on submodular function minimization.



Hiroko Satoh
Hiroko Satoh received her Ph.D. degree from Ochanomizu University in 1996. She was a post-doctoral fellow at RIKEN Institute during 1996–1998 and conducted her project on data-driven chemistry under a PRESTO program of Japan Science in 2002 at the National Institute of Informatics (NII), Tokyo, Japan. In 2015 she moved to the University of Zurich and was concurrently appointed as project Associate Professor at the Research Organization of Information and Systems (ROIS). Her research interests cover a broad range of development and applications of computational and data-driven chemistry.

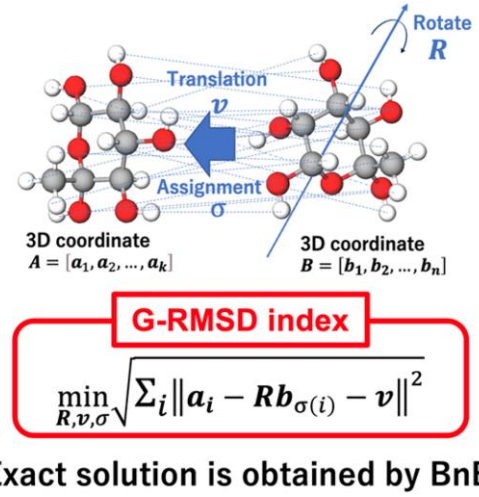
Abstract

Generalized root mean square deviation (G-RMSD) is an optimization method for three-dimensional molecular similarity determination. It calculates the minimum value of RMSD among all the possible one-to-one matchings between the atoms and positions of the molecules. The first paper on G-RMSD introduced two approaches called alternating optimization (AO) and tangent space relaxation (TSR) methods, which give local optimum solutions. We propose here a new method of G-RMSD using a branch-and-bound method (BnB) on isometric transformations, called IsometryOpt, which is mathematically proven to give an exact G-RMSD index, i.e. this method can reach the global optimum solution. The performance of IsometryOpt was compared to AO and TSR, as well as the MatchFastOpt method. IsometryOpt shows better performance than MatchFastOpt for molecules with the same number of atoms. AO and TSR fail to reach exact values in some cases. We also have developed two improved methods to search for all possible matches of a substructure in one or more molecules. One is called IsometrySearch, which uses BnB on isometric transformations. The other is a variant version of MatchFPT, called MatchFPT-delta. Computer experiments indicate that MatchFPT-delta performs better than MatchFPT and IsometrySearch.

Keywords: branch-and-bound method, molecular similarity, RMSD minimization, root mean square deviation (RMSD).

[Received on 16 November 2023; revised on 24 February 2024; accepted on 4 March 2024; corrected and typeset on 18 April 2024]

Graphical Abstract



Generalized root mean square deviation (G-RMSD) is an optimization method for three-dimensional molecular similarity determination. It calculates the minimum RMSD values without fixing both, atom mapping and superimposition. We propose a new method of G-RMSD using a branch-and-bound method which is mathematically proven to give an exact solution of the G-RMSD index.

1. Introduction

The measuring of molecular similarity is one of the fundamental tasks in chemical research. This task is employed daily, for example, in the study of structure-property, -activity or -reactivity relationships and in the search of chemical databases. Recently, as data-driven methods become more commonly used in chemistry, this task is becoming increasingly important. It is conducted using measuring methods or molecular descriptors, which have been developed in the chemoinformatics field intensively in the last half century.[1–4] Most of them are based on one or two-dimensional (1D or 2D) features of molecular structures. Although there are several useful 3D-based methods,[2,3,5–10] the necessity of further development has become clear as the need to handle 3D molecular structures increases.

Most of the molecular descriptors are not designed to handle chemical structures beyond valence bond (VB) theory, i.e. structures with non-standard bond lengths, such as structures in transition states (TSs) or dissociation channels (DCs), clusters, metal complexes, and molecules on a surface or in a cage. These types of molecules often play a central role in catalytic systems. The distance of 3D similarity is commonly measured with root mean square deviation (RMSD), which can also be applied to these types of structures beyond VB theory. However, RMSD usually has to be calculated with a unique atom mapping[11] or fixed superimposition.[12] These standard RMSD calculations cannot be applied for more flexible use, such as in the search of databases with 3D molecules,[13–23] especially including molecules beyond VB theory.[20–22]

To address this issue, Fukutani and some of the current authors developed the generalized RMSD (G-RMSD) method, which calculates the minimum value of RMSD among all the possible one-to-one matchings between the atoms and positions of the molecules.[24] The first report on G-RMSD in 2021 introduced two heuristic approaches called alternating optimization (AO) and tangent space relaxation (TSR) methods. These methods efficiently optimize the similarity between two molecules to give a similarity index, called the G-RMSD index, without fixing atom mapping and superimposition. The only necessary input is the Cartesian coordinates of atoms. The methods were successfully applied to the clustering of conformers of D-glucose and to full- and partial 3D structure searches of a reaction map database, RMapDB.[21] They produced similarity index values that are sufficient for practical use. These efficient methods are also useful for the rapid screening of similarities for a large dataset, like trajectories from molecular dynamics (MD) simulations.[25]

However, there is a trade-off between efficiency and robustness, i.e. there is no guarantee that the results obtained from the AO and TSR are exact minimum values. When stating scientific conclusions or benchmarking measuring methods, it is desirable to obtain the optimal or a most probable optimal solution, even if there is a higher computation cost.

In the field of computer vision, 3D point-set registration, which is similar to the optimization issue of G-RMSD, has been studied for a long time. This problem is mainly solved using an alternating optimization method called the ICP algorithm.[26] It is known that this algorithm tends to fall into local optimum solutions and often fails to find a global optimum solution. Yang et al.[27] devised the Go-ICP method to obtain an exact solution for the registration problem by using the branch-and-bound method (BnB) for rotations and translations while using the ICP algorithm. BnB is a common mathematical algorithm for finding the optimal solution by enumerating all possible solutions, dividing the solution space into problem subsets, and then pruning the subsets using lower bounds (Fig. 1).

The methods for minimization of RMSD[24,28] can be applied to the substructure search problem, namely, the problem to find a given substructure in larger chemical structures. In chemistry, the *substructure search* problem is generally defined to find a substructure in one or more molecules. The problem involves both finding the best match and enumerating all possible matches. However, in mathematical algorithms, these are treated as two different types of tasks. The former

task is defined as finding the optimal G-RMSD between two molecules with different sizes. The latter task is defined as the problem of enumerating all possible atom pairs under certain conditions. Therefore, to avoid possible confusion in the descriptions of algorithms, we here name the former and latter tasks *substructure match* and *substructure enumeration*, respectively. Both of the tasks are performed according to the G-RMSD concept, i.e. no need to fix atom-mapping and superimposition for the best match and enumeration.

In this paper, we introduce a new rigorous algorithm, called IsometryOpt, which can be used to obtain a global optimal solution of G-RMSD calculations. The algorithm employs BnB for isomeric transformations, analogous to the Go-ICP method for registration. By using BnB, this method can be mathematically proven to give an exact G-RMSD index, i.e. the global optimal solution.

For the substructure enumeration task, we also have developed two new methods. One is called IsometrySearch, which uses BnB on isometric transformations, and the other is called MatchFPT-delta, which is a variant version of MatchFPT developed by Sasaki et al.[28]

First, we will describe detailed algorithms of IsometryOpt, IsometrySearch, and MatchFPT-delta methods. In the second half of this paper, we will discuss their performance by computer experiments using artificial and chemical data. IsometryOpt has been tested and compared to AO and TSR for molecules with equal and different sizes. IsometrySearch and MatchFPT-delta together with MatchFPT have also been tested by numerical experiments as well as using chemical data.

2. Definition of G-RMSD

G-RMSD is the concept that we proposed in a previous paper[24] and is used as the name of code. The similarity index output from G-RMSD is called G-RMSD index. We refer to other methods that calculate RMSD values based on a similar concept, such as MatchFastOpt,[28] as the methods in the G-RMSD concept. To avoid possible confusion, we name all the similarity index calculated based on the G-RMSD concept or a similar concept as generalized similarity (GS)-index here.

Consider a comparison between molecule A with k atoms and molecule B with n ($\geq k$) atoms. Suppose that the coordinates of the i-th atom of A are $\boldsymbol{a}_i (i = 1, \ldots, k)$ and the coordinates of the i-th atom of B are $\boldsymbol{b}_i (i = 1, \ldots, n)$. Let $\boldsymbol{A}$, $\boldsymbol{B}$ be the matrices $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k]$, $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n]$. In this case, the RMSD between molecules A and B is defined by

$$\mathrm{RMSD}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{R}, \boldsymbol{v}, \sigma) = \sqrt{\frac{1}{k}\sum_{i=1}^{k} ||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_{\sigma(i)} - \boldsymbol{v}||^2}, \quad (1)$$

where $\boldsymbol{R}$ is a 3×3 orthogonal matrix, $\boldsymbol{v}$ is a 3-dimensional vector, and σ is an injection from $\{1, \ldots, k\}$ to $\{1, \ldots, n\}$. Namely, $\boldsymbol{R}$, $\boldsymbol{v}$, and σ correspond to rotation, translation, and atom mapping, respectively. For any $\{1, \ldots, k\}$, we may add the condition that the i-th atom of A and the $\sigma(i)$ -th atom of B are the same.

The G-RMSD index[24] is defined to be

$$\min_{\boldsymbol{R},\boldsymbol{v},\sigma} \mathrm{RMSD}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{R}, \boldsymbol{v}, \sigma). \quad (2)$$

This minimizes RMSD among isometric transformations of B and matching from A to B (Fig. 2a and 2b). When this value is small, it indicates that the two molecules are similar.

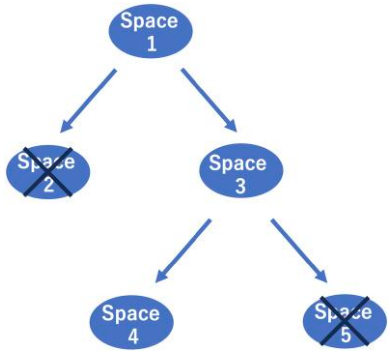


Fig. 1. Optimization flow of BnB. It involves enumerating all possible solutions, dividing the solution space into problem subsets (e.g. Space 2 and Space 3) and then pruning the subsets using lower bounds. In this figure, Space 2 and Space 5 are pruned and the entire solution space (Space 1) is limited to a subset Space 4.

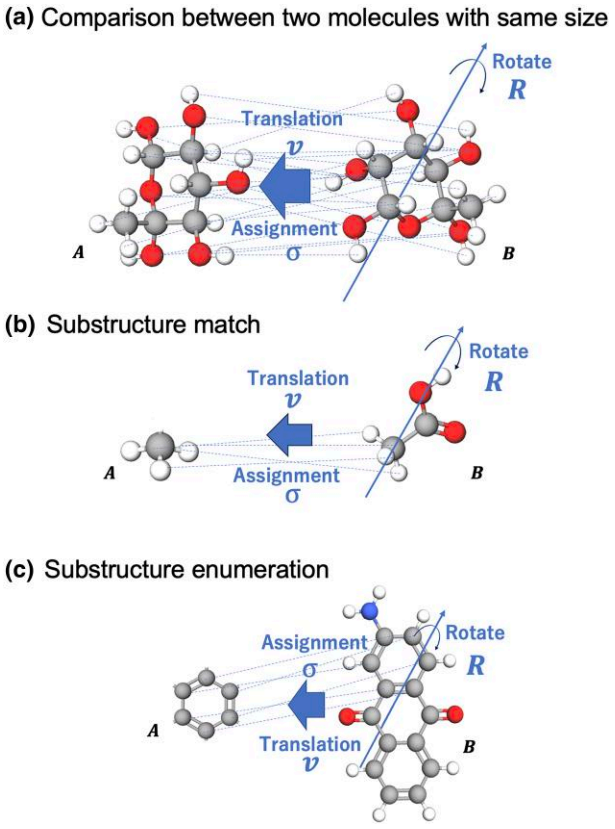


Fig. 2. Determination of 3D molecular similarity according to the G-RMSD concept: no fixing atom mapping and superimposition. a) Comparison between two molecules with equal size. The atom-pair [atom mapping (assignment) σ] and superimposition (rotation $\boldsymbol{R}$ and translation $\boldsymbol{v}$) are optimized to minimize the RMSD value. b) Substructure match. Two molecules with different size are compared, and the atom-pair [atom mapping (assignment) σ] and superimposition (rotation $\boldsymbol{R}$ and translation $\boldsymbol{v}$) are optimized to minimize the RMSD value. c) Substructure enumeration. All possible atom-pairs are enumerated in one or more molecules by changing superimposition (rotation $\boldsymbol{R}$ and translation $\boldsymbol{v}$). In this instance, two benzene rings should be found in 2-aminoanthracene-9,10-dione (*B*). The enumeration is performed in 3D.

In particular, when $n \geq k$ it is possible to determine whether molecule B has a similar substructure to A.

The G-RMSD concept can be applied to the substructure enumeration (Fig. 2c) to find all injections σ satisfying

$$\min_{R,v} \sqrt{\frac{1}{k}\sum_{i=1}^{k} ||a_i - Rb_{\sigma(i)} - v||^2} \leq \varepsilon, \tag{3}$$

where $\varepsilon \geq 0$.[28] The other variables are the same as in eq. (2). We call the set of all injections σ, which satisfies eq. (3) as GS_{sub} -index.

3. Mathematical background

3.1 Partial optimizations of RMSD

3.1.1 Partial optimization for matching. For fixed R and v, consider finding σ that minimizes RMSD, i.e. solving the optimization problem:

$$\min_{\sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k} || a_i - Rb_{\sigma(i)} - v||^2}. \tag{4}$$

This is equivalent to finding a minimum weight maximum matching in a bipartite graph with vertex sets $\{1, \ldots, k\}$ and $\{1, \ldots, n\}$, where the weight of edge (i, i') is $||a_i - Rb_{i'} - v||^2$. This problem can be solved by the Hungarian method[29] in $O(k^2 n)$ time.

3.1.2 Partial optimization for translations. For a fixed σ, consider finding R and v that minimize RMSD, i.e. solving the optimization problem:

$$\min_{R,v} \sqrt{\frac{1}{k}\sum_{i=1}^{k} || a_i - Rb_{\sigma(i)} - v||^2}. \tag{5}$$

In fact, it is generally easy to obtain an optimal translation v. Then eq. (5) is equivalent to the following problem:

$$\min_{R} \sqrt{\frac{1}{k}\sum_{i=1}^{k} || a'_i - Rb'_{\sigma(i)}||^2}. \tag{6}$$

This can be solved via the singular value decomposition.[30]

3.2 Alternate optimization (AO) method

3.2.1 Similarity between molecules of equal size. One of the methods proposed in our previous paper[24] is called the alternate optimization (AO) method. When the number of atoms is the same $k = n$, eq. (1) is equivalent to the following problem:

$$\min_{R,\sigma} \frac{1}{k}\sum_{i=1}^{k} ||a'_i - Rb'_{\sigma}(i)||^2. \tag{7}$$

We can obtain a local optimum solution by repeating,

$$\sigma_{l+1} = \operatorname*{argmin}_{\sigma} \frac{1}{k}\sum_{i=1}^{k} || a'_i - R_l b'_{\sigma(i)}||^2, \tag{8}$$

$$R_{l+1} = \operatorname*{argmin}_{R} \frac{1}{k}\sum_{i=1}^{k} || a'_i - Rb'_{\sigma_{l+1}(i)}||^2, \tag{9}$$

from an arbitrary R_0 until $\sigma_{l+1} = \sigma_l$. The first paper on G-RMSD[24] proposed to adopt 120 orthogonal matrices that make the icosahedron invariant as initial matrices of R_0. A local optimal solution is obtained for each of them, and the smallest RMSD among them is tentatively used as the G-RMSD index.

3.2.2 Substructure match between molecules of different size. When $k \neq n$, we obtain a local optimum solution by repeating

$$\sigma_{l+1} = \operatorname*{argmin}_{\sigma} \frac{1}{k}\sum_{i=1}^{k} ||a_i - R_l b_{\sigma(i)} - v_l||^2, \tag{10}$$

$$R_{l+1}, v_{l+1} = \operatorname*{argmin}_{R,v} \frac{1}{k}\sum_{i=1}^{k} ||a_i - Rb_{\sigma_{l+1}(i)} - v||^2, \tag{11}$$

from R_0, v_0, until $\sigma_{l+1} = \sigma_l$. Here 120 orthogonal matrices are also adopted to make the icosahedron invariant and $b_{i'} - a_i$ $(i = 1, \ldots, n, \quad i' = 1, \ldots, k)$, as initial matrices R_0 and initial vector v_0, respectively.[24] A local optimal solution is obtained for each of them, and the smallest RMSD value of those is tentatively used as the G-RMSD index.

3.3 Tangent space relaxation (TSR) method

Another method we proposed in the previous paper is called the tangent space relaxation (TSR) method. When $k = n$, eq. (1) is equivalent to the following problem:

$$\max_{R,P} \operatorname{trace}(RB'PA'^{\top}), \tag{12}$$

where $A' = [a'_1, a'_2, \ldots, a'_k]$, $B' = [b'_1, b'_2, \ldots, b'_n]$, and P is an $n \times n$ permutation matrix. Let $\mathcal{S}$ denote the set of all 3×3 skew-symmetric matrices. A local optimum solution is obtained by repeating

$$P_{i+1} = \operatorname*{argmin}_{P} \operatorname{trace}(R_i B'PA'^{\top}), \tag{13}$$

$$S_{i+1} = \operatorname*{argmin}_{S\in\mathcal{S}, ||S||_F = 1} \operatorname{trace}(SP_{i+1}A'B'R_iB'^{\top}), \tag{14}$$

$$R_{i+1} = \operatorname*{argmin}_{R} \operatorname{trace}(RB(P_{i+1} + \alpha SP_{i+1})A^{\top}), \tag{15}$$

from an arbitrary R_0 until $P_{i+1} = P_i$, where $||S||_F$ denotes the Frobenius norm of S. This is called the TSR method,[24] in which it is suggested to set α and to adopt 120 orthogonal matrices that make the icosahedron invariant as initial matrices R_0. A local optimal solution is obtained for each of them, and the smallest RMSD value of those is tentatively used as the G-RMSD index.

For further information on AO and TSR, refer to the first report on G-RMSD.[24]

3.4 BnB on translation

3.4.1 Parameterization of translation. Since the translation vector v is 3D, the possible values of translation can be expressed in 3D space. Assuming that we re-coordinate

$\boldsymbol{A}$ so that the center of gravity is at the origin, $\boldsymbol{v}$ can be expressed as

$$\boldsymbol{v}^* = -\frac{1}{k}\boldsymbol{R}\left(\sum_{i=1}^{k}\boldsymbol{b}_{\sigma(i)}\right) \tag{16}$$

and

$$||\boldsymbol{v}^*|| \leq \max_{i=1,\ldots,n} ||\boldsymbol{b}_i|| \tag{17}$$

holds. Thus, $\boldsymbol{v}$ is contained in a cube centered at the origin with a side length of $2\max_{i=1,\ldots,n}||\boldsymbol{b}_i||$.

By dividing each side into two equal parts, we can divide the cube into eight equal parts.

3.4.2 Inequality for translation. Let $\boldsymbol{v}_0$ be the center of the cube C. For an arbitrary vector $\boldsymbol{v}$ contained in C, we have

$$||\boldsymbol{v} - \boldsymbol{v}_0|| \leq \gamma, \tag{18}$$

where γ is the distance between the center and a vertex of C. Here, the distance is the same for any vertices of C because C is a cube. Then it follows from the triangular inequality that

$$||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_j - \boldsymbol{v}|| \geq \max\left(||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_j - \boldsymbol{v}_0|| - \gamma, 0\right) \tag{19}$$

holds for any i, j. Thus, we have

$$\min_{\boldsymbol{R},\boldsymbol{v}\in C,\sigma} \mathrm{RMSD}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{R}, \boldsymbol{v}, \sigma)$$

$$\geq \min_{\boldsymbol{R},\sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k}\max\left(||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_j - \boldsymbol{v}_0|| - \gamma, 0\right)^2}. \tag{20}$$

Also, we have

$$\min_{\boldsymbol{R},\boldsymbol{v}\in C,\sigma} \mathrm{RMSD}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{R}, \boldsymbol{v}, \sigma) \leq \min_{\boldsymbol{R},\ \sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k}||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_j - \boldsymbol{v}_0||^2}. \tag{21}$$

3.5 BnB method on orthogonal matrices

3.5.1 Parameterization of orthogonal matrices. A 3×3 orthogonal matrix with positive determinant represents a 3D rotation, which can be expressed in terms of the direction of the axis and the angle of the rotation. A rotation by θ around the axis $\ell = (\ell_1, \ell_2, \ell_3)$ corresponds to

$$r = \frac{\theta}{||\ell||}(\ell_1, \ell_2, \ell_3)^{\top}. \tag{22}$$

Then we can represent the set of rotations by a ball of radius π. Similarly, a 3×3 orthogonal matrix with negative determinant can be represented by a ball of radius π. We put each of these balls of radius π into the cube $[-\pi, \pi]^3$ and divide the cube into eight equal parts by dividing each side of the cube into two equal parts.

3.5.2 Inequality for orthogonal matrices. Since we represent the space of orthogonal matrices by a cube and divide the cube into cubes, we consider an inequality for a vector in a cube representing an orthogonal matrix. Let $\boldsymbol{R}_r$ correspond to $\boldsymbol{r}$ and $\boldsymbol{r}_0$ be the center of the cube D. If 2η is a side of the cube, then we have

$$||\boldsymbol{R}_r\boldsymbol{b}_i - \boldsymbol{R}_{r_0}\boldsymbol{b}_j|| \leq 2\sin\left(\min\left(\sqrt{3}\eta/2, \pi/2\right)\right)||\boldsymbol{b}_i|| \stackrel{\mathrm{def}}{=} \xi_i. \tag{23}$$

Then, for any i, j,

$$||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_j - \boldsymbol{v}|| \geq \max\left(||\boldsymbol{a}_i - \boldsymbol{R}_{r_0}\boldsymbol{b}_j - \boldsymbol{v}_0|| - \xi_i, 0\right) \tag{24}$$

holds.[27] Thus, we have

$$\min_{\boldsymbol{r}\in D,\boldsymbol{v}\in C,\sigma} \mathrm{RMSD}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{R}, \boldsymbol{v}, \sigma)$$

$$\geq \min_{\sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k}\max\left(||\boldsymbol{a}_i - \boldsymbol{R}_{r_0}\boldsymbol{b}_j - \boldsymbol{v}_0|| - \gamma - \xi_i, 0\right)^2} \tag{25}$$

and

$$\min_{\boldsymbol{r}\in D,\sigma} \mathrm{RMSD}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{R}, \boldsymbol{v}_0, \sigma)$$

$$\leq \min_{\boldsymbol{R},\ \sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k}\max\left(||\boldsymbol{a}_i - \boldsymbol{R}_{r_0}\boldsymbol{b}_j - \boldsymbol{v}_0|| - \xi_i, 0\right)^2}, \tag{26}$$

where D is the cube with center $\boldsymbol{r}_0$ and a side length of 2η.

3.6 BnB on assignment of atom-pairs

Let k and n be integers with $n \geq k$. The parameter space of injections from $\{1, \ldots, k\}$ to $\{1, \ldots, n\}$ is of size $n!/k!$. When $l < k$, we can divide $\{\sigma|\sigma(1) = k_1, \ldots, \sigma(l) = k_l\}$

$$\{\sigma|\sigma(1) = k_1, \ldots, \sigma(l) = k_l\} = \bigcup_{t\notin\{1,\ldots,k_l\}} \{\sigma|\sigma(1) = k_1, \ldots, \sigma(l) = k_l, \ \sigma(l+1) = k_{l+1}\} \tag{27}$$

If $\sigma \in \{\sigma|\sigma(1) = k_1, \ldots, \sigma(l) = k_l\}$, then

$$\min_{\boldsymbol{R},\boldsymbol{v}} \sqrt{\frac{1}{k}\sum_{i=1}^{l}||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_{k_i} - \boldsymbol{v}||^2} \leq \min_{\boldsymbol{R},\boldsymbol{v},\sigma} \mathrm{RMSD}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{R}, \boldsymbol{v}, \sigma) \tag{28}$$

holds.[28]

4. Existing methods for exact solution of GS-index

In this section, we describe the details of existing algorithms for obtaining an exact solution of GS-index and $\mathrm{GS}_{\mathrm{sub}}$ -index, called MatchFastOpt and MatchFPT, which were developed by Sasaki et al.[28]

4.1 MatchFastOpt

MatchFastOpt is for obtaining the exact solution of GS-index developed by Sasaki et al.[28] Eq. (2) can be solved by BnB for assignment. When $\{\sigma|\sigma(1) = k_1, \ldots, \sigma(l) = k_l\}$ is given as a search space, we define $\underline{E}$ by

$$\underline{E} = \min_{\boldsymbol{R},\boldsymbol{v}} \sqrt{\frac{1}{k}\sum_{i=1}^{l}||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_{k_i} - \boldsymbol{v}||^2}. \tag{29}$$

A specific method is given by Algorithm 1. MatchFastOpt can also solve eq. (7) by setting $\boldsymbol{v} = 0$.

Algorithm 1: MatchFastOpt

Input: Coordinate data of two molecules $\boldsymbol{A}$, $\boldsymbol{B}$.
Output: Minimum value E^* of RMSD and the corresponding $\boldsymbol{R}^*$, $\boldsymbol{v}^*$, σ^*.
1. Put the set of injections from $\{1, \ldots, k\}$ to $\{1, \ldots, n\}$ into a set Q.
2. Let $E^* = \infty$.
3. **while** Q is not empty **do**
4. Read out C we put most recently from Q.
5. Divide C into $\mathcal{C}$ based on Sect. 2.7.
6. **for each** C in $\mathcal{C}$ **do**
7. Compute $\underline{E}$.
8. **if** $\underline{E} < E^*$ **then**
9. **if** $|C| = 1$ **then**
10. Let $E^* = \underline{E}$. Update $\boldsymbol{R}^*$, $\boldsymbol{v}^*$, σ^*.
11. **else**
12. Put C into Q.
13. **end if**
14. **end if**
15. **end for**
16. **end while**

Algorithm 2: MatchFPT

Input: Coordinate data of two molecules $\boldsymbol{A}$, $\boldsymbol{B}$, and ε which determines the substructure.
Output: $\boldsymbol{R}^*$, $\boldsymbol{v}^*$, σ^* which represent substructures.
1. **for** $i \in \{1, \ldots, n\}$ **do**
2. Find $\boldsymbol{b}_{i_1}, \ldots, \boldsymbol{b}_{i_l}$ in the ball with radius $2\left(k^{\frac{1}{2}}\varepsilon + \ell\right)$
3. Put the set of injections from $\{1, \ldots, k\}$ to $\{i, \ldots, i_l\}$ into a set Q.
4. Let $E^* = \infty$.
5. **while** Q is not empty **do**
6. Read out C we put most recently from Q.
7. Divide C into $\mathcal{C}$ based on Sect. 2.7.
8. **for each** C in $\mathcal{C}$ **do**
9. Compute $\underline{E}$.
10. **if** $\underline{E} < \varepsilon$ **then**
11. **if** $|C| = k$ **then**
12. σ corresponding to $\underline{E}$ is one of the solution.
13. **else**
14. Put C into Q.
15. **end if**
16. **end if**
17. **end for**
18. **end while**
19. **end for**

4.2 MatchFPT

MatchFPT is for obtaining the exact solution of GS_{sub} -index to enumerate the substructure. Eq. (3) can be solved by BnB for assignment. To speed up, Sasaki et al.[28] used the following theorem.

Theorem 1[28]
Let $\boldsymbol{A} \in \mathbb{R}^{3\times k}$, $\boldsymbol{B} \in \mathbb{R}^{3\times n}$ $(k \leq n)$ be coordinates of molecules. Suppose that $\boldsymbol{A}$ is included in a ball centered at c with radius ℓ.
If $\mathrm{RMSD}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{R}, \boldsymbol{v}, \sigma) \leq r$ and $i \in \{\sigma(j)\}_{j=1}^{k}$ *holds, then*
$\boldsymbol{b}_{\sigma(1)}, \ldots, \boldsymbol{b}_{\sigma(k)}$ are contained within the ball centered at $\boldsymbol{b}_i$ with radius $2\left(k^{\frac{1}{2}}r + \ell\right)$.

Given $\{\sigma | \sigma(1) = k_1, \ldots, \sigma(l) = k_l\}$, we define $\underline{E}$ as follows:

$$\underline{E} = \min_{\boldsymbol{R},\boldsymbol{v}} \sqrt{\frac{1}{k}\sum_{i=1}^{l} ||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_{k_i} - \boldsymbol{v}||^2}. \tag{30}$$

The specific method is given by Algorithm 2.

5. New G-RMSD methods for exact solution

We have developed new G-RMSD methods for obtaining the exact solution of G-RMSD (or GS) index and GS_{sub} -index, called IsometryOpt, MatchFPT-delta, and IsometrySearch. IsometryOpt can be applied to both, measuring the similarity between equal size of molecules and substructure-matching between different size of molecules. MatchFPT-delta is a variant version of MatchFPT, which is for substructure enumeration. IsometrySearch is also for substructure enumeration. The details of these new methods are described in the following sections.

5.1 IsometryOpt to determine similarity between molecules of equal size

In this section, we consider the exact solution of eq. (7). Since there is no need to consider translation when the number of atoms is the same, we consider rotation and mirroring among the isometric transformations. Therefore, we can solve the problem using BnB for orthogonal matrices. Let D be a cube with center $\boldsymbol{r}_0$ and side length 2 η, and we assume that the orthogonal matrix is represented by $\boldsymbol{r}$ contained in D. We define $\overline{E}$ and $\underline{E}$ as follows:

$$\overline{E} = \min_{\sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k} ||\boldsymbol{a}_i - \boldsymbol{R}_{r_0}\boldsymbol{b}_{\sigma(i)}||^2} \tag{31}$$

$$\underline{E} = \min_{\sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k} \max\left(||\boldsymbol{a}_i - \boldsymbol{R}_{r_0}\boldsymbol{b}_{\sigma(i)}|| - \xi_i, 0\right)^2}. \tag{32}$$

For a positive integer K, let σ_K denote the K -th best matching of eq. (32) and $\underline{E}_K$ denote its weight. In particular, we have $\underline{E}_1 = \underline{E}$. We can calculate the K -th best matching in $O(Kn^3)$ time.[31] A specific method is given by Algorithm 3, and the conceptual flow is shown in Fig. 3. The details of this method are as follows.

5.1.1 Integration with the AO Method. When $\overline{E} \leq E^*$, there is a better solution than the tentative solution in the space we are considering. In lines 8 to 11 of algorithm 1, we use the AO method from the $\boldsymbol{R}$ obtained when finding $\overline{E}$. The speed at which the tentative solution becomes the global optimum is improved by using the AO method.

5.1.2 Use of $\boldsymbol{K}$ -th best matching. The algorithm for finding the K -th smallest weight maximum matching is used in lines 12 to 23 of Algorithm 3. Here, the *bound* of BnB is performed. We show the proof for this *bound* in the online supplementary material.

5.2 IsometryOpt to determine substructure-match between molecules of different size

G-RMSD for molecules of different size to find the best substructure match can be solved by using BnB for orthogonal matrices and translations. Recall the assumptions that D is a cube with center $\boldsymbol{r}_0$ and side length 2 η, the orthogonal matrix is represented by $\boldsymbol{r}$ contained in D, C is a cube with center $\boldsymbol{v}_0$ and side length γ, and the translation vector is contained in C.

Algorithm 3: IsometryOpt for molecule with same size

Input: Coordinate data of two molecules $\boldsymbol{A}$, $\boldsymbol{B}$.
Output: Minimum value E^* of RMSD and the corresponding $\boldsymbol{R}^*$, σ^*.
1. Put two cubes with center at origin and side 2π into a priority queue Q.
2. Let $E^* = \infty$.
3. **while** Q is not empty **do**
4. Read out a cube D with lowest lower-bound from Q.
5. Divide D into eight sub-cubes $\mathfrak{D}$.
6. **for each** D in $\mathfrak{D}$ **do**
7. Compute $\bar{E}$.
8. **if** $\bar{E} < E^*$ **then**
9. Run the AO method from the center of D.
10. Update E^*, $\boldsymbol{R}^*$, σ^* with the result of AO.
11. **end if**
12. Let 2η be the side of D and compute $k := \log_2 \frac{\pi}{\eta}$.
13. **for** $i \in \{1, \ldots, k\}$ **do**
14. Compute $\underline{E}_i$, σ_i
15. **if** $\underline{E}_i < E^*$ **then**
16. With σ_i fixed, compute minimum value of RMSD E_σ.
17. **if** $E_\sigma < E^*$ **then**
18. Update E^*, $\boldsymbol{R}^*$, σ^* with E_σ.
19. **end if**
20. **else**
21. Go to line 6.
22. **end if**
23. **end for**
24. Put D into Q.
25. **end for**
26. **end while**

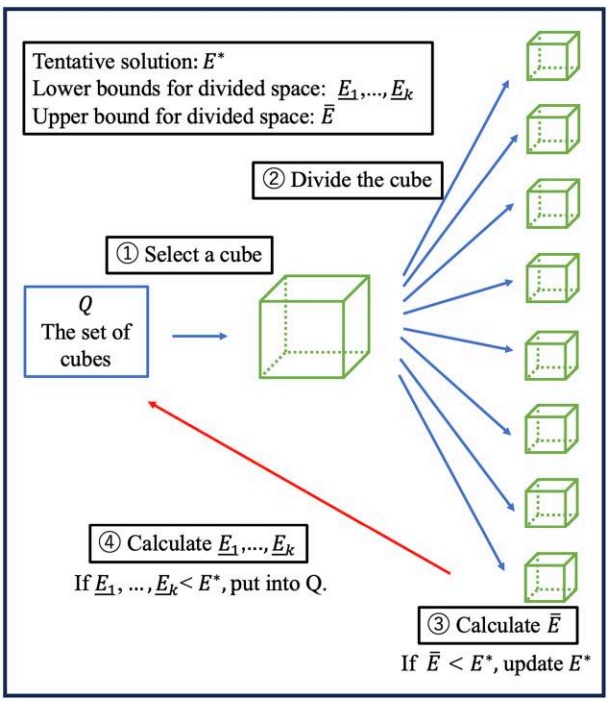


Fig. 3. Conceptual flow of IsometryOpt method for molecules of equal size (Section 5.1). A cube represents the subspace of rotation matrix. IsometryOpt for substructure match between molecules of different size (Section 5.2). The IsometrySearch (Section 5.4), which is for substructure enumeration, also follows this flow.

Algorithm 4: IsometryOpt for molecule with different size

Input: Coordinate data of two molecules $\boldsymbol{A}$, $\boldsymbol{B}$.
Output: Minimum value E^* of RMSD and the corresponding $\boldsymbol{R}^*$, $\boldsymbol{v}^*$, σ^*.
1. Put C_0 defined in Sect. 2.5 into the priority queue Q, and let ς_0 be the side of C_0.
2. Let $E^* = \infty$.
3. **while** Q is not empty **do**
4. Read out a cube C with lowest lower-bound from Q.
5. Divide C into eight sub-cubes $\mathcal{C}$.
6. **for each** C in $\mathcal{C}$ **do**
7. Let ς be the side of C and compute $k := \log_2 \frac{\varsigma_0}{\varsigma}$.
8. Compute $\hat{\underline{E}}_l$ by Algorithm 4 with k ($\leq l$).
9. **if** $\hat{\underline{E}}_l < E^*$ **then**
10. Put C into Q.
11. **end if**
12. **end for**
13. **end while**

Then $\overline{E}$, $\underline{E}$, and $\hat{\underline{E}}$ are defined as follows:

$$\overline{E} = \min_{\sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k} \max\left(||\boldsymbol{a}_i - \boldsymbol{R}_{r_0}\boldsymbol{b}_{\sigma(i)} - \boldsymbol{v}_0|| - \gamma, 0\right)^2}, \quad (33)$$

$$\underline{E} = \min_{\sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k} \max\left(||\boldsymbol{a}_i - \boldsymbol{R}_{r_0}\boldsymbol{b}_{\sigma(i)} - \boldsymbol{v}_0|| - \gamma - \xi_i, 0\right)^2}, \quad (34)$$

$$\hat{\underline{E}} = \min_{\sigma,\boldsymbol{r}} \sqrt{\frac{1}{k}\sum_{i=1}^{k} \max\left(||\boldsymbol{a}_i - \boldsymbol{R}_{\boldsymbol{r}}\boldsymbol{b}_{\sigma(i)} - \boldsymbol{v}_0|| - \gamma, 0\right)^2}. \quad (35)$$

For a positive integer K, let $\bar{E}_K$, $\underline{E}_K$, $\hat{\underline{E}}_K$, and σ_K denote the weights of the K -th best matching of eq. (33), eq. (34), eq. (35), and the K -th best matching of eq. (33) respectively. A specific method is given by Algorithm 4.

The entire search space consists of pairs of orthogonal matrices and translations. Thus, it has six dimensions. In order to avoid searching in such a high dimensional space, we apply BnB in two steps. When we perform BnB on the space of parallel translations, we need to find $\hat{\underline{E}}$. For this purpose, we perform BnB on the space of orthogonal matrices to obtain $\hat{\underline{E}}$.

In the two-step BnB, the K -th best matching can be used as in Algorithm 5. We show the proof to guarantee the method in the online supplementary material.

5.3 MatchFPT-delta for substructure enumeration between molecules of different size

To speed up the algorithm, we consider the following bound methods. Specifically, eq. (3) can be solved by Algorithm 6.

5.3.1 Bound method 1. Suppose that A is included in a ball centered at c with radius ℓ. From Theorem 1, when $\{\sigma|\sigma(1) = k_1, \ldots, \sigma(l) = k_l\}$ is given, we only need to consider σ such that holds for $i = 1, \ldots, l, j = l+1, \ldots, k$. A graphical image of the theorem is shown in Fig. 4.

Algorithm 5: BnB on orthogonal matrices

Input: Coordinate data of two molecules $\boldsymbol{A}$, $\boldsymbol{B}$, an integer k, tentative solution E^*, the space of translation C_v.
Output: Tentative solution E^*, $\hat{\underline{E}}_l^*$ $(k \leq l)$.

1. Let $\hat{\underline{E}}_l^* = \infty$.
2. Put two cubes with center at origin and side 2π into a priority queue Q.
3. **while** Q is not empty **do**
4. Read out a cube D with lowest lower-bound from Q.
5. Divide D into eight sub-cubes $\mathfrak{D}$.
6. **for each** D in $\mathfrak{D}$ **do**
7. Compute $\bar{E}_K$.
8. **if** $\bar{E}_K < E^*$ **then**
9. $\hat{\underline{E}}_l^* = \bar{E}_K$
10. **end if**
11. Let 2η be the side of D, and $K := k + \log_2 \frac{\pi}{\eta}$.
12. **for** $i \in \{1, \ldots, K\}$ **do**
13. Compute $\underline{E}_i$, σ_i.
14. **if** $\underline{E}_i < \hat{\underline{E}}_l^*$ **then**
15. With σ_i fixed, compute E_σ.
16. **if** $E_\sigma < E^*$ **then**
17. Update E^*, $\boldsymbol{R}^*$, σ^* with E_σ.
18. **end if**
19. **else**
20. Go to line 3.
21. **end if**
22. **end for**
23. Put D into Q.
24. **end for**
25. **end while**

Algorithm 6: MatchFPT-delta

Input: Coordinate data of two molecules $\boldsymbol{A}$, $\boldsymbol{B}$, and ε which determines the substructure.
Output: $\boldsymbol{R}^*$, $\boldsymbol{v}^*$, σ^* which represent substructures.

1. Put the set of injections from $\{1, \ldots, k\}$ to $\{1, \ldots, n\}$ into a set Q.
2. Let $E^* = \infty$.
3. **while** Q is not empty **do**
4. Read out C we put most recently from Q.
5. Divide C into $\mathcal{C}$ based on Sect. 2.7.
6. **for each** C in $\mathcal{C}$ **do**
7. Compute $\underline{E}$.
8. **if** $\underline{E} < \varepsilon$ **then**
9. **if** $|C| = k$ **then**
10. σ corresponding to $\underline{E}$ is one of the solution.
11. **else**
12. Limit C by bound method 1 and 2. Put C into Q.
13. **end if**
14. **end if**
15. **end for**
16. **end while**

5.3.2 Bound method 2. We can use the following Theorem 2.

Theorem 2 *Suppose*

$$\sum_{i=1}^{l} \boldsymbol{a}_i = 0, \ \sum_{i=1}^{l} \boldsymbol{b}_{\sigma(i)} = 0. \tag{36}$$

If

$$\sqrt{\frac{1}{k}\sum_{i=1}^{l+1} ||\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{b}_{\sigma(i)} - \boldsymbol{v}||^2} \leq \varepsilon, \tag{37}$$

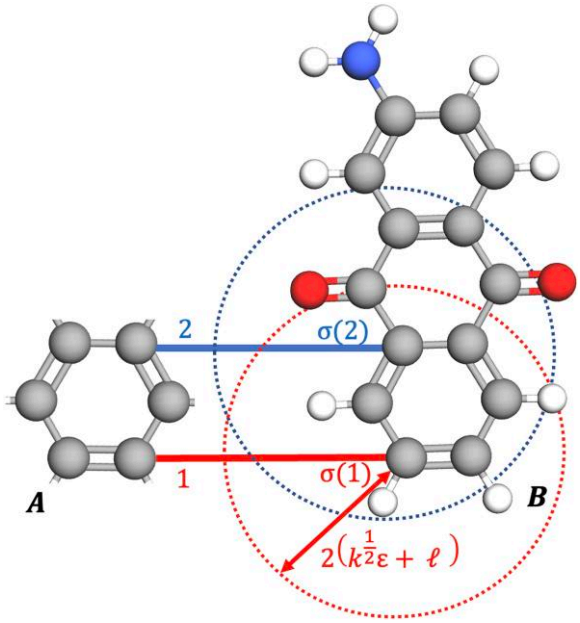


Fig. 4. Graphical image of Theorem 1 and Bound Method 1. Suppose that $\sigma(1)$ and $\sigma(2)$ is fixed as in the figure. In MatchFPT, the candidate for pairing is limited in the red circle. In MatchFPT-delta, the candidate for pairing is limited in the intersection between a red circle and a blue circle. As new pairs are determined, the search area becomes smaller. This process is repeated for all possible atom pairs between A and B in the implementation. In this case, two benzene rings are enumerated in B.

then,

$$\frac{l}{l+1}(|a_{l+1}| - |b_{\sigma(l+1)}|)^2 + \sum_{i=1}^{l}(|\boldsymbol{a}_i| - |\boldsymbol{b}_{\sigma(i)}|)^2 \leq k\varepsilon^2 \tag{38}$$

holds.

The proof is described in the online supplementary material. We can suppose $\sum_{i=1}^{l} \boldsymbol{a}_i = 0$ and $\sum_{i=1}^{l} \boldsymbol{b}_{\sigma(i)} = 0$ by translating $\boldsymbol{A}' = \boldsymbol{A} - \frac{1}{l}\sum_{i=1}^{l} \boldsymbol{a}_i$ and $\boldsymbol{B}' = \boldsymbol{B} - \frac{1}{l}\sum_{i=1}^{l} \boldsymbol{b}_{\sigma(i)}$, respectively.

5.4 IsometrySearch for substructure enumeration between molecules of different size

We consider Theorem 3 to reduce the problem to the minimization for orthogonal matrices.

Theorem 3 *If*

$$\sqrt{\frac{1}{k}\sum_{i=1}^{k} ||\boldsymbol{R}(\boldsymbol{a}_i - \boldsymbol{a}_j) + \boldsymbol{b}_{\sigma(j)} - \boldsymbol{b}_{\sigma(i)}||^2} \leq \varepsilon, \tag{39}$$

for some j, the following holds:

$$\sqrt{\frac{1}{k}\sum_{i=1}^{k} ||\boldsymbol{R}(\boldsymbol{a}_i - \boldsymbol{a}_j) + \boldsymbol{b}_{\sigma(j)} - \boldsymbol{b}_{\sigma(i)}||^2} \leq \sqrt{2}\varepsilon. \tag{40}$$

The proof is described in the online supplementary material. By Theorem 3, under the assumption that $\sigma(i) = j$, we can perform BnB for orthogonal matrices to search for substructures.

Algorithm 7: IsometrySearch

Input: Coordinate data of two molecules $\boldsymbol{A}$, $\boldsymbol{B}$, and ε which determines the substructure.
Output: $\boldsymbol{R}^*$, $\boldsymbol{v}^*$, σ^* which represent substructures.

```
1.  for (i, j) ∈ {1, ..., k} × {1, ..., n} do
2.    Suppose σ(i) = j.
3.    Limit B by Theorem 3.
4.    Put two cubes with center at origin and side 2π into a priority
      queue Q.
5.    Let E* = ∞.
6.    while Q is not empty do
7.      Read out a cube D with lowest lower-bound from Q.
8.      Divide D into eight sub-cubes 𝔇.
9.      for each D in 𝔇 do
10.       Let 2η be the side of D and compute k := log_2 π/η .
11.       for i ∈ {1, ..., k} do
12.         Compute E_i, σ_i
13.         if E_i < √2ε then
14.           With σ_i fixed, compute minimum value of RMSD E_σ.
15.           if E_σ < E* then
16.             Update E*, R*, σ* with E_σ.
17.           end if
18.         else
19.           Go to line 6.
20.         end if
21.       end for
22.         Put D into Q
23.       end for
24.     end while
25.   end for
```

We define $\bar{E}$ and $\underline{E}$ as follows, when $\sigma(i) = j$, D is a cube with center $\boldsymbol{r}_0$ and side length 2η, and the orthogonal matrix is represented by $\boldsymbol{r}$ contained in D:

$$\overline{E} = \min_{\sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k} ||\boldsymbol{R}(\boldsymbol{a}_i - \boldsymbol{a}_j) + \boldsymbol{b}_{\sigma(j)} - \boldsymbol{b}_{\sigma(i)}||^2}. \tag{41}$$

$$\underline{E} = \min_{\sigma} \sqrt{\frac{1}{k}\sum_{i=1}^{k} \max\left(||\boldsymbol{R}(\boldsymbol{a}_i - \boldsymbol{a}_j) + \boldsymbol{b}_{\sigma(j)} - \boldsymbol{b}_{\sigma(i)}|| - \xi_i, 0\right)^2}. \tag{42}$$

Let $\underline{E}_K$ be the weight of the K -th best matching of eq. (42). Thus, a specific method is given by the Algorithm 7.

6. Computer experiments

We have performed computer experiments of the methods described in Sections 3, 4, and 5 and compared their output and computational time. Since the methods for the exact solutions described in Sections 4 and 5 are mathematically proven to give the exact solutions, we consider the results from these methods as correct and evaluate the other methods. In Sections 6.1.2 and 6.2.2, we use 3D molecular data from RMapDB,[22] a database of reaction pathways obtained by quantum mechanical calculations. In Section 6.3.2, we use 3D molecular data of NCI's compound database.[32]

The computer used in these experiments is a dual-core Intel Core i5 2.3 GHz CPU with 8 GB of memory. The programs of the new methods are written in C++ and compiled with the optimization option -O3. For this implementation, we used the Eigen package[33] and the Hungarian method.[34]

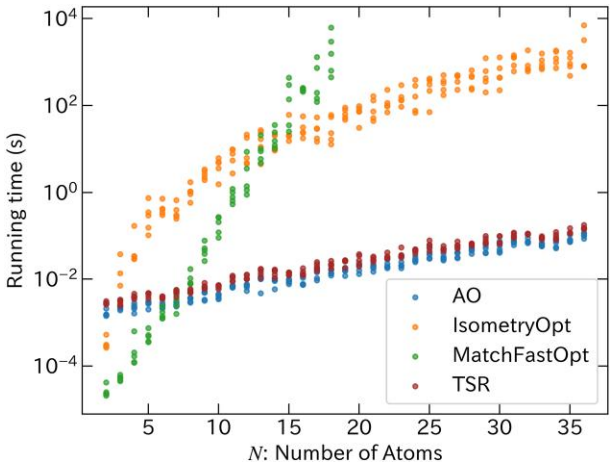


Fig. 5. Running time of AO, TSR, IsometryOpt, and MatchFastOpt for calculations of GS-index between two molecules of equal size of N atoms, of which coordinates are uniformly random.

In the following, the error of the output value is defined by

$$\text{error} = \frac{(\text{output value}) - (\text{global minimum})}{(\text{global minimum})} \tag{43}$$

where the results from our new proposed methods are used as the global minimum, because it is mathematically proven that their output is the optimal solution, as mentioned above.

6.1 Similarity between molecules of equal size

6.1.1 Experiment with artificial data. We assume that each component of the 3D molecular data $\boldsymbol{A}$ and $\boldsymbol{B}$ follows a uniform distribution of [0, 1], i.e. the coordinates of atoms were generated uniformly in random. GS-index was calculated between two molecules A and B with N atoms by AO (Section 3.2.1), TSR (Section 3.3), IsometryOpt (Section 5.1), and MatchFastOpt (Section 4.1). Five calculations per method were performed for each N. The calculations used only the Cartesian coordinates of atoms. No bond information was in the input.

Figure 5 plots the running time against the number of atoms. As the number of atoms increases, all methods require more computational time, but the rates of increase are different. MatchFastOpt shows the highest increment rate, while the rate of IsometryOpt decelerates as the molecular size rises, resulting in a smoother curve. This is because IsometryOpt performs BnB on a 3D space, while MatchFastOpt performs BnB on a $n!/k!$ dimensional space. AO and TSR are faster than IsometryOpt and MatchFastOpt as expected and show a flatter increase in computational time in relation to molecular size.

Figure 6 shows the average error values of GS-index obtained by AO and TSR. Since IsometryOpt and MatchFastOpt are proven to give the exact value, we used these methods as the benchmark, i.e. the global minimum in eq. (43), and determined the error values by calculating the discrepancy between the outcomes of them and AO or TSR. One hundred calculations per method were performed for each N.

It shows that as the number of atoms increases, the average error values increase. The GS-index value is so sensitive against the difference of geometries that the similarity between

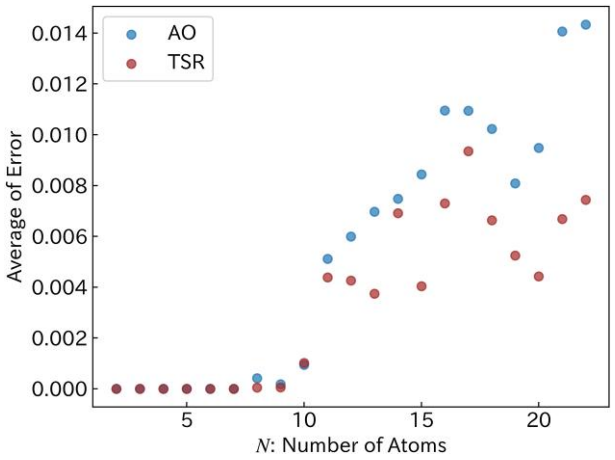


Fig. 6. The average errors of GS-index values from AO and TSR for two molecules with the same number of atoms *N*, of which coordinates are uniformly random.

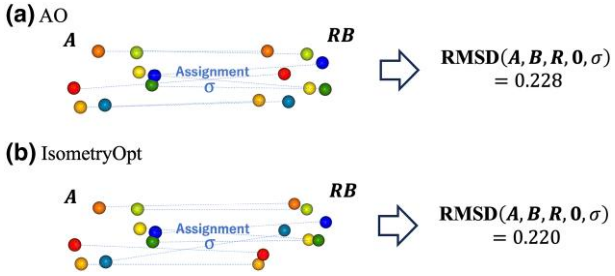


Fig. 7. An example where the AO method failed to find the optimal solution for the comparison between molecules of equal size. a) Solution of $\boldsymbol{R}$ and σ obtained by AO method. b) Exact solution of $\boldsymbol{R}$ and σ obtained by IsometryOpt method. $\boldsymbol{A}$ denotes the coordinates of molecules *A*. $\boldsymbol{RB}$ denotes the coordinates of rotated molecules *B*. The color of the sphere indicates assignments σ.

similar geometries is discussed between two and three decimal places. Therefore, this range of error may lead to incorrect conclusions. The results suggest that an algorithm for finding the global minimum, like IsometryOpt or MatchFastOpt, should be used, especially for large molecules if rigorous discussions with exact values are necessary. The results also show that TSR gives smaller errors than AO.

Refer to the previous paper[24] as well as its supporting information for the relationships between the GS-index values and the difference of geometries.

Figure 7 and Table 1 show an instance where AO failed to output the optimal solution. The optimization problem of these cases is so complex that it is difficult to reach the optimal solution only by a simple alternating optimization method.

6.1.2 Experiment with chemical data. Another assessment was carried out using chemical data as well. We calculated GS-index between one $^4C_1$ conformer (**1a**, Fig. 8b) of α-D-glucose (**1**, Fig. 8a) and the other 651 conformers, which are the same data used in the previous paper.[24] The calculations used only the Cartesian coordinates of atoms. Neither bond information nor atom types were taken into account.

The accuracy of AO and TSR was assessed by comparing their outcome to that of IsometryOpt. AO and TSR failed to reach accurate values for 43 and 36 types of conformers, respectively. Figure 9 shows the distribution of error values of AO and TSR, where some cases show rather large error values.

Table 1. Coordinates and assignments of atoms of Fig. 7.

Molecule *A*		Molecule *B* after optimization ($\boldsymbol{RB}$)					
		(a) AO			(b) IsometryOpt		
i	$\boldsymbol{a}_i$	$\sigma(i)$	$\boldsymbol{Rb}_{\sigma(i)}$	Dev_i	$\sigma(i)$	$\boldsymbol{Rb}_{\sigma(i)}$	Dev_i
1	$\begin{pmatrix}0.47\\-0.08\\-0.21\end{pmatrix}$	2	$\begin{pmatrix}0.09\\0.07\\0.03\end{pmatrix}$	0.47	8	$\begin{pmatrix}0.36\\-0.25\\-0.24\end{pmatrix}$	0.20
2	$\begin{pmatrix}0.18\\0.36\\0.39\end{pmatrix}$	8	$\begin{pmatrix}0.30\\0.35\\0.37\end{pmatrix}$	0.12	5	$\begin{pmatrix}-0.03\\0.40\\0.34\end{pmatrix}$	0.22
3	$\begin{pmatrix}0.39\\-0.30\\0.45\end{pmatrix}$	4	$\begin{pmatrix}0.40\\-0.25\\0.49\end{pmatrix}$	0.07	4	$\begin{pmatrix}0.42\\-0.37\\0.36\end{pmatrix}$	0.12
4	$\begin{pmatrix}-0.30\\0.11\\0.08\end{pmatrix}$	3	$\begin{pmatrix}-0.21\\-0.10\\0.11\end{pmatrix}$	0.23	3	$\begin{pmatrix}-0.22\\-0.06\\0.16\end{pmatrix}$	0.20
5	$\begin{pmatrix}-0.27\\0.34\\-0.14\end{pmatrix}$	1	$\begin{pmatrix}-0.18\\0.31\\-0.27\end{pmatrix}$	0.17	1	$\begin{pmatrix}-0.19\\0.33\\-0.25\end{pmatrix}$	0.14
6	$\begin{pmatrix}-0.45\\-0.04\\-0.03\end{pmatrix}$	6	$\begin{pmatrix}-0.40\\-0.12\\-0.14\end{pmatrix}$	0.15	7	$\begin{pmatrix}-0.34\\-0.07\\-0.17\end{pmatrix}$	0.18
7	$\begin{pmatrix}0.10\\-0.28\\-0.17\end{pmatrix}$	5	$\begin{pmatrix}0.04\\-0.27\\-0.32\end{pmatrix}$	0.17	2	$\begin{pmatrix}0.09\\0.06\\0.01\end{pmatrix}$	0.38
8	$\begin{pmatrix}-0.48\\0.07\\-0.03\end{pmatrix}$	7	$\begin{pmatrix}-0.37\\0.21\\0.1\end{pmatrix}$	0.22	6	$\begin{pmatrix}-0.44\\0.16\\0.15\end{pmatrix}$	0.21

$\boldsymbol{a}_i$ represents the coordinates of the i-th atom of **1a**. $\boldsymbol{Rb\_AO}_{\sigma(i)}$ and $\boldsymbol{Rb\_Iso}_{\sigma(i)}$ are the coordinates of i-th atom of **1b** after the optimization against σ and $\boldsymbol{R}$ obtained by AO and IsometryOpt, respectively. Dev_i means the deviation $\|\boldsymbol{a}_i - \boldsymbol{Rb}_{\sigma(i)}\|$.

An instance where AO failed to reach the optimal solution is shown in Fig. 8c to e. This is the case where **1a** was compared with **1b**, which adopts a $^OS_2$ conformer (Fig. 8c). The results of AO and IsometryOpt are shown in Fig. 8d and 8e, respectively. The hydrogen atoms which were not matched to those with the same serial numbers of **1a** are marked in blue. The coordinates and assignments of atoms are shown in Table 2.

In this case, all of the carbon and oxygen atoms were matched between the same serial numbers both by AO and IsometryOpt. However, as shown in the last row of each of Table 2-(I) and 2-(II), the exact solution by IsometryOpt shows smaller RMSD values both for the carbon and oxygen atoms. All of the hydrogen atoms except for four atoms (H14, H15, H20, and H21) were matched to those of the same serial numbers of **1a** by IsometryOpt, while eight atoms (H13, H14, H15, H16, H20, H21, H22, and H23) were not matched to those of the same serial numbers of **1a** by AO. The RMSD values for hydrogen atoms were 1.27 and 0.84 by AO and IsometryOpt, respectively, which shows the largest difference between AO and IsometryOpt. The G-RMSD index by AO and IsometryOpt were 1.17 and 0.930, respectively (Fig. 9d to e).

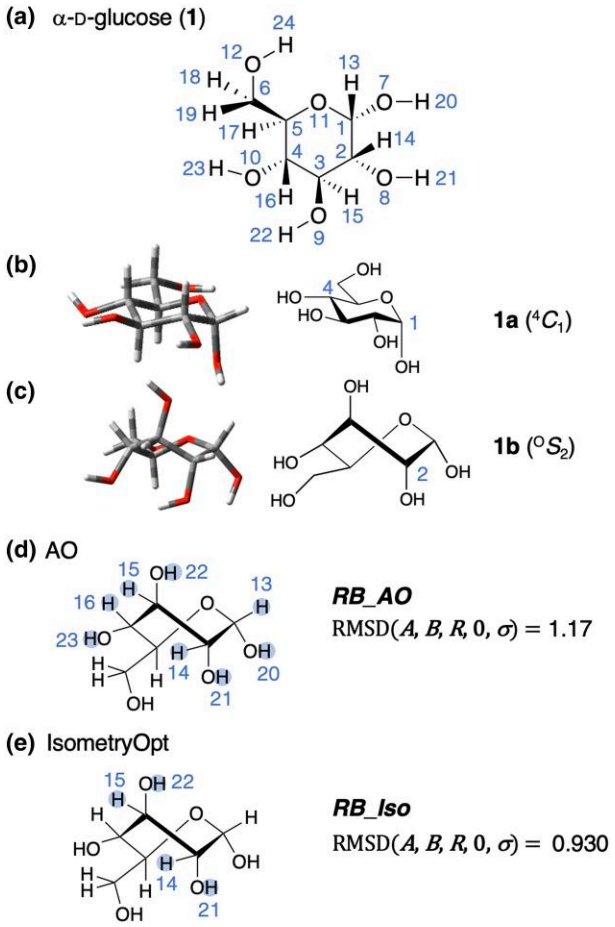


Fig. 8. An example where the AO method failed to find the optimal solution for the comparison between conformers of α -D-glucose (**1**). a) Structure of α -D-glucose (**1**) with serial numbers of atoms. These numbers correspond to those in Table 2. b) 3D structures of **1a**, which adopts a $^4C_1$ chair conformation. 651 conformers were compared with this structure. c) 3D structures of **1b**, which adopts a $^OS_2$ twist/skew-boat conformation. This is one of the structures where the AO method failed to reach the optimal solution. d) Result of optimization against $\boldsymbol{R}$ and σ by AO. $\boldsymbol{A}$ and $\boldsymbol{B}$ denote the coordinates of **1a** and **1b**, respectively. ***RB_AO*** denotes the optimized coordinates by the AO method. The hydrogen atoms marked in blue were not matched to those with the same serial numbers of **1a.** e) Result of optimization against $\boldsymbol{R}$ and σ by IsometryOpt. $\boldsymbol{A}$ and $\boldsymbol{B}$ denote the coordinates of **1a** and **1b**, respectively. ***RB_Iso*** denotes the optimized coordinates by the IsometryOpt method. The hydrogen atoms marked in blue were not matched to those with the same serial numbers of **1a.**

These results indicate certainly the necessity of an algorithm like IsometryOpt, which makes it possible to obtain the exact solution.

Regarding computational time, IsometryOpt took longer than AO and TSR as shown in Fig. 10. The running time using IsometryOpt shows a broader range in comparison to that using AO and TSR.

Note that this problem could not be solved with MatchFastOpt due to the long calculation time.

6.2 Substructure match between molecules of different size

6.2.1 Experiment with artificial data. We assume that each component of 3D coordinate data $\boldsymbol{A} \in \mathbb{R}^{3\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{3\times 2n}$ follows a uniform distribution of [0.1]. While changing N, the similarity between A and B was determined by AO (Section 3.2.2), IsometryOpt (Section 5.2), and MatchFastOpt (Section 4.1). Five calculations per method were performed for each N. To assess the error values, 100 calculations per method were carried out for each N. The calculations used only the Cartesian coordinates of atoms. No bond information was in the input.

Figure 11 plots the running time against the number of atoms. Compared to the case where the number of atoms is the same, the IsometryOpt method is slower due to the consideration of translation. For $2 \le N \le 15$, MatchFastOpt is faster than IsometryOpt. Theoretically, IsometryOpt should be faster when N becomes larger, because this method performs BnB in the 6D-space, while MatchFastOpt performs BnB on a $n!/k!$ dimensional space. However, with this problem setting, we could not confirm this trend due to a limitation of computer resource.

Figure 12 shows the average error values of AO. In the applications to molecules of different size, the initial settings of AO for rotation ($\boldsymbol{R}_0$) and translation ($\boldsymbol{v}_0$) are 120 and k times n, respectively, where k and n are the number of atoms of each molecule. This initial setting covers more exploration points compared to the applications to molecules of equal size. As the number N increases, the number of these initial values become larger. Thus, unlike the case of the same number of atoms, AO could obtain the exact value in most cases. Figure 13 and Table 3 show an instance where AO failed to reach the optimal solution.

6.2.2 Experiment with chemical data. As in the case of molecules of equal size, assessment using chemical data was also carried out. We calculated GS-index between CH_3 cation and 689 isomers of $C_2H_4O_2$, which were the same dataset used in the previous study.[24] The calculations used only the Cartesian coordinates of atoms. No bond information was taken into account.

The accuracy of AO was assessed by comparing its outcome to that of IsometryOpt. In this case, AO reached the optimal solution.

Regarding computational time, the fastest method was MatchFastOpt, followed by AO and IsometryOpt as shown in Fig. 14.

6.3 Substructure enumeration between molecules of different size

6.3.1 Experiment with artificial data. We assessed the performance of three methods for substructure enumeration, MatchFPT (Section 4.2), MatchFPT-delta (Section 5.3), and IsometrySearch (Section 5.4) using artificial data. We defined the artificial 3D data of two molecules A and B as $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N]$, $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{2N}]$, where $\boldsymbol{a}_1, \ldots, \boldsymbol{b}_N$ follow a uniform random number $[0, 1]^3$ and $\boldsymbol{b}_{N+1}, \ldots, \boldsymbol{b}_{2N}$ are obtained from $\boldsymbol{A}$ by applying a random rotation and translation with a vector uniformly distributed in $[-1, 1]^3$, and then adding a uniformly distributed random error in the range $\left[-0.03/\sqrt{3}, 0.03/\sqrt{3}\right]^3$.

Using this dataset, we calculated the GS_{sub} -index if B contains A with $\varepsilon = 0.03$. Five calculations per method were performed for each N. The calculations used only the Cartesian coordinates of atoms. No bond information was in the input.

Figure 15 plots the running time against N, i.e. the number of atoms of A. The fastest method was MatchFPT-delta, followed by MatchFPT and IsometrySearch. MatchFPT-delta

and MatchFPT are significantly faster than IsometrySearch. However, since IsometrySearch performs BnB in the 3D-space, it is expected that IsometrySearch becomes faster than the others for larger N. But we could not conduct calculations with IsometrySearch for $N > 21$, because it was too large to perform in this setting.

6.3.2 Experiment with chemical data. The performance of the three methods (MatchFPT, MatchFPT-delta, and IsometrySearch) was assessed also using chemical data. We used a dataset of 25,052 3D-molecular data obtained from the NCI Open Database Compounds.[32] Since the aim is to assess their performance, we chose a benzene ring as a simple

Table 2. Coordinates and assignments of atoms of Fig. 8.

(I) C atoms

1a		Optimized 1b					
		(a) AO			(b) IsometryOpt		
i	$\boldsymbol{a}_i$	$\sigma(i)$	$\boldsymbol{Rb\_AO}_{\sigma(i)}$	Dev_i	$\sigma(i)$	$\boldsymbol{Rb\_Iso}_{\sigma(i)}$	Dev_i
1	$\begin{pmatrix}0.48\\-1.58\\-0.27\end{pmatrix}$	1	$\begin{pmatrix}0.68\\-1.26\\-1.03\end{pmatrix}$	0.85	1	$\begin{pmatrix}0.75\\-1.54\\-0.20\end{pmatrix}$	0.28
2	$\begin{pmatrix}1.53\\-0.51\\-0.56\end{pmatrix}$	2	$\begin{pmatrix}1.66\\-0.40\\-0.22\end{pmatrix}$	0.38	2	$\begin{pmatrix}1.67\\-0.37\\0.18\end{pmatrix}$	0.76
3	$\begin{pmatrix}1.19\\0.78\\0.16\end{pmatrix}$	3	$\begin{pmatrix}1.24\\1.06\\-0.26\end{pmatrix}$	0.51	3	$\begin{pmatrix}1.27\\0.88\\-0.59\end{pmatrix}$	0.76
4	$\begin{pmatrix}-0.22\\1.24\\-0.17\end{pmatrix}$	4	$\begin{pmatrix}-0.14\\1.24\\0.39\end{pmatrix}$	0.57	4	$\begin{pmatrix}-0.16\\1.31\\-0.21\end{pmatrix}$	0.10
5	$\begin{pmatrix}-1.21\\-0.13\\0.17\end{pmatrix}$	5	$\begin{pmatrix}-1.04\\0.01\\0.20\end{pmatrix}$	0.21	5	$\begin{pmatrix}-1.06\\0.11\\0.13\end{pmatrix}$	0.16
6	$\begin{pmatrix}-2.63\\0.38\\-0.30\end{pmatrix}$	6	$\begin{pmatrix}-2.49\\0.36\\-0.03\end{pmatrix}$	0.30	6	$\begin{pmatrix}-2.48\\0.25\\-0.36\end{pmatrix}$	0.20
$\sqrt{\frac{1}{6}\sum_i \mathrm{Dev}_i^2}$				0.51			0.47

(II) O atoms

1a		Optimized 1b					
		(a) AO			(b) IsometryOpt		
i	$\boldsymbol{a}_i$	$\sigma(i)$	$\boldsymbol{Rb\_AO}_{\sigma(i)}$	Dev_i	$\sigma(i)$	$\boldsymbol{Rb\_Iso}_{\sigma(i)}$	Dev_i
7	$\begin{pmatrix}0.53\\-1.96\\1.09\end{pmatrix}$	7	$\begin{pmatrix}0.72\\-2.59\\-0.59\end{pmatrix}$	1.80	7	$\begin{pmatrix}0.74\\-2.50\\0.82\end{pmatrix}$	0.63
8	$\begin{pmatrix}2.77\\-1.04\\-0.09\end{pmatrix}$	8	$\begin{pmatrix}1.69\\-0.88\\1.13\end{pmatrix}$	1.64	8	$\begin{pmatrix}1.57\\-0.15\\1.60\end{pmatrix}$	2.26
9	$\begin{pmatrix}2.16\\1.73\\-0.25\end{pmatrix}$	9	$\begin{pmatrix}1.24\\1.42\\-1.64\end{pmatrix}$	1.70	9	$\begin{pmatrix}1.39\\0.55\\-1.96\end{pmatrix}$	2.22
10	$\begin{pmatrix}-0.43\\2.41\\0.62\end{pmatrix}$	10	$\begin{pmatrix}0.06\\1.36\\1.82\end{pmatrix}$	1.66	10	$\begin{pmatrix}-0.09\\2.1\\0.99\end{pmatrix}$	0.59
11	$\begin{pmatrix}-0.80\\-1.06\\-0.55\end{pmatrix}$	11	$\begin{pmatrix}-0.64\\-0.72\\-0.96\end{pmatrix}$	0.56	11	$\begin{pmatrix}-0.56\\-1.07\\-0.51\end{pmatrix}$	0.24
12	$\begin{pmatrix}-3.47\\-0.72\\0.03\end{pmatrix}$	12	$\begin{pmatrix}-3.27\\-0.83\\-0.16\end{pmatrix}$	0.30	12	$\begin{pmatrix}-3.26\\-0.87\\0.02\end{pmatrix}$	0.27
$\sqrt{\frac{1}{6}\sum_i \mathrm{Dev}_i^2}$				1.41			1.35

(*continued*)

(III) H atoms							
1a		Optimized 1b					
		(a) AO			(b) IsometryOpt		
i	$\boldsymbol{a}_i$	$\sigma(i)$	$\boldsymbol{Rb\_AO}_{\sigma(i)}$	Dev_i	$\sigma(i)$	$\boldsymbol{Rb\_Iso}_{\sigma(i)}$	Dev_i
13	$\begin{pmatrix}0.61\\-2.44\\-0.92\end{pmatrix}$	20	$\begin{pmatrix}0.91\\-2.66\\0.34\end{pmatrix}$	1.31	13	$\begin{pmatrix}1.08\\-2.03\\-1.09\end{pmatrix}$	0.65
14	$\begin{pmatrix}1.57\\-0.32\\-1.62\end{pmatrix}$	13	$\begin{pmatrix}0.92\\-1.26\\-2.07\end{pmatrix}$	1.22	22	$\begin{pmatrix}1.38\\1.29\\-2.56\end{pmatrix}$	1.88
15	$\begin{pmatrix}1.25\\0.62\\1.23.\end{pmatrix}$	23	$\begin{pmatrix}0.13\\2.25\\2.15\end{pmatrix}$	2.18	21	$\begin{pmatrix}1.17\\0.68\\1.85\end{pmatrix}$	0.63
16	$\begin{pmatrix}-0.28\\1.48\\-1.23\end{pmatrix}$	22	$\begin{pmatrix}1.20\\2.36\\-1.81\end{pmatrix}$	1.81	16	$\begin{pmatrix}-0.58\\1.89\\-1.02\end{pmatrix}$	0.55
17	$\begin{pmatrix}-1.20\\-0.07\\1.23\end{pmatrix}$	17	$\begin{pmatrix}-0.95\\-0.61\\1.08\end{pmatrix}$	0.61	17	$\begin{pmatrix}-1.06\\-0.02\\1.20\end{pmatrix}$	0.15
18	$\begin{pmatrix}-3.05\\1.24\\0.20\end{pmatrix}$	18	$\begin{pmatrix}-2.90\\0.90\\0.81\end{pmatrix}$	0.72	18	$\begin{pmatrix}-2.95\\1.12\\0.08\end{pmatrix}$	0.20
19	$\begin{pmatrix}-2.63\\0.55\\-1.37\end{pmatrix}$	19	$\begin{pmatrix}-2.58\\0.97\\-0.92\end{pmatrix}$	0.61	19	$\begin{pmatrix}-2.48\\0.37\\-1.44\end{pmatrix}$	0.25
20	$\begin{pmatrix}1.42\\-2.15\\1.37\end{pmatrix}$	21	$\begin{pmatrix}1.34\\-0.28\\1.78\end{pmatrix}$	1.92	20	$\begin{pmatrix}0.84\\-2.11\\1.69\end{pmatrix}$	0.67
21	$\begin{pmatrix}3.49\\-0.41\\-0.19\end{pmatrix}$	14	$\begin{pmatrix}2.65\\-0.52\\-0.63\end{pmatrix}$	0.95	14	$\begin{pmatrix}2.69\\-0.63\\-0.04\end{pmatrix}$	0.84
22	$\begin{pmatrix}2.04\\2.58\\0.18\end{pmatrix}$	15	$\begin{pmatrix}1.96\\1.66\\0.29\end{pmatrix}$	0.93	15	$\begin{pmatrix}1.94\\1.69\\-0.33\end{pmatrix}$	1.02
23	$\begin{pmatrix}-1.14\\2.97\\0.32\end{pmatrix}$	16	$\begin{pmatrix}-0.61\\2.13\\0.00\end{pmatrix}$	1.04	23	$\begin{pmatrix}-0.05\\3.04\\0.86\end{pmatrix}$	1.22
24	$\begin{pmatrix}-3.09\\-1.54\\-0.25\end{pmatrix}$	24	$\begin{pmatrix}-2.90\\-1.41\\-0.82\end{pmatrix}$	0.61	24	$\begin{pmatrix}-2.83\\-1.69\\-0.24\end{pmatrix}$	0.30
$\sqrt{\frac{1}{12}\sum_i \mathrm{Dev}_i^2}$				1.27			0.84

$\boldsymbol{a}_i$ represents the coordinates of the i-th atom of 1a. $\boldsymbol{Rb\_AO}_{\sigma(i)}$ and $\boldsymbol{Rb\_Iso}_{\sigma(i)}$ are the coordinates of i-th atom of **1b** after the optimization against σ and $\boldsymbol{R}$ obtained by AO and IsometryOpt, respectively. Dev_i means the deviation $\|\boldsymbol{a}_i - \boldsymbol{Rb}_{\sigma(i)}\|$.

query. Benzene is a planar molecule but the enumeration was carried out in 3D. In this dataset, 15,124 benzene rings were found with $\varepsilon = 0.1$Å.

Figure 16 plots the running time against the number of atoms of each molecule in the dataset. Since the query size is 6, the computation time of IsometrySearch is larger than that of other methods. However, since IsometrySearch performs BnB in the 3D-space, it is expected that IsometrySearch becomes faster than the others for larger query molecules.

The characteristics and performance of all the methods are summarized in Table 4.

7. Conclusion

We have developed three new exact optimization methods of G-RMSD to obtain global minimum solutions for measuring molecular similarity. These methods, IsometryOpt, IsometrySearch, and MatchFPT-delta, are all mathematically proven to give an exact G-RMSD index (GS-index) and $\mathrm{GS}_{\mathrm{sub}}$ -index, i.e. these methods can reach the global optimum solution. IsometryOpt employs BnB for isometric transformations. This method revealed that AO and TSR, which were the first methods of G-RMSD, may not reach the global optimal solution in all cases. In terms of computation time,

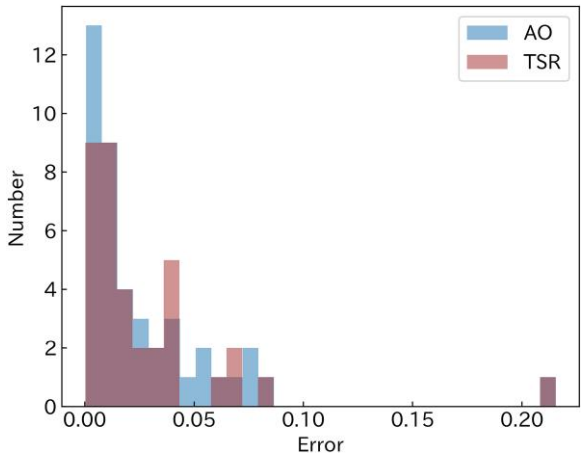


Fig. 9. Error values of G-RMSD index between one conformer and the other 651 conformers of α -glucose (**1**).

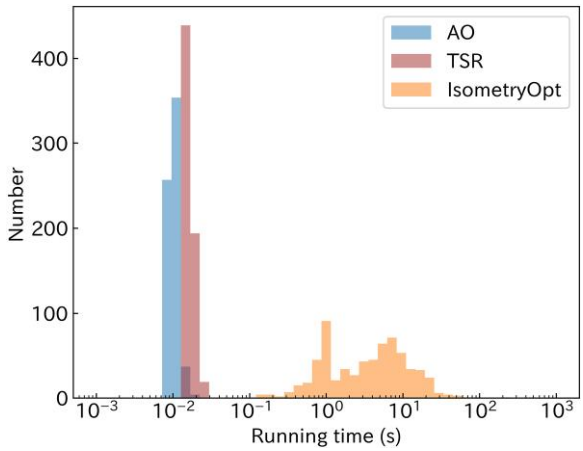


Fig. 10. Running time for the calculations of G-RMSD for 652 conformers of α -glucose.

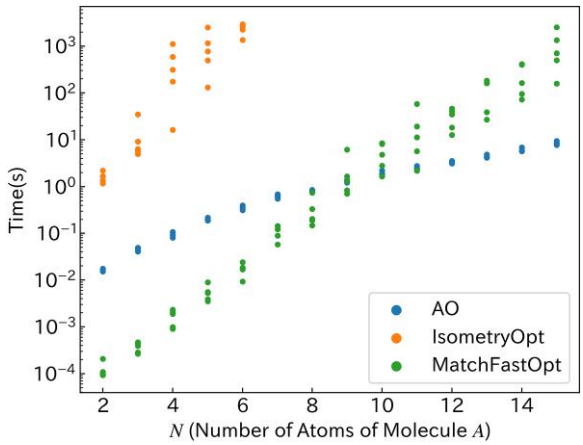


Fig. 11. Running time for substructure match between molecule A with N atoms and molecule B with $2N$ atoms.

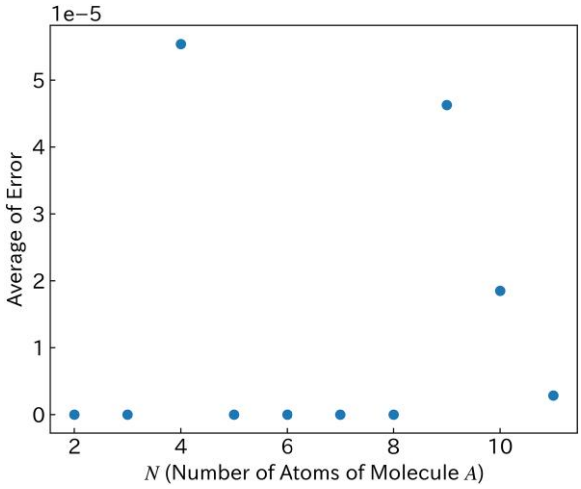


Fig. 12. Average error values of AO in the optimization between an N-atom molecule and $2N$ -atom molecule. For all N, 100 runs were performed.

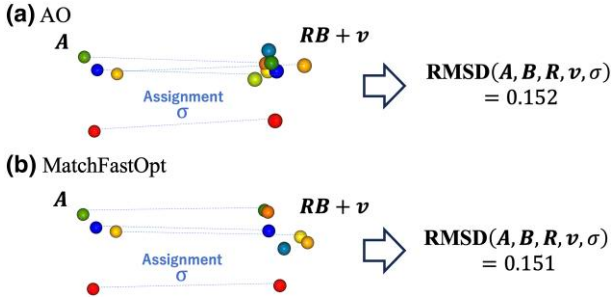


Fig. 13. An example where the AO method failed to find the optimal solution for molecules of different size. a) Solution of $\boldsymbol{R}$ and σ obtained by AO method. b) Exact solution of $\boldsymbol{R}$ and σ obtained by MatchFastOpt method. $\boldsymbol{A}$ denotes the coordinates of molecules A. $\boldsymbol{RB} + \boldsymbol{v}$ denotes the coordinates of rotated and translated molecules B. The color of the sphere indicates assignments σ.

Table 3. Coordinates and assignments of atoms of Fig. 13.

Molecule A		Molecule B after optimization ($\boldsymbol{RB}$)					
		(a) AO			(b) IsometryOpt		
i	$\boldsymbol{a}_i$	$\sigma(i)$	$\boldsymbol{Rb}_{\sigma(i)} + \boldsymbol{v}$	Dev_i	$\sigma(i)$	$\boldsymbol{Rb}_{\sigma(i)} + \boldsymbol{v}$	Dev_i
1	$\begin{pmatrix}-0.01\\-0.45\\0.36\end{pmatrix}$	2	$\begin{pmatrix}0.01\\-0.28\\0.26\end{pmatrix}$	0.20	1	$\begin{pmatrix}-0.06\\-0.37\\-0.27\end{pmatrix}$	0.13
2	$\begin{pmatrix}-0.29\\0.26\\-0.50\end{pmatrix}$	1	$\begin{pmatrix}-0.30\\0.30\\-0.53\end{pmatrix}$	0.05	2	$\begin{pmatrix}-0.38\\0.14\\-0.56\end{pmatrix}$	0.16
3	$\begin{pmatrix}0.11\\0.47\\-0.38\end{pmatrix}$	6	$\begin{pmatrix}0.05\\0.32\\-0.24\end{pmatrix}$	0.21	3	$\begin{pmatrix}0.13\\0.52\\-0.40\end{pmatrix}$	0.06
4	$\begin{pmatrix}-0.05\\0.31\\-0.12\end{pmatrix}$	7	$\begin{pmatrix}-0.00\\0.24\\-0.12\end{pmatrix}$	0.08	8	$\begin{pmatrix}-0.08\\0.28\\0.05\end{pmatrix}$	0.21

$\boldsymbol{a}_i$ represent the coordinates of the i-th atom of A, and $\boldsymbol{b}_{\sigma(i)}$ represent the coordinates of the $\sigma(i)$-th atom of B. σ, $\boldsymbol{R}$, and $\boldsymbol{v}$ represent the solution obtained by (a) the AO method and (b) the IsometryOpt method. Dev_i means the deviation $\|\boldsymbol{a}_i - \boldsymbol{Rb}_{\sigma(i)} - \boldsymbol{v}\|$.

MatchFastOpt, which was developed by Sasaki et al. is faster when the number of atoms is small. However, IsometryOpt performs better for equal size of molecules when the number of atoms is more than about 15. For substructure match between molecules with different number of atoms, within the scope of the computer experiments, MatchFastOpt performed

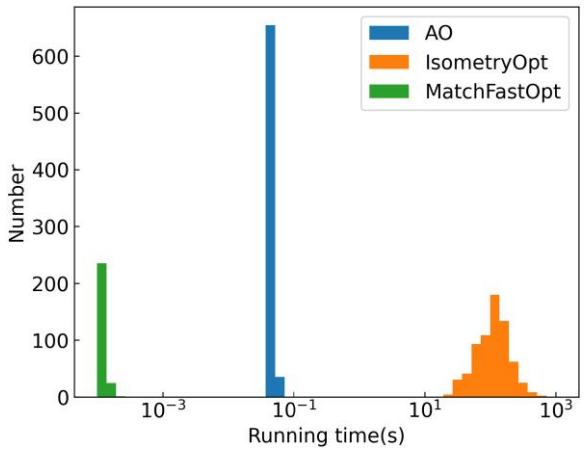


Fig. 14. Running time for calculations of GS-index between CH_3 cation and 689 isomers of $C_2H_4O_2$.

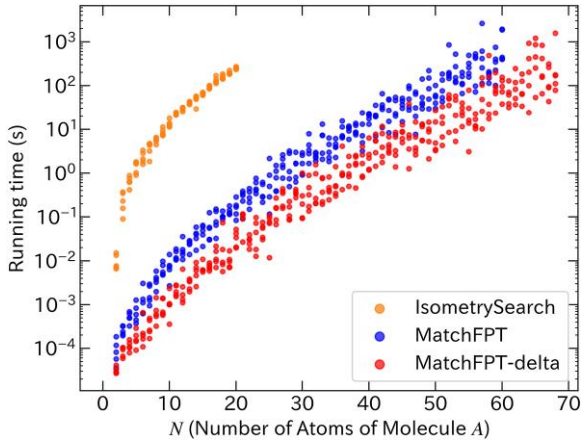


Fig. 15. Running time to obtain the solution if molecule B with $2N$ atoms contains molecule A with N atoms.

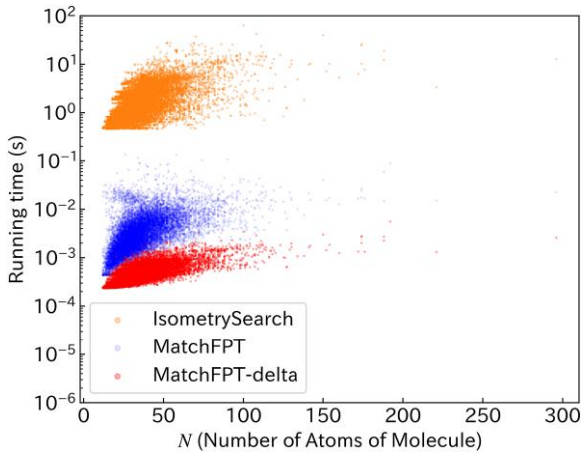


Fig. 16. Running time to determine whether a benzene ring (query) is contained in each of the molecules in the dataset from NCI Open Database Compounds.[32]

considerably faster than IsometryOpt. Although IsometryOpt should be theoretically faster for larger molecules, it could not be confirmed due to a limitation of computer resource.

Table 4. Overview of proposed and conventional methods. (a) Four methods for comparison between molecules of equal size and substructure match between molecules of different size. (b) Three methods for substructure enumeration. Proposed methods are written in bold font.

(a)

Method	Characteristic	Optimality	Time	
			Same size	Different size
AO	Alternate optimization	No	Excellent	Excellent
TSR	Tangent space relaxation	No	Excellent	Not applicable
IsometryOpt	BnB for isometry	Yes	Good	Bad
MatchFastOpt	BnB for assignment	Yes	Fair	Fair

(b)

Method	Characteristic	Time
IsometrySearch	BnB for isometry	Fair
MatchFPT	BnB for assignment	Good
MatchFPT-delta	Improvement of the MatchFPT method	Excellent

Two new methods, IsometrySearch and MatchFPT-delta, for the improvement of substructure enumeration also perform better than the conventional methods. IsometrySearch employs BnB for isometric transformations, and MatchFPT-delta is a variant of MatchFPT (Sasaki et al.), which is also a method for substructure enumeration. Computer experiments have shown that MatchFPT-delta reduces computational time compared to the original MatchFPT method. As far as the assessment we have conducted, MatchFPT-delta is much faster than IsometrySearch. However, IsometrySearch is expected to become faster with larger query molecules.

These new methods increase the rigorousness of G-RMSD. The user can now choose efficiency (AO, TSR) or accuracy (IsometryOpt or MatchFastOpt) for measuring the similarity between molecules of equal or different size and can use MatchFPT-delta for substructure enumeration, depending on the problem settings. These G-RMSD algorithms are useful for various purposes of measuring molecular similarity, including search of 3D chemical structure databases, analysis of a large number of trajectories from MD simulations, evaluation of molecular similarity measures, and exploration of the abundant chemical space.

Acknowledgments

The authors would like to thank Yoichi Sasaki for providing his code of MatchFPT for computational experiments.

Supplementary data

Supplementary material is available at *Bulletin of the Chemical Society of Japan* online.

Funding

This research is supported by JSPS KAKENHI Grant Numbers JP21K11776.

Conflict of interest statement. None declared.

Data availability

The exact G-RMSD program code and the dataset used for the numerical experiment are available. The availability and requirements are as follows:

Program name: X-GRMSD
Program home page: https://github.com/striwata/X-GRMSD
Program language: C++ language
License: MIT license

References

1. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**, p. 395–403.
2. T. Engel, *Chemoinformatics, a Textbook*, eds. by J. Gasteiger, T. Engel, Wiley-VCH, Weinheim, **2003**, pp. 15–168.
3. N. Kochev, V. Monev, I. Bangov, *Chemoinformatics, a textbook*, eds. by J. Gasteiger, T. Engel, Wiley-VCH, Weinheim, **2003**, pp. 291–318.
4. P. Willett. Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance, *WIREs Data Min. Knowl. Discovery*. **2011**, *1*, 241. https://doi.org/10.1002/widm.26.
5. H. Satoh, H. Koshino, K. Funatsu, T. Nakata. Novel canonical coding method for representation of three-dimensional structures, *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 622. https://doi.org/10.1021/ci990147d.
6. H. Satoh, H. Koshino, K. Funatsu, T. Nakata. Representation of molecular configurations by CAST coding method, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1106. https://doi.org/10.1021/ci000136g.
7. H. Satoh, H. Koshino, T. Nakata. Extended CAST coding method for exact search of stereochemical structures, *J. Comput. Aided Chem*. **2002**, *3*, 48. https://doi.org/10.2751/jcac.3.48.
8. H. Satoh. Numerical representation of three-dimensional stereochemical environments using FRAU-descriptors, *Croat. Chem. Acta*. **2007**, *80*, 217. https://doi.org/10.3929/ethz-b-000007143.
9. S. De, A. P. Bartók, G. Csányi, M. Ceriotti. Comparing molecules and solids across structural and alchemical space, *Phys. Chem. Chem. Phys*. **2016**, *18*, 13754. https://doi.org/10.1039/C6CP00415F.
10. S. De, F. Musil, C. Baldauf, M. Ceriotti. Mapping and classifying molecules from a high-throughput structural database, *J. Cheminf*. **2017**, *9*, 6. https://doi.org/10.1186/s13321-017-0192-4.
11. A. D. McLachlan. A mathematical procedure for superimposing atomic coordinates of proteins, *Acta Cryst*. **1972**, *28*, 656. https://doi.org/10.1107/S0567739472001627.
12. J. M. Vásquez-Pérez, G. U. G. Martínez, A. M. Köster, P. Calaminici. The discovery of unexpected isomers in sodium heptamers by born-oppenheimer molecular dynamics, *J. Chem. Phys*. **2009**, *131*, 124126. https://doi.org/10.1063/1.3231134.
13. K. Hori. A data base for transition states. Ranking of synthesis routes by using a system combined computational with information chemistry, *J. Comput. Aided Chem*. **2001**, *2*, 37. https://doi.org/10.2751/jcac.2.37.
14. G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space, *New J. Phys*. **2013**, *15*, 095003. https://doi.org/10.1088/1367-2630/15/9/095003.
15. R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules, *Sci. Data*. **2014**, *1*, 140022. https://doi.org/10.1038/sdata.2014.22.
16. R. Ramakrishnan, M. Hartmann, E. Tapavicza, O. A. von Lilienfeld. Electronic spectra from TDDFT and machine learning in chemical space, *J. Chem. Phys*. **2015**, *143*, 084111. https://doi.org/10.1063/1.4928757.
17. M. Nakata. The PubChemQC project: a large chemical database from the first principle calculations, *AIP Conf. Proc*. **2015**, *1702*, 090058. https://doi.org/10.1063/1.4938866.
18. M. Nakata, T. Shimazaki. Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry, *J. Chem. Inf. Model*. **2017**, *57*, 1300. https://doi.org/10.1021/acs.jcim.7b00083.
19. M. Nakata, The PubChemQC project [accessed 2023 October 1]. https://nakatamaho.riken.jp/pubchemqc.riken.jp/.
20. H. Satoh, T. Oda, K. Nakakoji, T. Uno, S. Iwata, K. Ohno. "Maizo"-chemistry project: toward molecular- and reaction discovery from quantum mechanical global reaction route mappings, *J. Comput. Chem. Jpn*. **2015**, *14*, 77. https://doi.org/10.2477/jccj.2015-0048.
21. H. Satoh, T. Oda, K. Nakakoji, T. Uno, S. Iwata, K. Ohno, Reaction map database server [accessed 2023 October 1]. https://github.com/ReactionMap/RMapServer.
22. H. Satoh, T. Oda, K. Nakakoji, T. Uno, S. Iwata, K. Ohno. Rmapdb: chemical reaction route map data for quantum mechanical-based data chemistry, *Materials Cloud Archive*. **2020**, *2020*, 138. https://doi.org/10.24435/materialscloud:5f-14.
23. F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, K.-R. Müller. Bypassing the Kohn-Sham equations with machine learning, *Nat. Commun*. **2017**, *8*, 872. https://doi.org/10.1038/s41467-017-00839-3.
24. T. Fukutani, K. Miyazawa, S. Iwata, H. Satoh. G-RMSD: root mean square deviation based method for three-dimensional molecular similarity determination, *Bull. Chem. Soc. Jpn*. **2020**, *94*, 655. https://doi.org/10.1246/bcsj.20200258.
25. T. Tsutsumi, Y. Ono, T. Taketsugu. Visualization of reaction route map and dynamical trajectory in reduced dimension, *Chem. Commun*. **2021**, *57*, 11734. https://doi.org/10.1039/D1CC04667E.
26. P. J. Besl, N. D. McKay. Method for registration of 3-D shapes, *IEEE Trans. Pattern Anal. Mach. Intell*. **1992**, *14*, 239. https://doi.org/10.1109/34.121791.
27. J. Yang, H. Li, D. Campbell, Y. Jia. Go-ICP: a globally optimal solution to 3D ICP point-set registration, *IEEE Trans. Pattern Anal. Mach. Intell*. **2016**, *38*, 2241. https://doi.org/10.1109/TPAMI.2015.2513405.
28. Y. Sasaki, T. Shibuya, K. Ito, H. Arimura. Efficient approximate 3-dimensional, point set matching using root-mean-square deviation score, *IEICE Trans. Fundamentals*. **2019**, *E102-A*, 1159. https://doi.org/10.1587/transfun.E102.A.1159.
29. B. H. Korte, J. Vygen, *Combinatorial Optimization*, Springer, Berlin, **2011**.
30. K. S. Arun, T. S. Huang, S. D. Blostein. Least-squares fitting of two 3-D point sets, *IEEE Trans. Pattern Anal. Mach. Intell*. **1987**, *9*, 698. https://doi.org/10.1109/TPAMI.1987.4767965.
31. C. R. Chegireddy, H. W. Hamacher. Algorithms for finding K-best perfect matchings, *Discrete Appl. Math*. **1987**, *18*, 155. https://doi.org/10.1016/0166-218X(87)90017-5.
32. NCI Open Database Compounds, National Cancer Institute. [accessed 2023 October 1]. https://cactus.nci.nih.gov/download/nci/index.html.
33. G. Guennebaud, B. Jacob. Eigen v3. [accessed 2021 October 30]. https://eigen.tuxfamily.org.
34. M. Buehren. mcximing/hungarian-algorithm-cpp: a C++ wrapper for a hungarian algorithm implementation", GitHub. [accessed 2022 February 2]. https://github.com/mcximing/hungarian-algorithm-cpp.