Statistics of spatial rotations and projection directions considering molecular symmetry in 3D electron cryo-microscopy

Qi Zhang¹, ², ³, ⁴, Hai Lin⁵, ⁶, and Mingxu Hu¹, ², ³, ⁴

¹Key Laboratory of Protein Sciences (Tsinghua University), Ministry of Education, Beijing, China
²School of Life Science, Tsinghua University, Beijing, China
³Beijing Advanced Innovation Center for Structural Biology
⁴Beijing Frontier Research Center for Biological Structure
⁵Department of Mathematical Sciences, Tsinghua University, Beijing, China
⁶Yau Mathematical Sciences Center, Tsinghua University, Beijing, China

†Correspondence should be addressed to M.H. (humingxu@mail.tsinghua.edu.cn) and H.L. (hailin@mail.tsinghua.edu.cn).
Abstract

Electron cryo-microscopy (cryoEM) three-dimensional (3D) reconstruction is based on estimations of orientations of projection images or 3D volumes. It is common that the macromolecules studied by cryoEM have molecular symmetry, which, unfortunately, has not been taken into consideration by any statistics for either spatial rotations or projection directions at this point. Meanwhile, there are growing needs to adopt advanced statistical methods, and further, modern machine learning techniques in cryoEM. Since those methodologies are built heavily upon statistical learning cornerstones, the absence of their domain-specific statistical justification limits their applications in cryoEM. In this research, based on the concept of non-unique-games (NUG), we propose two key statistical measurements, the mean and the variance, of both spatial rotations and projection directions when molecular symmetry is considered. Such methods are implemented in the open-source python package pySymStat.
1 Introduction

In the three-dimensional (3D) reconstruction of electron cryo-microscopy (cryoEM), particle images are projections of biological samples at random orientations. Therefore, the analysis of such orientations is the key step in the reconstruction of 3D density maps[1, 2, 3, 4, 5, 6, 7, 8]. Accordingly, the analysis of both spatial rotations and projection directions is fundamentally important for cryoEM. In the cryoEM field, it is common that the macromolecules studied have molecular symmetry[9, 10]. However, the true topology induced by the molecular symmetry is currently neglected as there are no statistics considering molecular symmetry.

Meanwhile, there are growing demands to adopt advanced statistical methods, and further, machine learning techniques[11, 12] in cryoEM. Modern machine learning practices, including deep learning, are built upon statistical learning cornerstones. Therefore, such absence of statistics tools limits the applications of advanced statistical methods and machine learning techniques in cryoEM, in the circumstances that the macromolecules studied have molecular symmetry. Those methodologies would likely fail if true topology induced by the molecular symmetry is neglected, especially when the order of molecular symmetry is high.

In this study, we propose statistics of both spatial rotations and projection directions which take molecular symmetry into consideration. Such methods are realized in the open-source python package pySymStat. Our solution starts with representing spatial rotations using unit quaternions. We then estimate the mean and variance in a non-unique-games (NUG) framework, which is devised for solving a certain type of optimization problems defined over a locally compact group. The solution to our optimization problem is the convex combination[13] of all possible estimates, therefore we utilize a leaderboard algorithm to recover the final true result, i.e., the global optimal solution, from the NUG estimates.

The unit quaternion[14, 15, 16] is widely used for describing rotations. Section 2 gives a brief overview of this concept as well as clarifies the notations used in this work. In Section 3, we define distance, variance, and mean in the context of considering molecular symmetry. In Section 4, we demonstrate how such variance and mean problems can be solved by a method developed based on NUG[17, 18]. In Section 5, the performance of the proposed methods is analyzed. After determining the distance, variance, and mean, clustering is introduced in Section 6. The potential applications of the proposed methods are discussed in Section 7.

2 Unit quaternion description

The unit quaternion description of rotation has been well studied and applied in many fields, including robotics[19, 20, 21, 22], aviation[23, 24, 25, 26], computational vision[27, 28, 29], as well as cryoEM[30, 31, 32, 33, 34, 7, 35, 5, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45]. In the present work, we give a very brief introduction of the unit quaternion description of spatial rotations. For general introductions of quaternion, we refer the reader to [15, 16, 46, 47].

Notations in this article are specified according to the following rules. The unit quaternion in vector form, which is a unit four-dimensional (4D) vector, is denoted by a bold letter, e.g., \( \mathbf{q} \). The molecular
symmetry group is denoted by $G$. An element in the molecular symmetry group $G$ is denoted by $g$. Such an element is also in the unit quaternion description. Meanwhile, the number of elements in $G$ is denoted by $|G|$. The set containing all 3D spatial rotations is denoted by $SO(3)$.

A quaternion is an extension of a complex number, composed of a real part and three imaginary parts. The quaternion is commonly viewed as a 4D vector. If its norm equals 1, i.e., $|q| = 1$, the quaternion is called a unit quaternion. Quaternion multiplication follows special rules, whose definition can be found in [15, 16]. The quaternion multiplication operator is denoted by $\otimes$ in the present work.

The rotation of a 3D vector $\vec{v}$ to another vector $\vec{v}'$ can be achieved by quaternion multiplication. By viewing $\vec{v}$ as $\vec{v} = (0 \vec{v})$ in quaternion form, the rotation induced by a unit quaternion is defined by[15]

$$\vec{v}' = \begin{pmatrix} 0 \\ \vec{v}' \end{pmatrix} = \mathbf{q} \otimes \begin{pmatrix} 0 \\ \vec{v} \end{pmatrix} \otimes \mathbf{q}^*, \quad (1)$$

where $\mathbf{q}^*$ is the conjugate[47, 15, 16] of $\mathbf{q}$. Therefore, $\mathbf{q}$ is a description of such spatial rotation. Such rotation is also denoted by

$$\vec{v}' = R_{\mathbf{q}}(\vec{v}). \quad (2)$$

For sequential rotations, such as a spatial rotation $\mathbf{q}_1$ followed by a spatial rotation $\mathbf{q}_2$, the overall rotation is calculated as $\mathbf{q}_2 \otimes \mathbf{q}_1$[15]. When molecular symmetry is not considered, the distance between the two spatial rotations $\mathbf{q}_1$ and $\mathbf{q}_2$ can be defined by[47].

$$d_{SO(3)}(\mathbf{q}_1, \mathbf{q}_2) := \max_{\vec{v} \in S^2} d_{S^2}(R_{\mathbf{q}_1}(\vec{v}), R_{\mathbf{q}_2}(\vec{v})), \quad (3)$$

where $d_{S^2}$ is the natural distance in the unit sphere $S^2$.

Meanwhile, there are several different definitions of the projection direction in the cryoEM field. Some of them, however, are ambiguous. Hence, we explicitly define the projection direction as follows, in the present work. We denote the operator of projection for a spatial rotation $\mathbf{q}$ as $\mathbf{P}^\mathbf{q}$. In other words, $\mathbf{P}^\mathbf{q}(V)$ stands for the image obtained as follows. The map $V$ is rotated by spatial rotation $\mathbf{q}$; then, the rotated map is projected along the $Z$ axis. The defining of projection direction $\vec{v}_p$ can be achieved by defining the set of projected images along $\vec{v}_p$, which is denoted by $\mathbf{P}_{\vec{v}_p}$. This set is defined by

$$\mathbf{P}_{\vec{v}_p} := \{\mathbf{P}^\mathbf{q}(V)\}_{\mathbf{q} : R_{\mathbf{q}}(\vec{v}_p) = \vec{v}_n}, \quad (4)$$

where $\vec{v}_n = (0, 0, 1)^T$ is the unit vector along the $Z$ axis. In other words, every projection image in this set is projected along the $Z$ axis from a rotated volume. Such a rotation transforms $\vec{v}_p$ to $\vec{v}_n$. For two projection directions, $\vec{v}_{p_1}$ and $\vec{v}_{p_2}$, they are equal if and only if their corresponding sets of projected images are equal, i.e., $\mathbf{P}_{\vec{v}_{p_1}} = \mathbf{P}_{\vec{v}_{p_2}}$. The definition of the distance between the two projection directions $\vec{v}_{p_1}$ and $\vec{v}_{p_2}$ is assigned by the natural distance in the unit sphere $S^2$, when molecular symmetry is not taken into consideration.
3 Definitions of distance, variance, and mean considering molecular symmetry

In this section, we redefine two most basic statistical measurements, the mean and the variance, in the context of considering molecular symmetry. We first define the equivalence relations on the sets of spatial rotations and projection directions, respectively (Section 3.1). Their inter-set distances are then used to measure the distances between two spatial rotations or two spatial rotations, respectively (Section 3.2), based on which we define the mean and the variance accordingly (Section 3.3).

3.1 Equivalence relations considering molecular symmetry

The symmetry property of a molecule can be represented by a molecular symmetry group. For a molecule with symmetry $G$, the spatial rotation $q$ is equivalent to $q \otimes g_m$ for any $g_m$ in a molecular symmetry group $G$. This is proven in Appendix I. We denote all of these equivalent spatial rotations of $q$ as $q \otimes G$, defined by

$$q \otimes G := \{q \otimes g_m\}_{1 \leq m \leq |G|}.$$  (5)

In mathematical terms, $q \otimes G$ is a left coset of $G$ in $SO(3)$. An element in the set of equivalent spatial rotations $q \otimes G$, i.e., $q \otimes g_m$, is referred to as a representative of this set.

Meanwhile, when the molecular symmetry $G$ is taken into consideration, the projection direction $\vec{v}_p$ is equivalent to $R_{g_m}^{-1}(\vec{v}_p)$ for any $g_m$ in $G$; this is proven in Appendix II. Here, an inverse operation of $g_m$ exists for the completeness of the concept. $R_{g_m}^{-1}(\vec{v}_p)$ is the equivalent projection direction generated by rotating the map $V$ by the symmetry element $g_m$ (see Appendix II).

We denote all of these equivalent projection directions of $\vec{v}_p$ as $R_{G^{-1}}(\vec{v}_p)$, defined by

$$R_{G^{-1}}(\vec{v}_p) := \{R_{g_m}^{-1}(\vec{v}_p)\}_{1 \leq m \leq |G|}.$$  (6)

An element in the set of equivalent projection direction $R_{G^{-1}}(\vec{v}_p)$, i.e., $R_{g_m}^{-1}(\vec{v}_p)$, is referred to as a representative of this set.

For a set of spatial rotations $\{q_i\}_{1 \leq i \leq N}$ and a set of projection directions $\{\vec{v}_p\}_{1 \leq i \leq N}$, their statistical analyses are actually performed on $\{q_i \otimes G\}_{1 \leq i \leq N}$ and $\{R_{G^{-1}}(\vec{v}_p)\}_{1 \leq i \leq N}$ respectively, when the molecular symmetry $G$ is taken into consideration.

3.2 Distance considering molecular symmetry

Definitions of the distance between two sets of equivalent spatial rotations and the distance between two sets of equivalent projection directions are the prerequisites of considering the statistics of spatial rotations and projection directions, respectively, when molecular symmetry is taken into consideration.

It is a common approach to define the distance between two sets by the distance between their representatives[50]. For spatial rotations, we define the distance between $q_1 \otimes G$ and $q_2 \otimes G$ by

$$d_{SO(3)/G}(q_1 \otimes G, q_2 \otimes G) := \min_{g_{t_1}, g_{t_2} \in G} d_{SO(3)}(q_1 \otimes g_{t_1}, q_2 \otimes g_{t_2}).$$  (7)
which is the shortest distance in $SO(3)$ by selecting representatives $q_1 \otimes g_{t_1}$ and $q_2 \otimes g_{t_2}$ from $q_1 \otimes G$ and $q_2 \otimes G$, respectively.

Similar to the distance definition of two spatial rotations, we define the distance between two sets of equivalent projection directions $R_G (\vec{v}_{p_1})$ and $R_G (\vec{v}_{p_2})$ by

$$d_{S^2/G} (R_G (\vec{v}_{p_1}), R_G (\vec{v}_{p_2})) := \min_{g_{t_1}, g_{t_2} \in G} d_{S^2} \left( R_{g_{t_1}}^{-1} (\vec{v}_{p_1}), R_{g_{t_2}}^{-1} (\vec{v}_{p_2}) \right),$$

which is the shortest distance in $S^2$ by selecting representatives $R_{g_{t_1}}^{-1} (\vec{v}_{p_1})$ and $R_{g_{t_2}}^{-1} (\vec{v}_{p_2})$ from $R_{G^{-1}} (\vec{v}_{p_1})$ and $R_{G^{-1}} (\vec{v}_{p_2})$, respectively. $d_{S^2}$ is the natural distance in the unit sphere $S^2$.

These definitions, i.e., Equation (7) and Equation (8), both satisfy all of the requirements of the distance definition in mathematics (see Appendix III and Appendix IV, respectively), namely, the symmetry property, the positive definiteness property, and the triangle inequality property.

### 3.3 Mean and variance considering molecular symmetry

In the same spirit of the distance definition, we define the variance of spatial projections $\{q_i \otimes G\}_{1 \leq i \leq N}$ and projection directions $\{R_G (\vec{v}_{p_i})\}_{1 \leq i \leq N}$, respectively as

$$\text{Var} \left( \{q_i \otimes G\}_{1 \leq i \leq N} \right) := \min_{g_{t_1}, g_{t_2}, \ldots, g_{t_N} \in G} \left( \sum_{1 \leq i,j \leq N} d^2_{SO(3)} (q_i \otimes g_{t_i}, q_j \otimes g_{t_j}) \right)$$

(9)

and

$$\text{Var} \left( \{R_G (\vec{v}_{p_i})\}_{1 \leq i \leq N} \right) := \min_{g_{t_1}, g_{t_2}, \ldots, g_{t_N} \in G} \left( \sum_{1 \leq i,j \leq N} d^2_{SO(3)} \left( R_{g_{t_i}}^{-1} (\vec{v}_{p_i}), R_{g_{t_j}}^{-1} (\vec{v}_{p_j}) \right) \right).$$

(10)

In other words, for spatial rotations, a representative $q_i \otimes g_{t_i}$ is chosen from each set of equivalent spatial rotations $q_i \otimes G$, such that all representatives are as close as possible overall. The projection direction follows the same principle. Furthermore, we define the mean of spatial rotations $\{q_i \otimes G\}_{1 \leq i \leq N}$ and projection directions $\{R_{G^{-1}} (\vec{v}_{p_i})\}_{1 \leq i \leq N}$, respectively as

$$\text{Mean} \left( \{q_i \otimes G\}_{1 \leq i \leq N} \right) := \left( \arg\min_{q_i} \min_{g_{t_1}, g_{t_2}, \ldots, g_{t_N} \in G} \sum_{i=1}^{N} d^2_{SO(3)} (q_i, q_i \otimes g_{t_i}) \right) \otimes G$$

(11)

and

$$\text{Mean} \left( \{R_{G^{-1}} (\vec{v}_{p_i})\}_{1 \leq i \leq N} \right) := R_{G^{-1}} \left( \arg\min_{\vec{v}_{p_i}} \min_{g_{t_1}, g_{t_2}, \ldots, g_{t_N} \in G} \sum_{i=1}^{N} d^2_{SO(3)} (\vec{v}_{p_i}, R_{g_{t_i}}^{-1} \vec{v}_{p_i}) \right).$$

(12)

In other words, for spatial rotations, the mean of $\{q_i \otimes G\}_{1 \leq i \leq N}$ is defined by the spatial rotation closest to $\{q_i \otimes G\}_{1 \leq i \leq N}$. The mean of the projection direction follows the same principle.

The problems of determining the variance and the mean are in some sense equivalent. For spatial rotations, the mean definition by Equation (11) can be approximated by (see Appendix V)

$$\text{Mean}(\{q_i \otimes G\}_{1 \leq i \leq N}) \approx \left( \arg\min_{q_i} \sum_{i=1}^{N} d^2_{SO(3)} (q_i, q_i \otimes \tilde{g}_{t_i}) \right) \otimes G,$$

(13)
where
\[
\{\tilde{g}_i\}_{1 \leq i \leq N} = \arg\min_{g_1, g_2, \ldots, g_N} \left( \sum_{1 \leq i, j \leq N} d^{2}_{SO(3)}(q_i \otimes \tilde{g}_i, q_j \otimes \tilde{g}_j) \right)
\]  
(14)

and \(\arg\min_{q} \sum_{i=1}^{N} d^{2}_{SO(3)}(q, q_i \otimes \tilde{g}_i)\) is solvable by an established method[30] once \(\{\tilde{g}_i\}_{1 \leq i \leq N}\) are determined. Meanwhile, for projection directions, the mean definition by Equation (12) can be approximated by (see Appendix VI)

\[
\text{Mean}(\{R_{G^{-1}}(\tilde{v}_p)\}_{1 \leq i \leq N}) \approx R_{G^{-1}}\left(\arg\min_{\tilde{v}_p} \sum_{i=1}^{N} d^{2}_{S2} \left( \tilde{v}_p, R_{g_i^{-1}}(\tilde{v}_p) \right) \right),
\]  
(15)

where
\[
\{\tilde{g}_i\}_{1 \leq i \leq N} = \arg\min_{g_1, g_2, \ldots, g_N} \left( \sum_{1 \leq i, j \leq N} d^{2}_{S2} \left( R_{g_i^{-1}}(\tilde{v}_p), R_{g_j^{-1}}(\tilde{v}_p) \right) \right)
\]  
(16)

and \(\arg\min_{\tilde{v}_p} \sum_{i=1}^{N} d^{2}_{S2} \left( \tilde{v}_p, R_{g_i^{-1}}(\tilde{v}_p) \right)\) is also solvable by an established method[51] once \(\{\tilde{g}_i\}_{1 \leq i \leq N}\) are determined.

In other words, the representatives chosen in the determination of the variances, i.e., Equation (9) and (10), determine \(\{\tilde{g}_i\}_{1 \leq i \leq N}\). The means of spatial rotations and projection directions can be approximated by Equation (13) and Equation (15), by using \(\{\tilde{g}_i\}_{1 \leq i \leq N}\).

In conclusion, choosing the overall-closest representatives, i.e., \(q_1 \otimes \tilde{g}_1, q_2 \otimes \tilde{g}_2, \ldots, q_N \otimes \tilde{g}_N\) for spatial rotations, and \(R_{g_1^{-1}}(\tilde{v}_p_1), R_{g_2^{-1}}(\tilde{v}_p_2), \ldots, R_{g_N^{-1}}(\tilde{v}_p_N)\) for projection directions, where \(\{\tilde{g}_i\}_{1 \leq i \leq N}\) are selected by Equation (14) and Equation (16), is the key step in computing the statistics of spatial rotations and projection directions, respectively, when molecular symmetry is taken into consideration. In other words, once Equation (14) and Equation (16) are solved, the variance and the mean of \(\{q_i \otimes G\}_{1 \leq i \leq N}\) and \(\{R_{G^{-1}}(\tilde{v}_p)\}_{1 \leq i \leq N}\) can be calculated, respectively.

## 4 NUG and the determination of representatives

In this section, we present an approximation algorithm for determining \(\{\tilde{g}_i\}_{1 \leq i \leq N}\) in both Equation (14) and Equation (16), which is equivalent to the variance and the mean determination, as discussed above. This approximation algorithm is based on the framework of non-unique games (NUG), developed by Lederman and Singer[17, 18]. Therefore, we refer to this algorithm as the NUG approach. The \(\{\tilde{g}_i\}_{1 \leq i \leq N}\) determination in both Equation (14) and Equation (16) is first translated into a NUG problem (see Section 4.1). Then, by making use of representation theory[52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 17, 18, 63] (see Section 4.2 for a brief overview), the NUG problem is relaxed into a semi-positive definite programming (SDP) problem[64, 65] (see Section 4.3 and Section 4.4), which can be solved by a convex optimization technique[64, 13]. However, the solution of the SDP problem is a “relaxation” of the NUG problem. In other words, the constraint in the SDP problem is looser than that of the NUG problem. Therefore, in the present work, we propose a leaderboard algorithm for “re-constraining” the SDP solution into a proper solution for the NUG problem (see Section 4.5). Currently, such algorithms are mainly adopted as leaderboard peptide sequencing in bioinformatics[66, 67].
4.1 Determination of representatives is a NUG problem

The original NUG problem is defined by the following optimization problem

$$
\arg\min_{x_1, x_2, \ldots, x_N \in \mathcal{X}} \sum_{1 \leq i, j \leq N} f_{ij}(x_i x_j^{-1}),
$$

(17)

where $\mathcal{X}$ is a given locally compact group, and $\{f_{ij}\}_{1 \leq i, j \leq N}$ are $N^2$ given real valued functions on $\mathcal{X}$.

The $\{\tilde{g}_t\}_{1 \leq i \leq N}$ determination for both spatial rotation statistics, i.e., Equation (14), and projection direction statistics, i.e., Equation (16), can be derived in the same form (see Appendix VII and Appendix VIII)

$$
\{\tilde{g}_t\}_{1 \leq i \leq N} = \arg\min_{g_1, g_2, \ldots, g_N \in G} \left( \sum_{1 \leq i, j \leq N} f_{ij}(g_i \otimes g_j^{-1}) \right).
$$

(18)

In the above equation, $f_{ij}(g_i \otimes g_j^{-1})$ is assigned as (see Appendix VII),

$$
f_{ij}(g_i \otimes g_j^{-1}) = d_{SO(3)}^2(g_i \otimes g_j^{-1}, q_i^{-1} \otimes q_j)
$$

(19)

for the representative determination of spatial rotations. Meanwhile, for the representative determination of projection directions, $f_{ij}(g_i \otimes g_j^{-1})$ is assigned as (see Appendix VIII),

$$
f_{ij}(g_i \otimes g_j^{-1}) = d_{S^2}^2 (\vec{v}_p, R_{g_i \otimes g_j^{-1}}(\vec{v}_p)).
$$

(20)

Moreover, the molecular symmetry group $G$ is a finite group. Hence, the determination of representatives meets all of the requirements of being a NUG problem. In the following, we solve this NUG problem for determining the variance and the mean for both spatial rotations and projection directions.

4.2 Recall of representation theory

As in the NUG problem, the variables in the optimization are elements of the molecular symmetry group $G$. However, it is often difficult to optimize over a finite group. Proposed by Singer and Lederman[17, 18], such problems can be converted into optimizations over matrices using representation theory[52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63]. Representation theory provides a general Fourier transform between a group and matrices, making such conversions possible.

We would like to briefly present some useful concepts and methods of representation theory in this section. As the molecular symmetry group $G$ is considered in the present work and it is a finite group, this overview is limited to representation theory of the finite group[59, 60, 61, 62, 63].

For a finite group $G$, it corresponds to several intrinsically different irreducible representations. The number of irreducible representations of $G$ is denoted by $K$. The irreducible representations of $G$ are denoted by $\{\rho_k\}_{1 \leq k \leq K}$. The irreducible representation $\rho_k$ is a function, which maps every element in a group $G$ to a real or complex $d_k \times d_k$ unitary matrix, where $d_k$ is the dimensionality of the $k$-th irreducible representation. When $G$ is given, its irreducible representations, i.e., $\{\rho_k(g_m)\}_{1 \leq k \leq K}$ are known. The irreducible representations have the property described below.
By the Peter-Weyl theorem[68], a function \( f : G \to \mathbb{R} \) has a general Fourier transformation, as follows. The \( k \)-th Fourier coefficient is defined by

\[
\hat{f}^{(k)} := \frac{1}{|G|} \sum_{m=1}^{|G|} f(g_m) \rho_k(g_m)^* ,
\]

(21)

where \( \rho_k(g_m)^* \) is the conjugate transpose of \( \rho_k(g_m) \). Then, the general Fourier expansion of \( f \) is

\[
f(g_m) = \sum_{k=0}^{K-1} d_k \text{tr} \left( \hat{f}^{(k)} \rho_k(g_m) \right) ,
\]

(22)

where \( \text{tr} \left( \hat{f}^{(k)} \rho_k(g_m) \right) \) is the trace of the matrix \( \hat{f}^{(k)} \rho_k(g_m) \).

### 4.3 Matrix form of a NUG by representation theory

By a general Fourier transformation using representation theory, introduced as Equation (21) and Equation (22), the NUG problem, i.e., Equation (14), is converted into the matrix form as (see Appendix IX)

\[
\{ \mathbf{g}_t \}_{1 \leq t \leq N} = \arg\min_{\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_N \in G} \sum_{k=0}^{K-1} \text{tr} \left( \hat{F}^{(k)} \mathbf{X}^{(k)} \right)
\]

(23)

where

\[
\hat{F}^{(k)} = d_k \begin{pmatrix}
\hat{f}_{11}^{(k)} & \hat{f}_{12}^{(k)} & \cdots & \hat{f}_{1N}^{(k)} \\
\hat{f}_{21}^{(k)} & \hat{f}_{22}^{(k)} & \cdots & \hat{f}_{2N}^{(k)} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{f}_{1N}^{(k)} & \hat{f}_{2N}^{(k)} & \cdots & \hat{f}_{NN}^{(k)}
\end{pmatrix},
\]

(24)

in which

\[
\hat{f}_{ij}^{(k)} = \sum_{m=1}^{|G|} f_{ij}(g_m) \rho_k^*(g_m),
\]

(25)

and

\[
\mathbf{X}^{(k)} = \begin{pmatrix}
\rho_k(\mathbf{g}_1 \otimes \mathbf{g}_1^{-1}) & \rho_k(\mathbf{g}_1 \otimes \mathbf{g}_2^{-1}) & \cdots & \rho_k(\mathbf{g}_1 \otimes \mathbf{g}_N^{-1}) \\
\rho_k(\mathbf{g}_2 \otimes \mathbf{g}_1^{-1}) & \rho_k(\mathbf{g}_2 \otimes \mathbf{g}_2^{-1}) & \cdots & \rho_k(\mathbf{g}_2 \otimes \mathbf{g}_N^{-1}) \\
\vdots & \vdots & \ddots & \vdots \\
\rho_k(\mathbf{g}_N \otimes \mathbf{g}_1^{-1}) & \rho_k(\mathbf{g}_N \otimes \mathbf{g}_2^{-1}) & \cdots & \rho_k(\mathbf{g}_N \otimes \mathbf{g}_N^{-1})
\end{pmatrix}.
\]

(26)

### 4.4 Convex relaxation of NUG to SDP

In the matrix form of a NUG problem, i.e., Equation (23), \( \hat{F}^{(k)} \) is known. As shown by Equation (25), the element of \( \hat{F}^{(k)} \), i.e., \( \hat{f}_{ij}^{(k)} \), is determined by the given symmetry group \( G \) and the values of function \( f_{ij} \). In both cases (spatial rotations and projection directions), the values of the function \( f_{ij} \) are pre-determined by \( \{ q_i \}_{1 \leq i \leq N} \) and \( \{ v_{pi} \}_{1 \leq i \leq N} \) by Equation (19) and Equation (20), respectively. Meanwhile, \( \mathbf{X}^{(k)} \) is composed of the variables, i.e., \( \{ \mathbf{g}_i \}_{1 \leq i \leq N} \), in the optimization problem. Thus, Equation (23) is an optimization problem over the variables \( \{ \mathbf{X}^{(k)} \}_{0 \leq k \leq K-1} \), subject to Equation (26) as constraints.

However, these constraints make the defined domain in this optimization problem in general non-convex. Thus, convex optimization techniques are not suitable, leading to the difficulty of determining a
global optimal solution. However, such matrices $\{X^{(k)}\}_{1 \leq k \leq K}$ can be relaxed into their convex hull[18, 69, 70] (see Appendix X), converting this NUG problem to an SDP problem. It is a classic problem to solve the solution of an SDP problem. In convex optimization, the global optimal solution is guaranteed. One can obtain the solution using open-source packages[71, 72, 73, 74]. Thus, instead of solving $\{\tilde{g}_i\}_{1 \leq i \leq N}$, the solution matrices $\{X^{(k)}\}_{0 \leq k \leq K-1}$ are obtained. Such solution matrices $\{\tilde{X}^{(k)}\}_{0 \leq k \leq K-1}$ are in the convex hull of the matrices defined in Equation (26).

4.5 Recovery of representatives from the solution of SDP

As the constraint in the SDP problem is looser than that in the NUG problem, the solution matrices of the SDP problem may no longer comply with the constraints of the NUG problem, i.e., Equation (26). In other words, $\{\tilde{X}^{(k)}\}_{0 \leq k \leq K-1}$, i.e., the solution matrices of the SDP problem, may violate Equation (26). However, determination of representatives, i.e., determination of $\{\tilde{g}_i\}_{1 \leq i \leq N}$, is needed for calculating mean and variance of rotations considering molecular symmetry. Thus, a recovery algorithm is needed for “re-constraining” $\{\tilde{X}^{(k)}\}_{0 \leq k \leq K-1}$ to obtain $\{\tilde{g}_i\}_{1 \leq i \leq N}$.

We propose this recovery algorithm based on the following. Each block of the matrices $\{X^{(k)}\}_{0 \leq k \leq K-1}$, i.e., $X^{(k)}_{ij}$, is constrained as

$$X^{(k)}_{ij} = \rho_k (g_{ij} \otimes g_{ij}^{-1}),$$

shown in Equation (26). Meanwhile, such a NUG problem is relaxed into an SDP problem by relaxing the above constraints into their convex hull. Based on the nature of a convex hull, the solution matrices of an SDP problem are the convex combinations of its constraints. In other words, each block $\tilde{X}^{(k)}_{ij}$ of $\{\tilde{X}^{(k)}\}_{0 \leq k \leq K-1}$ is the convex combination of $\rho_k (g_m)$ as

$$\tilde{X}^{(k)}_{ij} = \sum_{m=1}^{G} \lambda^{rm}_{ij} \rho_k (g_m),$$

$$\lambda^{rm}_{ij} \geq 0, \quad 1 \leq i, j \leq N, \quad 1 \leq m \leq |G|$$

$$\sum_{m=1}^{G} \lambda^{rm}_{ij} = 1, \quad 1 \leq i, j \leq N,$$

where $\lambda^{rm}_{ij}$ is the coefficient of the convex combination. $\lambda^{rm}_{ij}$ is determined as

$$\lambda^{rm}_{ij} = \frac{1}{|G|} \left( \sum_{k=0}^{K-1} d_k \text{tr} \left( \rho_k (g_m) \tilde{X}^{(k)}_{ij} \right) \right)$$

by the Peter-Weyl theorem[68], i.e., Equation (21) and Equation (22).

Comparing the constraints of the block matrices of the NUG problem, i.e., Equation (27), and the solution block matrices of the SDP problem as the convex combinations of $\rho_k (g_m)$, i.e., Equation (28), the convex combination coefficient $\lambda^{rm}_{ij}$ can be considered as the “probability” of $g_{ij} \otimes g_{ij}^{-1}$ being $g_m$.

For determining $\{\tilde{g}_i\}_{1 \leq i \leq N}$ in the original NUG problem, $N-1$ pairs of relations between two of them are required. In other words, once proper $N-1$ relations, i.e., $\tilde{g}_i \otimes \tilde{g}_i^{-1}$, are determined, $\{\tilde{g}_i\}_{1 \leq i \leq N}$ can also be determined. Meanwhile, as previously stated, $\lambda^{rm}_{ij}$ indicates the “probability” of $\tilde{g}_i \otimes \tilde{g}_i^{-1}$
being \( g_m \). There are in total \( N \times N \times |G| \) guesses of every \( \tilde{g}_t_i \otimes \tilde{g}_t_j^{-1} \) being every \( g_m \) in \( G \); each has the “probability” \( \lambda_{ij}^m \). Hence, these guesses are redundant for solving \( \{ \tilde{g}_t_i \} _{1 \leq i \leq N} \).

In the present work, we propose a leaderboard algorithm for recovering \( \{ \tilde{g}_t_i \} _{1 \leq i \leq N} \) from such redundant guesses. This algorithm follows the following three principles.

The first principle of this algorithm is sequentially determining each \( \tilde{g}_t_i \). As \( G \) is a group, the following property holds. For all \( 1 \leq t_i, t_j, t_k \leq N \), the relations between \( \tilde{g}_t_i \) and \( \tilde{g}_t_j \), i.e., \( \tilde{g}_t_i \otimes \tilde{g}_t_j^{-1} \), and the relation between \( \tilde{g}_t_j \) and \( \tilde{g}_t_k \), i.e., \( \tilde{g}_t_j \otimes \tilde{g}_t_k^{-1} \), are recovered. In addition, the relation between \( \tilde{g}_t_i \) and \( \tilde{g}_t_k \), i.e., \( \tilde{g}_t_i \otimes \tilde{g}_t_k^{-1} \), is also recovered. Therefore, this sequential determination of \( \tilde{g}_t_i \) is a classic union-find[75]. In this union-find, if the relation between \( \tilde{g}_t_i \) and \( \tilde{g}_t_j \) is decided, \( \tilde{g}_t_i \) is connected to \( \tilde{g}_t_j \). Once every \( \tilde{g}_t_i \) is connected to any other \( \tilde{g}_t_j \), then \( \{ \tilde{g}_t_i \} _{1 \leq i \leq N} \) are determined.

The second principle of this algorithm is as follows. During sequentially deciding \( \{ \tilde{g}_t_i \} _{1 \leq i \leq N} \), the relations between \( \tilde{g}_t_i \) and \( \tilde{g}_t_j \) are sequentially determined in the descending order of their certainty. For example, \( \tilde{g}_t_i \otimes \tilde{g}_t_j^{-1} \) and \( \tilde{g}_t_i \otimes \tilde{g}_t_j^{-1} \) are two relations. As previously stated, the “probability” of \( \tilde{g}_t_i \otimes \tilde{g}_t_j^{-1} \) being \( g_m \) is \( \lambda_{ij}^m \). If the greatest value in \( \{ \lambda_{ij}^m \} _{1 \leq m \leq |G|} \) is greater than that in \( \{ \lambda_{ij}^m \} _{1 \leq m \leq |G|} \), then the relation between \( \tilde{g}_t_i \otimes \tilde{g}_t_j^{-1} \) is more certain than that between \( \tilde{g}_t_j \otimes \tilde{g}_t_j^{-1} \). Therefore, the relation between \( \tilde{g}_t_i \) and \( \tilde{g}_t_j \) must be determined before the relation between \( \tilde{g}_t_j \) and \( \tilde{g}_t_j \). The reason of adopting this principle is as follows. As previously stated, it is redundant for solving \( \{ \tilde{g}_t_i \} _{1 \leq i \leq N} \) by \( N \times N \times |G| \) guesses of all \( \tilde{g}_t_i \otimes \tilde{g}_t_j^{-1} \). By deciding relations with high certainty first, high quality information is used for solving the problem which contributes to the high confidence of the recovery.

The third principle of this algorithm is as follows. For the relation \( \tilde{g}_t_i \otimes \tilde{g}_t_j^{-1} \), instead of determining it solely by \( g_m \) of the single greatest value of \( \{ \lambda_{ij}^m \} _{1 \leq m \leq |G|} \), several trials are examined. These trials are based on multiple greatest values of \( \{ \lambda_{ij}^m \} _{1 \leq m \leq |G|} \). During the sequential determination of each \( \tilde{g}_t_i \), multiple sets of trials are tracked; these are selected by partial cost (see Appendix XI). Ultimately, the most qualified set of trials is selected as the final solution. This third principle makes the proposed algorithm in this work tally with the leaderboard algorithm.

5 Performance of the NUG approach

In this section, we empirically demonstrate that the NUG approach is capable of reaching the global optimal solution. The capability of solving Equation (14), i.e., selection of representatives for solving the variance and the mean of spatial rotations, is presented below (see Section 5.1). Moreover, since currently there are no other approximation algorithms for solving this problem, we have benchmarked our proposed NUG approach against a few intuitive alternatives, and have shown that the proposed NUG outperforms them all(see Section 5.2). For solving variance and mean of projection directions, the performance is similar, as Equation (16) has a similar construction to Equation (14).

5.1 Capability of reaching global optimal solution

Currently, there is no algorithm for obtaining a global optimal solution, other than the brute force approach. In the brute force approach, for each \( \tilde{g}_t_i \), there are \( |G| \) potential \( g_m \) to be considered. Therefore,
there are \(|G|^N\) trials to be examined. In other words, the computation cost grows exponentially as the number of spatial rotations grows. Hence, the global optimal is only practically solvable by the brute force approach when \(N\) is limited.

Several molecular symmetries are tested. For 3-fold cyclic \((C_3)\), 7-fold dihedral \((D_7)\), and tetrahedral \((T)\) symmetries, \(N = 5\) is used for performance tests due to the computational limitations of a modern workstation. Meanwhile, for octahedral \((O)\) and icosahedral \((I)\) symmetries, \(N\) is further limited to 3.

Such tests demonstrate that, in almost all cases, the NUG approach reaches a global optimal solution (see Table 1). Moreover, the capability of reaching a global optimal solution increases as the spatial rotations concentrate. Spatial rotations with four angular precision values[47] are tested. For the angular precision values higher than 7.9\(^\circ\), all cases tested reach a global optimal (see Table 1). Moreover, for cases in which global optimal solutions are not reached, the cost gaps between the obtained solution and the global optimal are small. For all cases tested, the maximum gap is approximately 5% (see Table 1).

5.2 Comparison with an intuitive approach

The NUG approach is compared with an intuitive approach we design, as currently there is no other approximation algorithm available for solving Equation (14). We design several intuitive approaches, and one intuitive approach performs better than the others. The superior intuitive approach is described in Appendix XII.

However, compared to the NUG approach, the intuitive approach performs rather poorly (see Table 1). For spatial rotations in a uniform distribution, i.e., angular precision equals to 90\(^\circ\), 177 \sim 533 out of 1000 tested cases do not reach a global optimal, under conditions of various molecular symmetries. Moreover, the cost gaps between the obtained solution and the global optimal solution varies from 23.0\% to 70.0\% with molecular symmetries. Such incapability to reach a global optimal is alleviated as the spatial rotations concentrate. However, even when angular precision is 1.6\(^\circ\), there are cases which fail to reach global optimal solutions. Especially, the maximum cost gaps between the obtained solutions and the global optimal remain large even though the spatial rotations concentrate.

6 Clustering considering molecular symmetry

As the distance is defined and the mean can be computed, the clustering of both spatial rotations and projection directions can be achieved by several algorithms, e.g., classic k-means clustering[76]. Due to the similar construction of the distance definition and the mean calculation between the spatial rotations and the project directions, we can restrict our exposition to projection directions; equivalent results hold for spatial rotations.

As an example, we classified projection directions into five classes with several molecular symmetries (Figure 1). One may notice that, for some classes, the regions seem to be not connected. For example, consider the classes indicated by blue and turquoise in \(C_3\), and the classes indicated by turquoise and red in \(D_7\). However, when molecular symmetry is considered, these regions are in fact connected. Taking the red class with \(D_7\) symmetry as an example, although the right small region seems apart from the left
large region, it can be rotated by a symmetry operation and then be attached to the left large region.

In contrast, the traditional method of projection directions clustering with molecular symmetry is as follows. A fundamental domain of $S^2/G$ is selected, and the projection directions are rotated by a symmetry operation into this domain. Clustering without considering molecular symmetries is then performed on these rotated projection directions. In contrast to the clustering method we propose, the true topology induced by the molecular symmetry is neglected. The closer the projection direction is to the edge of the fundamental domain, the more severe the consequences (i.e., error) of neglecting the molecular symmetry will be.

7 Discussion about potential applications

In the present work, we have enabled the statistics and clustering method for both spatial rotations and projection directions considering molecular symmetry, by developing an approximation algorithm for calculating the mean and the variance. The proposed statistics and clustering method have the potential to be adopted in a wide range of applications.

A quaternion-assisted angular reconstruction algorithm\[32, 33\] contributes to solving a series of structures at 10-30Å resolution in 90s\[34\]. The key step in this algorithm is solving the absolute orientation problem\[30, 31\], which is related to the determination of the mean of a series of spatial rotations. With this present work, it becomes feasible to solve such problems in the context of considering molecular symmetry. This might lead to rethinking the possibility of adopting a quaternion-assisted angular reconstruction algorithm in state-of-the-art cryoEM reconstruction. Furthermore, the statistics of spatial rotations may contribute to the adoption of advanced statistical method, such as particle filter\[77, 78, 8\] or Kalman filter algorithm\[79, 80, 81\], in cryoEM image processing.

For the statistics of spatial rotations, they may be adopted to resolve continuous structural changes by manifold embedding\[82, 83\]. This is because in such methods, images are grouped by projection directions for the purpose of mapping the continuum of the states of a molecule in each projection direction. Moreover, they can be used for analyzing molecule heterogeneity and misalignment. As an example, 124,810 single particle images of a Tc complex are classified by projection directions when $C_5$ symmetry is considered (Figure 2). The orientation heterogeneity of the tail exists, and approximately 12.5% images are misaligned. However, a 3.7Å resolution structure can be obtained from this dataset.

In summary, the statistics tools developed in this work form a system for analyzing spatial rotations and projection directions in cryoEM when molecular symmetry is considered. They provide more freedom and efficiency when processing and understanding images in cryoEM. We have realized the proposed methods in the open-source python package pySymStat, and it can be easily adopted or integrated in other algorithms or software.
Acknowledgements

This work was supported by fund from Advanced Innovation Center for Structural Biology (to M.H.), Yau Mathematical Sciences Center (to H.L.), and grant TH-533310008 of Tsinghua University (to H.L.). The authors are grateful to Dr. Xueming Li for providing the Tc complex dataset, and to Dr. Yichen Zhou for revising the manuscript.

Contributions

M.H., Q.Z. and H.L. initialized the project and designed the methods. Q.Z. and H.L. clarified mathematical issues and derived formulas. M.H. and Q.Z. wrote pySymStat and validated its performance. M.H. wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.
References


Table 1: Performance of the NUG approach, compared with an intuitive approach

<table>
<thead>
<tr>
<th>Method</th>
<th>Symmetry</th>
<th>$\epsilon \approx 90^\circ$</th>
<th>$\epsilon \approx 36^\circ$</th>
<th>$\epsilon \approx 7.9^\circ$</th>
<th>$\epsilon \approx 1.6^\circ$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$nW/nT$</td>
<td>$\delta$</td>
<td>$nW/nT$</td>
<td>$\delta$</td>
</tr>
<tr>
<td>NUG approach</td>
<td>C3</td>
<td>1/1000  1.78%</td>
<td>1/1000 1.15%</td>
<td>0/1000 0</td>
<td>0/1000 0</td>
</tr>
<tr>
<td></td>
<td>D7</td>
<td>3/1000  5.07%</td>
<td>1/1000 0.005%</td>
<td>0/1000 0</td>
<td>0/1000 0</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0/1000 0</td>
<td>0/1000 0</td>
<td>0/1000 0</td>
<td>0/1000 0</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>1/1000 0.002%</td>
<td>0/1000 0</td>
<td>0/1000 0</td>
<td>0/1000 0</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>0/1000 0</td>
<td>1/1000 1.74%</td>
<td>0/1000 0</td>
<td>0/1000 0</td>
</tr>
<tr>
<td>intuitive approach</td>
<td>C3</td>
<td>533/1000 23.0%</td>
<td>260/1000 33.7%</td>
<td>27/1000 8.6%</td>
<td>1/1000 18.6%</td>
</tr>
<tr>
<td></td>
<td>D7</td>
<td>511/1000 33.7%</td>
<td>441/1000 39.2%</td>
<td>90/1000 44.3%</td>
<td>5/1000 10.6%</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>444/1000 44.8%</td>
<td>367/1000 51.2%</td>
<td>57/1000 43.9%</td>
<td>2/1000 3.27%</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>485/1000 59.0%</td>
<td>429/1000 55.2%</td>
<td>78/1000 49.6%</td>
<td>6/1000 36.6%</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>177/1000 70.0%</td>
<td>177/1000 75.6%</td>
<td>47/1000 60.5%</td>
<td>3/1000 30.1%</td>
</tr>
</tbody>
</table>

* $\epsilon$ stands for the angular precision[47] of the testing spatial rotations.
** $nW$ stands for the number of cases that fail to reach a global optimum. $nT$ stands for the number of cases tested.
*** $\delta$ stands for the maximum relative cost gap between the obtained solutions and the global optimal.
Figure 1: Clustering of projection directions considering molecular symmetries. With several molecular symmetries, the projection directions are classified into five classes; these are distinguished by color. For presentation purposes, in each set of equivalent projection directions $R_{G^{-1}}(\vec{v}_p)$, an element within a given fundamental domain is selected for display. The spheres are rotated in order to place projection directions towards the reader.
Figure 2: Projection direction clustering of a Tc complex. 

**a.** 124,810 single particle images of a Tc complex are classified by their projection directions into 100 classes, when $C_5$ molecular symmetry is taken into consideration. Classes are distinguished by colors. The images from a selected class (orange) are averaged and then displayed.

**b.** The images from the selected class are further classified into 20 classes. The orientation heterogeneity of the tail is indicated by an arrow. The misaligned images are boxed by red lines.
Appendix I  The sufficient condition of two spatial rotations being equivalent considering molecular symmetry

In this section, we prove the following statement. For a molecule with $G$ symmetry, the spatial rotation $q$ is equivalent to $q \otimes g_m$ for any $g_m$ in the molecular symmetry group $G$. This statement can be paraphrased as follows. For two spatial rotations $q_1$ and $q_2$, the sufficient condition of $q_1$ and $q_2$ being equivalent considering molecular symmetry $G$ is that there exists an element $g_m$ in $G$ such that $q_1 = q_2 \otimes g_m$.

A coordinate in the map is denoted by $\vec{v}$. The value at coordinate $\vec{v}$ in this map is denoted by $h(\vec{v})$. As this map has molecular symmetry $G$,

$$h(R_{g_m}(\vec{v})) = h(\vec{v}),$$  \hspace{1cm} (30)

for all $g_m$ in $G$.

The value at the coordinate $\vec{v}$ in the rotated map with a spatial rotation $q$ is denoted by $h_q$. By the definition of the spatial rotation,

$$h_q(\vec{v}) = h(R_{q^{-1}}(\vec{v})).$$  \hspace{1cm} (31)

The sufficiency is proved as follows. As there exists an element $g_m$ in $G$ such that $q_1 = q_2 \otimes g_m$, it can be concluded that

$$h_{q_1}(\vec{v}) = h(R_{q_1^{-1}}(\vec{v}))$$

$$= h(R_{g_m \otimes q_2^{-1}}(\vec{v}))$$

$$= h(R_{g_m}(R_{q_2^{-1}}(\vec{v})))$$  \hspace{1cm} (32)

By Equation (30), it can be further concluded that

$$h_{q_1}(\vec{v}) = h(R_{q_2^{-1}}(\vec{v}))$$

$$= h_{q_2}(\vec{v}).$$  \hspace{1cm} (33)

As stated in Equation (33), the $q_1$ rotated map is identical to the $q_2$ rotated map. Hence, $q_1$ is equivalent to $q_2$. Therefore, we have proven the sufficiency of the condition.

Appendix II  The sufficient condition of two projection directions being equivalent considering molecular symmetry

In this section, we continue to use the notations and conclusions presented in Appendix I. Next, we prove the following statement. For a molecule with $G$ symmetry, the projection direction $\vec{v}_p$ is equivalent to $R_{g_m^{-1}}(\vec{v}_p)$ for any $g_m$ in a molecular symmetry group $G$.

By the concept of the projection image set defined by Equation (4), the equivalence between two projec-
tion directions can be defined by the equivalence between the two sets of projected images. As the projection operator $\mathcal{P}^q$ is defined by the operation of projecting along $Z$ axis after after rotating the map $V$ by spatial rotation $q$ and the map $V$ has molecular symmetry $G$, it stands that

$$\mathcal{P}^q(V) = \mathcal{P}^q \otimes \mathcal{E}_m(V), \quad (34)$$

for any spatial rotation $q$ and any element $g_m \in G$. Therefore, by definition of the projection image set, i.e., Equation (4), it can be concluded that

$$\mathbf{P}_{\mathbf{v}_p} = \{\mathcal{P}^q(V)\} \{q | R_q(\mathbf{v}_p) = \mathbf{v}_n\}$$

$$= \{\mathcal{P}^q \otimes \mathcal{E}_m(V)\} \{q | R_q(\mathbf{v}_p) = \mathbf{v}_n\}$$

$$= \{\mathcal{P}^q(V)\} \{q | R_q^{-1}(\mathbf{v}_m) = \mathbf{v}_n\}$$

$$= \{\mathcal{P}^q(V)\} \{q | R_q^{-1}(\mathbf{v}_m) = \mathbf{v}_n\}$$

$$= \mathbf{P}_{R_q^{-1}(\mathbf{v}_p)} \quad (35)$$

Therefore, we have proven the sufficiency of the condition.

**Appendix III  Proof of well-definedness of the distance between two spatial rotations considering molecular symmetry**

In this section, Equation (7) is proved to be a well-defined distance definition. The three requirements of distance definition in mathematics, i.e., the symmetry property, the positive definiteness property, and the triangle inequality property, are proven to be satisfied for Equation (7).

The symmetry property is such that

$$d_{SO(3)/G}(q_1 \otimes G, q_2 \otimes G) = d_{SO(3)/G}(q_2 \otimes G, q_1 \otimes G) \quad (36)$$

holds for all $R_{q_1}, R_{q_2} \in SO(3)$. It is proven as follows.

$$d_{SO(3)/G}(q_1 \otimes G, q_2 \otimes G) = \min_{g_1', g_2'} d_{SO(3)}(q_1 \otimes g_1, q_2 \otimes g_2)$$

$$= \min_{g_1', g_2'} d_{SO(3)}(q_2 \otimes g_2, q_1 \otimes g_1) \quad (37)$$

as $d_{SO(3)}(q_1 \otimes g_1, q_2 \otimes g_2) = d_{SO(3)}(q_2 \otimes g_2, q_1 \otimes g_1)$ holds, because $d_{SO(3)}$ is also a distance, therefore
it has the symmetry property.

The positive definiteness property contains two parts. The first part is that

\[ d_{SO(3)/G}(q_1 \otimes G, q_2 \otimes G) \geq 0 \]

(38)

holds for all \( R_{q_1}, R_{q_2} \in SO(3) \), which can be derived by \( d_{SO(3)}(q) \geq 0 \) for all \( R_q \in SO(3) \). The second part is that

\[ d_{SO(3)/G}(q_1 \otimes G, q_2 \otimes G) = 0 \]

(39)

if and only if \( q_1 \otimes G = q_2 \otimes G \). This statement is proven as follows. When \( q_1 \otimes G = q_2 \otimes G \), there exist \( g_{t_1} \) and \( g_{t_2} \) such that \( q_1 \otimes g_{t_1} = q_2 \otimes g_{t_2} \). Thus, \( d_{SO(3)}(q_1 \otimes g_{t_1}, q_2 \otimes g_{t_2}) = 0 \), as \( d_{SO(3)} \) also has the positive definiteness property. Therefore, \( d_{SO(3)/G}(q_1 \otimes G, q_2 \otimes G) = 0 \). Conversely, when \( d_{SO(3)/G}(q_1 \otimes G, q_2 \otimes G) = 0 \), there exist \( g_{t_1} \) and \( g_{t_2} \) such that \( q_1 \otimes g_{t_1} = q_2 \otimes g_{t_2} \). If \( q_1 \otimes G \neq q_2 \otimes G \), without losing generality, there exists \( g_{t_1}' \) such that \( q_1 \otimes g_{t_1}' \notin q_2 \otimes G \). Therefore, \( q_2 = q_1 \otimes g_{t_1} \otimes g^{-1}_{t_2} = (q_1 \otimes g_{t_1}') \otimes (g^{-1}_{t_1} \otimes g_{t_1} \otimes g^{-1}_{t_2}) \) implies that \( g^{-1}_{t_1} \otimes g_{t_1} \otimes g^{-1}_{t_2} \notin G \), which conflicts with the property of the group \( G \). Thus, \( q_1 \otimes G = q_2 \otimes G \).

The triangle inequality property is that

\[ d_{SO(3)/G}(q_1 \otimes G, q_2 \otimes G) \leq d_{SO(3)/G}(q_1 \otimes G, q_3 \otimes G) + d_{SO(3)/G}(q_3 \otimes G, q_2 \otimes G) \]

(40)

holds for all \( R_{q_1}, R_{q_2}, R_{q_3} \) in \( SO(3) \). It is proven as follows.

\[
d_{SO(3)/G}(q_1 \otimes G, q_2 \otimes G) = \min_{g_{t_1}, g_{t_2} \in G} d_{SO(3)}(q_1 \otimes g_{t_1}, q_2 \otimes g_{t_2})
\]

(41)

\[
\leq \min_{g_{t_1}, g_{t_2}, g_{t_3} \in G} (d_{SO(3)}(q_1 \otimes g_{t_1}, q_3 \otimes g_{t_3}) + d_{SO(3)}(q_3 \otimes g_{t_3}, q_2 \otimes g_{t_2}))
\]

as \( d_{SO(3)}(q_1 \otimes g_{t_1}, q_2 \otimes g_{t_2}) \leq d_{SO(3)}(q_1 \otimes g_{t_1}, q_3 \otimes g_{t_3}) + d_{SO(3)}(q_3 \otimes g_{t_3}, q_2 \otimes g_{t_2}) \) holds for any \( g_{t_3} \in G \), because \( d_{SO(3)} \) also has the triangle inequality property.

Moreover, distance in \( SO(3) \), i.e., \( d_{SO(3)} \), is invariant under left and right multiplication[1]. In other words, \( d_{SO(3)} \) has the following properties,

\[ d_{SO(3)}(q_1, q_2) = d_{SO(3)}(q_1 \otimes q_3, q_2 \otimes q_3) \]

(42)

and

\[ d_{SO(3)}(q_1, q_2) = d_{SO(3)}(q_3 \otimes q_1, q_3 \otimes q_2) \]

(43)
for any \( R_{q_1}, R_{q_2}, \) and \( R_{q_3} \) in \( SO(3) \). Therefore, Equation (41) can be further converted into

\[
\begin{align*}
d_{SO(3) / G}(q_1 \otimes G, q_2 \otimes G) \\
&\leq \min_{g_1, g_2, g_3 \in G} (d_{SO(3)}(q_1, q_3 \otimes g_1 \otimes g_2^{-1}) + d_{SO(3)}(q_3 \otimes g_3 \otimes g_3^{-1}, q_2)) \\
&= \min_{g} (d_{SO(3)}(q_1, q_3 \otimes g) + d_{SO(3)}(q_3 \otimes g^*, q_2)) \\
&= \min_{g} d_{SO(3)}(q_1, q_3 \otimes g) + \min_{g} d_{SO(3)}(q_3 \otimes g, q_2) \\
&= \min_{g_1, g_2} d_{SO(3)}(q_1 \otimes g_1, q_3 \otimes g_2) + \min_{g_3, g_4} d_{SO(3)}(q_3 \otimes g_3, q_2 \otimes g_4).
\end{align*}
\]

Therefore, we have proven the suitability of Equation (7) as a distance definition.

**Appendix IV  Proof of well-definedness of distance definition between two projection directions considering molecular symmetry**

In this section, Equation (8) is proven to be a well-defined distance definition. Three requirements of distance definition in mathematics, i.e., the symmetry property, the positive definiteness property and the triangle inequality property, are proven to satisfy Equation (8).

As the distance definition between two projection directions considering molecular symmetry, i.e., Equation (8), follows the same structure of the definition between two spatial rotations considering molecular symmetry, i.e., Equation (7), the symmetry property, the first part of the positive definiteness, and the triangle inequality property can be proven with the same procedure in Appendix III.

The second part of the positive definiteness is that

\[
d_{SO(3)/G}(R_{G^{-1}}(\tilde{v}_{p_1}), R_{G^{-1}}(\tilde{v}_{p_2})) = 0
\]

if and only if \( R_{G^{-1}}(\tilde{v}_{p_1}) = R_{G^{-1}}(\tilde{v}_{p_2}) \). This statement is proven as follows. When \( R_{G^{-1}}(\tilde{v}_{p_1}) = R_{G^{-1}}(\tilde{v}_{p_2}) \), there exist \( g_{t_1} \) and \( g_{t_2} \) such that \( R_{g_{t_1}}^{-1}(\tilde{v}_{p_1}) = R_{g_{t_2}}^{-1}(\tilde{v}_{p_2}) \). As \( d_{SO(3)/G} \) also has positive definiteness property, \( d_{SO(3)/G}(R_{g_{t_1}}^{-1}(\tilde{v}_{p_1}), R_{g_{t_2}}^{-1}(\tilde{v}_{p_2})) = 0 \). Therefore, \( d_{SO(3)/G}(R_{G^{-1}}(\tilde{v}_{p_1}), R_{G^{-1}}(\tilde{v}_{p_2})) = 0 \). Conversely, when \( d_{SO(3)/G}(R_{G^{-1}}(\tilde{v}_{p_1}), R_{G^{-1}}(\tilde{v}_{p_2})) = 0 \), there exist \( g_{t_1} \) and \( g_{t_2} \) such that \( R_{g_{t_1}}^{-1}(\tilde{v}_{p_1}) = R_{g_{t_2}}^{-1}(\tilde{v}_{p_2}) \). If \( R_{G^{-1}}(\tilde{v}_{p_1}) \neq R_{G^{-1}}(\tilde{v}_{p_2}) \), without losing generality, there exists \( g_{t_1} \) such that \( R_{g_{t_1}}^{-1}(\tilde{v}_{p_1}) \notin R_{G^{-1}}(\tilde{v}_{p_2}) \). Therefore, \( R_{g_{t_1}}^{-1}(\tilde{v}_{p_1}) = R_{g_{t_1}^{-1} \otimes g_{t_1}}^{-1}(R_{g_{t_1}}^{-1}(\tilde{v}_{p_1})) = R_{g_{t_1}^{-1} \otimes g_{t_1}^{-1}}^{-1}(R_{g_{t_1}}^{-1}(\tilde{v}_{p_2})) = R_{g_{t_1}^{-1} \otimes g_{t_1}^{-1} \otimes g_{t_2}^{-1}}^{-1}(\tilde{v}_{p_2}) \notin R_{G}(\tilde{v}_{p_2}) \), which conflicts with the property of the group \( G \). Thus, \( R_{G^{-1}}(\tilde{v}_{p_1}) = R_{G^{-1}}(\tilde{v}_{p_2}) \).

Therefore, we have proven the suitability of Equation (8) as a distance definition.
Appendix V  Approximation of the mean of spatial rotations by representatives selected in variance determination

In this section, we state that it is rational to approximate the mean of spatial rotations, i.e., Equation (11), by Equation (13).

By observing the definition of mean and the definition of variance of spatial rotations, i.e., Equation (11) and (9), respectively, one can find that the spirit of the selection of \( g_{t_i} \) for the mean and the variance are the same. In other words, \( g_{t_i} \) is selected for each \( q_i \) to make \( q_i \otimes g_{t_i} \) as close as possible overall. Therefore, it is natural to replace the \( g_{t_i} \) in the definition of the mean by the \( g_{t_i} \) in the definition of the variance. Thus, Equation (13) is a rational approximation of Equation (11).

Appendix VI  Approximation of mean of projection directions by representatives selected in variance determination

In this section, we state that it is rational to approximate the mean of projection directions, i.e., Equation (12), by Equation (15).

As the relation between the mean and the variance of spatial rotations is identical to those of the projection directions, the rationality of approximating Equation (12) by Equation (15) can be explained by using the same procedure as in Appendix V.

Appendix VII  Derivation of representative determination of spatial rotations into a NUG form

In this section, we prove that Equation (14) can be derived into a NUG form, i.e., Equation (18).

Distance in \( SO(3) \), i.e., \( d_{SO(3)} \), is invariant under left and right multiplications[1], i.e., Equation (42) and Equation (43). Thus, Equation (14) can be derived as

\[
\{\tilde{g}_{t_i}\}_{1 \leq i \leq N} = \text{argmin}_{g_{t_1}, g_{t_2}, \ldots, g_{t_N} \in G} \left( \sum_{1 \leq i,j \leq N} d^2_{SO(3)}(q_i \otimes g_{t_i}, q_j \otimes g_{t_j}) \right)
\]

\[
= \text{argmin}_{g_{t_1}, g_{t_2}, \ldots, g_{t_N} \in G} \left( \sum_{1 \leq i,j \leq N} d^2_{SO(3)}(q_i \otimes g_{t_i}, g_{t_j}^{-1} \otimes q_j) \right)
\]

(46)

By assigning \( f_{ij}(g_{t_i} \otimes g_{t_j}^{-1}) \) as Equation (19), we have proven that Equation (14) can be derived into a NUG form, i.e., Equation (18).
Appendix VIII Derivation of representative determination of projection directions into a NUG form

In this section, we prove that Equation (16) can be derived into a NUG form, i.e., Equation (18).

The distance in $S^2$, i.e., $d_{S^2}$ is invariant under spatial rotation. In other words, $d_{S^2}$ has the following property,

$$d_{S^2}(\vec{v}_{p_1}, \vec{v}_{p_2}) = d_{S^2}(R_q(\vec{v}_{p_1}), R_q(\vec{v}_{p_2}))$$

(47)

for any projection directions $\vec{v}_{p_1}$ and $\vec{v}_{p_2}$, and any spatial rotation $q$.

Thus, Equation (16) can be derived into

$$\{\tilde{g}_t\}_{1 \leq t \leq N} = \arg\min_{g_{t_1}, g_{t_2}, \ldots, g_{t_N} \in G} \left( \sum_{1 \leq i, j \leq N} d_{S^2}^2 \left( R_{g_{t_i}^{-1}} (\vec{v}_{p_i}), R_{g_{t_j}^{-1}} (\vec{v}_{p_j}) \right) \right)$$

$$= \arg\min_{g_{t_1}, g_{t_2}, \ldots, g_{t_N} \in G} \left( \sum_{1 \leq i, j \leq N} d_{S^2}^2 \left( \vec{v}_{p_i}, R_{g_{t_i}} (R_{g_{t_j}^{-1}} (\vec{v}_{p_j})) \right) \right)$$

(48)

By assigning $f_{ij}(g_{t_i} \odot g_{t_j}^{-1})$ as Equation (20), we have proven that Equation (16) can be derived into a NUG form, i.e., Equation (18).

Appendix IX Fourier expansion of a NUG, and a matrix form

In this section, we prove that Equation (18) is equivalent to Equation (23).

As $G$ is a molecular symmetry group, $G$ is compact. By the Peter-Weyl Theorem[2], the generalized Fourier expansion of $f_{ij}(g_{t_i} \odot g_{t_j}^{-1})$ is

$$f_{ij}(g_{t_i} \odot g_{t_j}^{-1}) = \sum_{k=0}^{K-1} d_k \text{tr} \left( \hat{f}_{ij}^{(k)} (g_{t_i} \odot g_{t_j}^{-1}) \right)$$

(49)

where $\hat{f}_{ij}^{(k)}$ is determined by Equation (21).
Thus, Equation (18) can be derived into

$$\text{argmin}_{g_1, g_2, \ldots, g_N \in G} \sum_{1 \leq i, j \leq N} f_{ij}(g_t \otimes g_t^{-1})$$

$$= \text{argmin}_{g_1, g_2, \ldots, g_N \in G} \sum_{1 \leq i, j \leq N} \sum_{k=0}^{K-1} d_k \text{tr} \left( \hat{f}_{ij}^{(k)} \rho_k(g_t \otimes g_t^{-1}) \right)$$

$$= \text{argmin}_{g_1, g_2, \ldots, g_N \in G} \sum_{k=0}^{K-1} \sum_{1 \leq i, j \leq N} \text{tr} \left( \left( d_k \hat{f}_{ij}^{(k)} \right) \left( \rho_k(g_t \otimes g_t^{-1}) \right) \right)$$

$$= \text{argmin}_{g_1, g_2, \ldots, g_N \in G} \sum_{k=0}^{K-1} \text{tr} \left( \hat{F}^{(k)} X^{(k)} \right)$$

where $\hat{F}^{(k)}$ is determined by Equation (24) and $X^{(k)}$ is determined by Equation (26). Hence, we have proven that Equation (18) is equivalent to Equation (23).

**Appendix X  SDP relaxation of the matrix form of the NUG problem**

The variable matrices $\{X^{(k)}\}_{1 \leq k \leq K}$ is defined by Equation (26), in the matrix form of the NUG problem. However, the constraints of such matrices are in general non-convex. For the purpose of convex optimization, the constraints of $\{X^{(k)}\}_{1 \leq k \leq K}$ are relaxed by the method proposed in [3]

$$X^{(k)} \succeq 0 \quad 0 \leq k \leq K-1$$

$$X_{ii}^{(k)} = I_{d_k \times d_k} \quad 0 \leq k \leq K-1, \quad 1 \leq i \leq N$$

$$X_{ij}^{(0)} = I_{d_0 \times d_0} \quad 1 \leq i, j \leq N$$

$$\sum_{k=0}^{K-1} d_k \text{tr} \left( \rho_k(g_m)^* X_{ij}^{(k)} \right) \geq 0 \quad 1 \leq i, j \leq N, \quad 1 \leq m \leq |G|$$

where $X^{(k)} \succeq 0$ stands for that $\{X^{(k)}\}_{1 \leq k \leq K}$ are semi-definite positive (SDP) matrices.

Therefore, the NUG problem is relaxed into a SDP problem, as

$$\left\{ \tilde{X}^{(k)} \right\}_{0 \leq k \leq K-1} = \text{argmin}_{\{X^{(k)}\}_{0 \leq k \leq K-1}} \sum_{k=0}^{K-1} \text{tr} \left( \hat{F}^{(k)} X^{(k)} \right),$$

where $\{X^{(k)}\}_{1 \leq k \leq K}$ are subject to the above relaxed constraint. The solution matrices $\left\{ \tilde{X}^{(k)} \right\}_{0 \leq k \leq K-1}$ are in the convex hull of the matrices defined in Equation (26).
Appendix XI Partial cost

Solving $\{\tilde{g}_i\}_{1 \leq i \leq N}$ minimizes the cost function Equation (18). Thus, during the sequential recovery of representatives from the solution of the SDP problem, the partial cost must be calculated at each recovery step for the purpose of determining the keeping trials. As in Equation (18), the cost $\sum_{1 \leq i, j \leq N} f_{ij} (g_i \otimes g_j^{-1})$ is the summation of $f_{ij}$ of all relations $\{\tilde{g}_i \otimes \tilde{g}_j^{-1}\}_{1 \leq i, j \leq N}$. Therefore, the partial cost can be calculated as

$$\sum_{1 \leq i, j \leq N} f_{ij} (g_i \otimes g_j^{-1}).$$ (52)

In other words, the partial cost is the summation of $f_{ij}$ of all determined relations.

Appendix XII Design of an intuitive approach

In the section, an intuitive approach for solving Equation (14) is described.

The principle of obtaining $\{\tilde{g}_i\}_{1 \leq i \leq N}$ from Equation (14) is that a representative $q_i \otimes g_{ti}$ is chosen from each set of equivalent spatial rotations $q_i \otimes G$, such that all representatives are as close as possible overall. In this intuitive approach, this principle is realized as follows.

First, a spatial rotation $\tilde{q}$ is selected as follows. For each $q_i \otimes G$ from $\{q_i \otimes G\}_{1 \leq i \leq N}$, the overall distance between it and the other set of equivalent spatial rotations is calculated as

$$\sum_{1 \leq j \leq N, j \neq i} d_{SO(3)/G}(q_i \otimes G, q_j \otimes G).$$ (53)

The set of equivalent spatial rotations with minimum such distance can be considered at the “center” of $\{q_i \otimes G\}_{1 \leq i \leq N}$. Therefore, the spatial rotation $\tilde{q}$ is selected as an element of the “center” set of equivalent spatial rotations.

Then, from each set of equivalent spatial rotations $q_i \otimes G$, the one closest to the “center” $\tilde{q}$, i.e., $q_i \otimes \tilde{g}_i$, is selected as the representative. Therefore, $\tilde{g}_i$ is determined for a spatial rotation set $q_i \otimes G$.

By this means, a representative $q_i \otimes g_{ti}$ is chosen from each set of equivalent spatial rotations $q_i \otimes G$. Such a representative is closest to the “center” $\tilde{q}$. The principle of obtaining $\{\tilde{g}_i\}_{1 \leq i \leq N}$ from Equation (14) is realized to some extent.
References

