Twist and Scout: Analysis and Curation of Particles in Cryo-Electron Tomography Using TANGO

Markus Schreiber[a,b], Beata Turoňová[a,*]

[a]Department of Molecular Sociology, Max Planck Institute of Biophysics, Frankfurt am Main, Germany
[b]IMPRS on Cellular Biophysics, Frankfurt am Main, Germany

ARTICLE INFO

Keywords:
Point Clouds
cryo-ET
Subtomogram Averaging
Contextual analysis
Twist coordinates
Lie groups

ABSTRACT

Cryo-electron tomography (cryo-ET) enables the visualization of cellular structures in near-native environments, but its potential for spatial analysis has been underutilized due to a lack of versatile tools accommodating biological sample diversity. Available solutions often rely on case-specific or hypothesis-driven approaches, while holistic analyses remain challenging. In this work, we introduce TANGO (Twist-Aware Neighborhoods for Geometric Organization), a novel framework leveraging point cloud descriptors to analyze spatial arrangements of particles, such as macromolecular complexes, in cryo-ET. By encoding relative positions and orientations of particles as twist vectors, TANGO enables rotationally invariant feature extraction, including structured neighborhood occupancy, lattice topology, or angular deviations. Its modular design and user-friendly interface allow for customization of features, facilitating exploratory analyses of spatial patterns in diverse experimental datasets. With its open-source Python implementation, TANGO advances the ability to decode complex cellular architectures and their functional relationships, offering a valuable tool for the cryo-ET community.

1. Introduction

Cryo-electron tomography (cryo-ET) provides unique structural insights into the landscape of cells in their close to native environment. The basic workflow consists of the acquisition of 2-dimensional (2D) images of a frozen specimen at different tilt angles that are subsequently used for the reconstruction of a 3-dimensional (3D) volume called tomogram. To gain access to high-resolution information, cryo-ET is often used in combination with subtomogram averaging (STA). STA is based on finding a common alignment among instances, referred to as particles, of the same cellular feature, such as macromolecular complexes or membranes. The averaging of the aligned particles reduces the anisotropic resolution and improves the SNR, resulting in a structure that can nowadays reach resolution below 5Å.

However, the range of information extracted by STA goes beyond high-resolution structures. It provides us with positions and orientations of the particles of interest, which allows for further analysis of their mutual relationships, as well as of their spatial relations to other cellular features. Spatial analysis enhances our understanding of the cellular ultrastructure and can result in additional valuable insights other than the structure itself would provide [1, 2, 3, 4, 5]. Despite this great potential, the tools to perform such analyses remain limited to specialized scripts and hypothesis-driven solutions [1, 4]. A notable exception is presented by the work on statistical particle pattern analysis [6], which aims at the quantitative description of particle clustering based solely on their positions.

Among specialized tools, Jiang et al. [7] present a tool that was designed to study polyribosomes. It is based on the clustering of both relative positions and orientations; such an approach is related to the framework presented in our study, which is rooted in the interplay of positional and orientational information.

To our knowledge, there are no tools for the holistic spatial analysis of particles that are not tailored to a specific use case. This work aims at closing this gap by providing a framework based on point cloud descriptors [8].

This challenge may be addressed using point clouds that are discrete sets of points in space, for example 3D points describing the shape of an object by densely sampling its surface. In such a case, points can hold additional information such as normal vectors or assigned colors. Similar cases can be summed up under geometric shape analysis, which utilizes coordinates from various scanning methods [9, 10]. Other areas where point clouds can be applied include biology, where points can, for example, represent the positions of atoms in 3D Euclidean space [11, 12]. In another case, a point cloud can be embedded in a space representing virus genomes [13].

The underlying concept of point cloud descriptors is to characterize each point based on the geometric properties of its local neighborhood and encode such information in a feature vector. Practically, one first defines the desired local neighborhood, referred to as support. This is often a simple sphere or cube with a given size. Then, for each point, referred to as query point, all points within its support are fetched and used to compute the feature vector, also called descriptor. Features can range from simple ones, such as the number of points within the support, to very elaborate schemes, such descriptors based on spherical harmonics [14], and can capture patterns like surface curvature

*Corresponding author
Beata.Turonova@biophys.mpg.de (B. Turoňová)
ORCID(s): 0009-0008-4485-4792 (M. Schreiber); 0000-0002-5457-4478 (B. Turoňová)

[15], neighbor distribution, or orientations [16] in a well-structured, compact way. This allows for tasks such as object recognition, pattern matching, registration, surface segmentation, and similarity comparison to be performed reliably and efficiently [17, 18, 19]. A simple example for a point cloud descriptor is the signature of histograms of orientations (coined SHOT) [20], which captures local surface shape and orientation information by computing histograms of normals around a point on a predefined spherical support. These histograms build the feature vectors for each point which is used for pattern matching either by searching for a certain signature or by clustering the signatures into similar groups. In the area of structural biology, point descriptors have been used for analyzing protein surfaces and molecular docking by means of point cloud registration [21].

In cryo-ET, one can consider the positions of particles to form a point cloud. However, what sets particles apart from classical point clouds is the additional information on particle orientation. In general, point clouds do not need to carry orientation information; the closest notion of their orientation comes with normals referring to an underlying surface (if provided). In contrast, particles possess full rotational information and one can thus examine relative orientations among all particles within a support. Moreover, these rotations encode a canonical reference orientation which allows for all feature vectors to be designed in a rotational invariant manner.

In this work, we explore the potential of point cloud descriptors for holistic analyses of particles from cryo-ET. We apply so-called twist vectors which encode relative positions and relative orientations of pairs of particles [22] in a rotationally invariant manner. They build the core of the proposed framework called **TANGO - Twist-Aware Neighborhoods for Geometric Organization**. For each particle, twist vectors are determined for all neighbors within a fixed radius around that particle: the given particle's spherical support. This forms the basis that allows for an easy derivation of other features such as neighborhood occupancy, lattice geometry, angular deviations, and more. The choice of features to use depends on the data and questions at hand.

The outcome of the feature vector computation offers multiple possibilities for further data analysis as we demonstrate in this work. First, the results can be directly used to describe particles in a quantitative or statistical manner, or determine the presence of cyclical symmetry. Vectors can be further clustered to find affiliations of particles to their underlying objects, as demonstrated on microtubules (MTs) and nuclear pore complexes (NPCs). Clustering can be also used to find outliers in data to either exclude them or to examine them further. We show this on immature virus-like particles of the HIV-1 CA-SP1 lattice, where we could identify and exclude particles with this approach, and also on capsids of mature HIV-1, where the found outliers presented pentamers within the hexagonal lattices. Finally, we demonstrate on synthetic data, as well as on experimental data, how point descriptors can be used to find patterns among particles.

TANGO is not only a framework, but it also comes with an open-source Python-based implementation which is further enhanced by a graphical user interface (GUI). Data analysis starts with the computation of twist vectors on a sufficiently large spherical support. Then, to ensure flexibility, one can choose from a catalog of different supports (ranging from simple spheres to custom ones provided by a user) and features (relative orientation, relative position, lattice geometry, neighborhood occupancy, angular scores, ...) to proceed with the analysis. Due to the precomputed twist vectors, all features offered in the catalog can be efficiently computed, enabling an exploratory analysis of the data. The user can choose a single feature or combine multiple ones into one feature vector to see which works best for their data. This is facilitated by the GUI, where one can easily judge the outcome of the chosen type of support and features. Moreover, the modularity of the implementation allows for the addition of custom features and supports if needed.

Overall, the flexibility of TANGO allows for a more holistic characterization of spatial arrangements of particles and, together with its user-friendly interface, it can become a useful addition to data analysis and processing within the cryo-ET community.

The outcome of the feature vector computation offers multiple possibilities for further data analysis. It can be used for quantitative description, as well as more advanced applications such as symmetry detection, affiliation of particles to their underlying objects, or finding patterns among particles. We demonstrate the potential of this approach in our work using a variety of particle types, including microtubules, nuclear pore complexes, and virus-like particles such as HIV-1, showcasing its versatility and broad applicability.

TANGO is not only a framework, but it also comes with an open-source Python-based implementation which is further enhanced by a graphical user interface (GUI). The implementation allows for flexibility and customization, enabling users to choose from a range of features and supports to suit their analysis needs.

Overall, the development of TANGO addresses the need for a holistic characterization of spatial arrangements of particles in cryo-ET, providing a foundation for further research and analysis in this field. By enabling the analysis of particle affiliations and patterns, TANGO has the potential to reveal new insights into the organization and behavior of cellular structures, as demonstrated through its application in a range of examples.

2. Methods

2.1. Workflow

The approach proposed in this study is based on point cloud descriptors (PCD). Going through each point in the data set, the query points, data points with positions within a defined query point's neighborhood, the support, are collected and used for the computation of features that describe the neighborhood's geometric properties. The collection of features yields a feature vector for each query point, also

known as a point descriptor. These descriptors form the base for further analysis of point clouds. In our setup, a point cloud is created from a list of particles, which can stem directly from particle picking or from further iterative refinement using STA. In summation, apart from positions $p \in \mathbb{R}^3$, we also handle the orientation of each particle. The latter refers to a rotation that is needed to achieve a canonical orientation (an orientation of the reference used in STA or particle picking). Thus, particles are positioned in a fixed extrinsic frame of reference, inside of which particle orientations are compared with the canonical one.

To use the full potential of the available particle information, and to accommodate the diversity of the analysis needed, we introduce two changes into the classical PCD workflow: First, we consider each query particle in its canonical orientation and thus rotate the respective support accordingly. This guarantees orientational alignment of the individual supports, ensuring rotational invariance of the descriptors that are computed in this manner. The second change is the hierarchical approach. We first compute the twist descriptor (described below) on larger spherical supports. This descriptor provides essential information on neighborhood distributions and can then be used to derive additional features based on the data and question at hand. Features can be arbitrarily combined, thereby yielding new descriptors that best reflect the needs for the given data analysis. In this second step, an arbitrary support can be chosen either from the provided support catalog or custom-made in form of a binary mask. Furthermore, twist descriptors can be used to filter out points within the support based on, for example, their angular distance from the query point. The final step of our framework is the analysis of the computed descriptor(s), which is data-dependent. This step can include quantitative or statistical characterizations of the particles' spatial distribution, the clustering of descriptors in order to find patterns within the data, or the assignment of particles to their underlying objects. The complete workflow is depicted in Figure 1.

2.2. Twist vectors

In mathematical terms, orientations can be described by rotation matrices $\omega \in SO(3)$ [23], where the special orthogonal group $SO(3)$ consists of orthogonal matrices of determinant 1:

$$SO(3) = \{\omega \in \mathbb{R}^{3\times 3} \mid \omega^T = \omega^{-1}, \det(\omega) = 1\}. \quad (1)$$

An object characterized by coordinates centered around the origin $\mathbb{0}$ in 3D Euclidean space $\mathbb{R}^3$ can thus be rotated by having a matrix $\omega \in SO(3)$ act on each of these coordinates. The object itself remains rigid; it is neither sheared nor scaled. By combining this action with a translation as described by a vector $p \in \mathbb{R}^3$, transporting the object so that its coordinates are centered around p, results in a rigid motion [24]. A graphical depiction of such a motion can be found in Supplementary Figure3.

In this sense, a particle can be associated with an element of the group of rigid motions $(\omega, p) \in SE(3)$, which, as a manifold, is the product $SO(3) \times \mathbb{R}^3$. In matrix form, (ω, p) can be explicitly handled as

$$\begin{pmatrix} \omega & p \\ \mathbb{0} & 1 \end{pmatrix}. \quad (2)$$

The idea of an extrinsic frame of reference to a canonical particle orientation has a natural equivalent in this context. Among rigid motions, there is a distinguished point $\mathbb{1}_4 \in SE(3)$, the 4×4 identity matrix, which can be interpreted as a canonical particle, the one positioned at the origin of $\mathbb{R}^3$ equipped with the canonical orientation $\mathbb{1}_3 \in SO(3)$. By abuse of notation, $\mathbb{1}$ is used to denote the identity matrix when the dimension is implied.

The philosophy of using "rigid motions" and "particles" synonymously implies the existence of a product of particles, as well as that of an inverse of particles. Biologically speaking, this is of course nonsensical. However, from a computational point of view, the idea of aligning supports is rooted in such operations.

The rigid motions of $SE(3)$ act on particle lists (finite subsets of $SE(3)$) in the following way:

$$(\omega, p)^{-1} \cdot (\omega, p) = \mathbb{1}_4 \quad (3)$$

$$(\omega_1, p_1)^{-1} \cdot (\omega_2, p_2) = (\omega_1^T \omega_2, \omega_1^T (p_2 - p_1)) \quad (4)$$

Having the inverse of a query particle $(\omega_1, p_1) \in SE(3)$ act on a collection of particles as described in (4) is equivalent to shifting and rotating the collection so that the query particle aligns with $\mathbb{1}$, compare (3), keeping the relative positions and orientations unchanged. Thus, rotation invariance of descriptors building up on this concept is guaranteed. Such a normalization reduces the degrees of freedom when comparing local particle neighborhoods.

The most fundamental descriptor in our tool satisfies this condition in a natural way – twist vectors encoding relative orientations and relative positions between a query particle and all particles within the support. Forming a subset of a vector space, such sets lend themselves to the application of conventional statistical methods. They can be constructed by considering not only the algebraic structure of $SE(3)$, but also its geometry, for particle data analysis. The group of rigid motions is a smooth manifold, which implies that neighborhoods around each point $x \in SE(3)$ can be linearly approximated through the tangent space $T_x SE(3) \cong \mathbb{R}^6$ [25]. Of special interest for us is the tangent space at the distinguished point: $\mathfrak{se}(3) := T_{\mathbb{1}} SE(3) = T_{\mathbb{1}} SO(3) \times \mathbb{R}^3 = \mathfrak{so}(3) \times \mathbb{R}^3$. The reasoning behind this once more can be traced back to the desire to align supports.

Finally, the space $\mathfrak{se}(3)$ is considered as the space of twist vectors:

$$\xi = \begin{pmatrix} 0 & -\zeta_z & \zeta_y & p_x \\ \zeta_z & 0 & -\zeta_x & p_y \\ -\zeta_y & \zeta_x & 0 & p_z \\ 0 & 0 & 0 & 0 \end{pmatrix} \in \mathfrak{se}(3). \quad (5)$$

Note that any twist vector ξ splits into a tangent ζ to $SO(3)$ at $\mathbb{1}$, given by a skew-symmetric matrix, and a vector $p \in \mathbb{R}^3$.

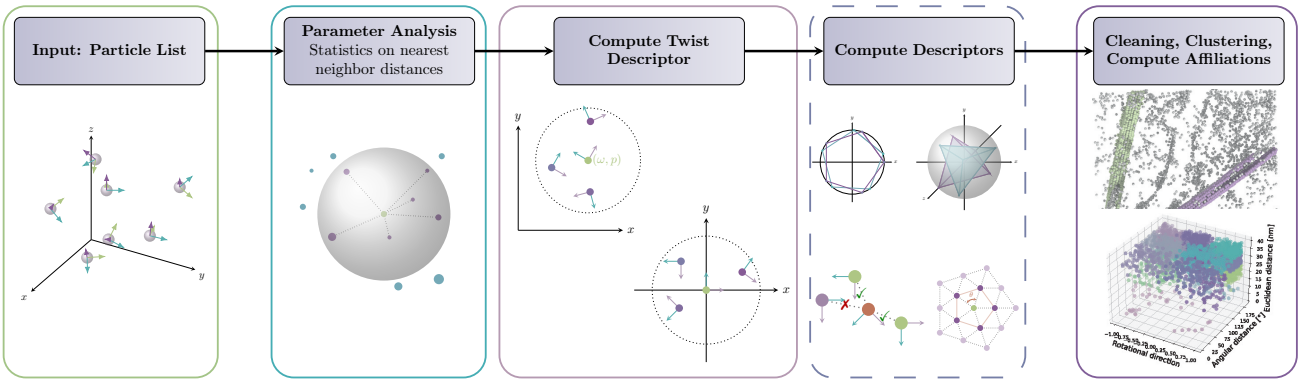

Figure 1: Starting with a particle list, parameters relevant for the analysis, such as nearest neighbor distances, are extracted. These help to define the support on which the twist descriptor is computed. By default, rigid motions ($(\omega, p)^{-1}$) are applied to each support to align them with the query particle's canonical orientation (middle panel). Using the twist descriptor as basis, additional descriptors might be computed if needed. The final step is the analysis of the descriptor(s) which can be tailored to the task at hand; be it finding patterns in the data using clustering of the descriptor features, identifying outliers, or assigning affiliations to underlying objects.

The coordinates $(\zeta_x, \zeta_y, \zeta_z, p_x, p_y, p_z) \in \mathbb{R}^6$ are so-called twist coordinates [22]. How exactly do twist coordinates characterize the geometric interplay of a pair of particles? In the context of particles, given $(\omega_i, p_i) \in SE(3), i = 1, 2$, equation 4 describes the relative pose, say $(\omega_1^T \omega_2, \omega_1^T (p_2 - p_1)) =: (\omega, p)$. The twist vector encoding the relevant information is thus the tangent vector at $\mathbb{1}$ pointing in the direction of (ω, p). Using the matrix logarithm $\log : SO(3) \to \mathfrak{so}(3)$ [26], the twist in question is given by $(\log(\omega), p)$.

In summation, twist coordinates encode to what degree relative positions are pronounced within the individual coordinate directions. The same holds for relative orientations, where the coordinate directions in $\mathfrak{so}(3)$ are related to rotations around the respective axes.

Further information to be gained from these tangent vectors are their norms, which can be interpreted as geodesic distances. In the case of the positional part, this yields the Euclidean distance. For the orientational part, the geodesic distance is also known as angular distance [27] and is given by

$$d_{SO(3)}(\omega_1, \omega_2) = \| \log(\omega_1^T \omega_2) \|_F, \tag{6}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. The overall norm of the twist vector is the product metric for $SE(3)$.

2.3. Particle symmetry

In addition to the information encoded in the twist descriptor, an extension especially for particles of two classes of symmetries is provided. The analysis of relative orientations between symmetric particles comes with some ambiguity. As an example, a particle showing six-fold rotational symmetry around a given axis has six different choices of rotation matrices representing the same orientation. This range of options needs to be accounted for when considering the orientational differences between two given particles of the same symmetry type. Naturally, one can consider the angular distances between all 36 possible pairs, but the computational expense would grow with more intricate types of symmetry. In our framework, we resolve this ambiguity for two classes of rotationally symmetric particles: cyclically symmetric ones (C_n-symmetry for $n > 1$) and those that show symmetry of a platonic solid (tetrahedron, cube, octahedron, dodecahedron, icosahedron [28]), see Figure 2.

Given a symmetry group from one of the above cases, the underlying idea for an unambiguous way to distinguish between orientations is to consider the geometric object that is fixed by the respective symmetry group instead of the explicit symmetries themselves. Without loss of generality, each of these objects (regular n-gons in the case of C_n-symmetry, or platonic solids) can be inscribed in a unit sphere of the appropriate dimension ($S^1 \subset \mathbb{R}^2$ for n-gons via coordinates $(\cos(\frac{2\pi}{n}k), \sin(\frac{2\pi}{n}k))$ for $k = 0, ..., n-1$, or $S^2 \subset \mathbb{R}^3$ for platonic solids). For every class among these objects, one can choose a representative with vertices $V \subset S^N, N = 1, 2$. As we will show later, the C_n-case is independent of this choice. However, in the case of platonic solids, this is not the case; a choice can be fixed for the analysis pertaining to a given particle. Then, given a particle equipped with the corresponding type of symmetry, a choice of representative orientation $\omega \in SO(3)$ acts on V as $\omega V = \{\omega v \mid v \in V\}$.

Special care is required in the C_n case. Here, ω can be decomposed into a product of the in-plane portion and the cone-portion. The in-plane portion $\omega_{\text{in-plane}}$ can then be projected to the group of rotations of the plane, $SO(2) \cong S^1$, to act on $V \subset S^1$. Given any other rotation $\omega' \in SO(3)$ that represents the same orientation of the symmetric particle, we have $\omega V = \omega' V$, motivating the definition of a function to unambiguously distinguish between the orientations of two particles $(\omega_1, p_1), (\omega_2, p_2) \in SE(3)$ of the symmetry type of V:

$$\delta_V(\omega_1, \omega_2) := d_H(\omega_1 V, \omega_2 V), \tag{7}$$

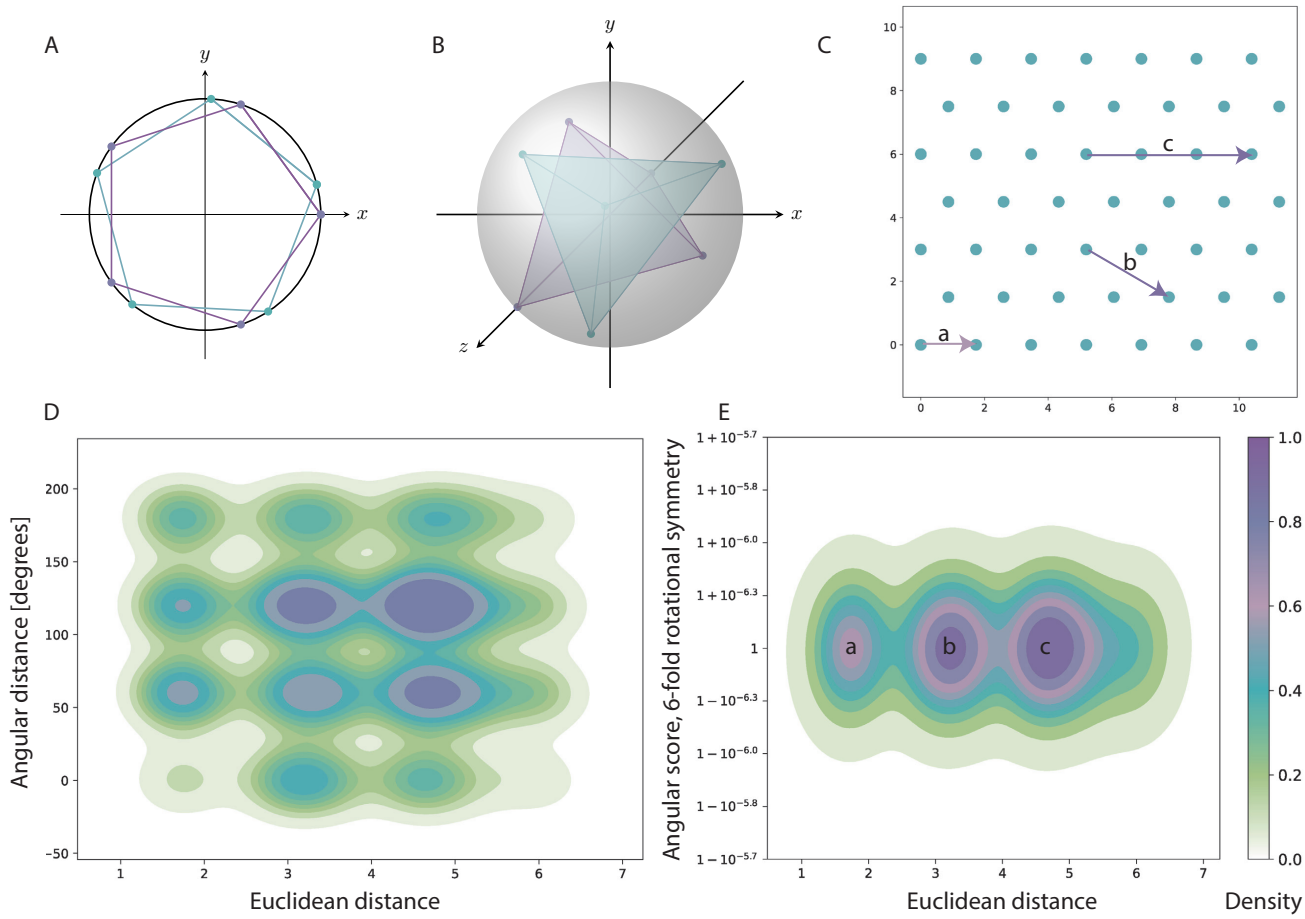

Figure 2: A, B. Depiction of symmetry types V together with rotated copies ωV for C_5-symmetry (**A**) and tetrahedral symmetry (**B**). **C**. An example for vertices in a regular, hexagonal lattice in the plane to demonstrate the unambiguous character of δ_V. Each vertex stands for the position of a C_6-symmetric particle; orientations are sampled from C_6. **D**. Kernel density estimation of options for angular distances measured around concentric shells of each particle from **C**. Peaks occur at multiples of 60°. **E**. By using σ_V instead of angular distance, the density peaks are concentrated around 1 (ideal hexagonal symmetry) at distance steps as indicated by the radii of concentric shells. A particle's first shell (a) is the smallest shell, followed in size by the second shell (b), and the third one (c). The labels a,b,c correspond to those in **C**. The depicted data does not account for boundary artifacts.

where d_H denotes the S^2-Hausdorff distance. The case for C_n-symmetry works analogously. In the case presented here, 7 simplifies to

$$\delta_V(\omega_1, \omega_2) = \max_{v \in V} d_{S^2}(\omega_1 v, \omega_2 V) \tag{8}$$

where $d_{S^2}(\cdot, \cdot)$ denotes the great-circle distance on S^2, and for a subset $W \subset S^2$ and a point $x \in S^2$, $d_{S^2}(x, W) = \inf_{w \in W} d_{S^2}(x, w)$. The same holds for the S^1-case for in-plane rotations.

Note that $\delta_V : SO(N+1) \times SO(N+1) \to \mathbb{R}$, $N = 1, 2$, does not define a metric, since positivity does not hold: the function is zero if $\omega_1 V = \omega_2 V$, which can occur for distinct rotations. However, δ_V induces a metric on the space that is obtained by identifying rotations whenever their relative orientation is a symmetry of V. Due to rotations being isometries on spheres, the S^N-Hausdorff-distance d_H remains unchanged under their actions, thus

$$\delta_V(\omega_1, \omega_2) = \delta_V(\omega\omega_1, \omega\omega_2) \tag{9}$$

for all $\omega \in SO(N+1)$.

If V were to represent a regular flat polygon, i.e. C_n-symmetry, commutativity of $SO(2)$ implies that δ_V is independent of the choice of regular n-gon inscribed in S^1:

$$\begin{aligned} \delta_{\omega V}(\omega_1, \omega_2) &= \delta_V(\omega_1\omega, \omega_2\omega) \\ &= \delta_V(\omega\omega_1, \omega\omega_2) \\ &= \delta_V(\omega_1, \omega_2) \end{aligned} \tag{10}$$

The situation for platonic solids can be remedied by supplying a fixed orientation to act on the given vertex set V in order for it to fit the user's data.

To obtain a normalized angular score, consider for each symmetry type V the maximum value of δ_V,

$$m_V := \max_{\omega_1, \omega_2 \in SO(N+1)} \delta_V(\omega_1, \omega_2). \tag{11}$$

We can thus define an angular score for particles of symmetry type V as

$$\sigma_V : SO(N+1) \times SO(N+1) \to [0,1],$$
$$\sigma_V(\omega_1, \omega_2) = 1 - \frac{\delta_V(\omega_1, \omega_2)}{m_V}. \quad (12)$$

2.4. Implementation

The presented framework is implemented in Python and embedded under the module named tango.py into the cryoCAT toolbox [29]. This module contains classes for particles, the twist descriptor and descriptor classes derived from it, features, supports, filters, and for the last three also their corresponding catalogs. Additional functionalities for analysis and visualization tailored to descriptor objects are provided as well.

The particle class contains necessary information for its processing, such as position, orientation, and symmetry as well as corresponding functionalities. Additionally, there is an option to add noise to a given object of that class. This can be orientational noise, positional noise, or both. In order to add orientational noise to a particle's associated rotation, a randomly generated 3×3 matrix is scaled with respect to some input noise level and added to the rotation matrix associated to the particle's orientation. Since the resulting matrix is generally not a rotation matrix, it is projected onto $SO(3)$ via its singular value decomposition [30]. In the case of positional noise, a vector is sampled from a normal distribution centered at the origin with standard deviation (STD) given by an input noise level. This vector is then added to the particle's position.

The twist descriptor class takes in either an already existing twist descriptor object or a particle list together with the radius for the initial support. Particle lists are read in using cryoCAT, in which support for the most common formats (TOM/AV3 [31, 32], Relion [33], STOPGAP [34], Dynamo [35], mod files [36]) is provided. It is assumed that particle lists do not contain overlapping particles and thus each entry in the data set represents a unique particle. This can be ensured as a pre-processing step using cryoCAT's distance-based or shape-based cleaning function [37]. The radius for the initial support can be derived from a statistical overview of nearest neighbor distances. For instance, the parameters that were used in the analyses demonstrated throughout this work depend on the medians and STDs of nearest neighbor distances. The twist vectors together with Euclidean distances, geodesic distances, and the product metric computed between the query particles and all particles within their respective spherical supports yields the "twist descriptor". Additionally, due to its practicality for many common use cases, the orientational deviation from the canonical rotational axes is included in the twist descriptor, as well as the in-plane angles of the query particle and each of its nearest neighbors.

For subsequent analysis, we provide a range of different supports such as spheres, cylinders, ellipsoids, cones, and tori. These are specified by coordinate axes which need not be restricted to the positional portions, but can also refer to orientational components. Additionally, thanks to the canonical orientation of the support, one can use custom binary masks of any shape (specified as a 3D array within python or loaded from an MRC or EM file). Implementation-wise, a support takes in a twist descriptor object and computes an intersection between itself and the twist positions, removing the particles outside the support.

Filters are based on similar principles as supports. They operate on descriptor objects and remove particles from the support based on required properties. For instance, one can set an allowed range for the angular distance from the query particle and filter out all particles from the support outside this range.

Features take in twist descriptor objects that have potentially already been reduced by the support or filtering if wanted, and compute the features belonging to each query particle. Unlike for the twist descriptor, where each query particle can have different numbers of neighboring particles within its support, the outcome of the feature has to have the same number of values for each query point. This ensures that the features can always be combined to form a new custom descriptor.

The custom descriptor class takes in a twist descriptor object, a list of desired features, and optionally a different support and list of filters to apply. If provided, the support is first used to reduce the twist descriptor object, followed by filtering, and finally the specified features are computed for each query point. Descriptors can be easily extended by new features (or reduced by removal of some of them) and they can also be merged with other descriptors that were computed on different supports. Besides the custom descriptor, additional descriptors are offered as well, see Supplementary section 1 for more details.

The computational cost for twist descriptors is influenced by the density of particle positions and the choice of search radii. While the initial computation can be expensive, further applications of features are based on simple, fast data frame operations. For user convenience, the twist descriptor can be saved into a file and loaded at a later time if needed.

Since TANGO is Python-based, it can be used within scripts, interactive Python environments, or Jupyter Notebooks. To make it more user-friendly, we implemented a GUI using the Plotly Dash library [38]. Dash allows for interactive data visualization in web browsers and thus greatly facilitates the descriptor analysis by allowing to easily explore different supports, filters, and features.

Finally, given the versatile character of cryo-ET data, we implemented TANGO in a way that users can easily extend it with their own features, supports, and filters. One can simply design e.g. a new feature by inheriting the Feature class and ensuring compliance with inputs and outputs as described above. Such a feature will automatically be added to the GUI as well, thanks to the feature catalog class that automatically collects all classes stemming from the parent Feature class. The same holds for supports and filters.

The full application programming interface (API) documentation, the tutorials, as well as an instructive video for the GUI are available as a part of the cryoCAT documentation.

3. Use Cases

3.1. Object Affiliation and Cleaning

Often, particle lists encode positions of structures that form larger complexes, such as positions of subunits of NPCs, or pleomorphic assemblies, such as $\alpha\beta$-tubulin dimers forming MTs, or subunits forming lattices on viral capsids (such as the immature HIV-1 CA-SP1 lattice). In the past, such particle lists typically came from manual particle picking of the underlying object. For instance, an NPC can be picked by specifying three points that define its position and orientation, aligned as a whole complex and then split into eight subunits or even into 24 subunits of the individual rings [39, 40]. The subunits of the immature HIV capsid can be picked by roughly finding the center of the underlying spherical capsid, measuring its radius, and using this information to create the spherical surface which is then used to create initial positions and orientations for STA. Similarly, the $\alpha\beta$-tubulin dimers are picked by tracing (either manually or automatically) the centers of the underlying MT and estimating their radii – the initial positions and orientations are then generated by oversampling the surface of the tube. These methods, while tedious, have two major advantages. First, the particle list is "clean", i.e., it contains only a limited number of junk particles that can easily be removed based on the object's geometry. The second advantage is the affiliation of each particle to its underlying object (e.g. to point a subunit to its associated complex). This is highly beneficial as it allows for the quantitative analysis of the affiliated objects [41, 42, 43] or for the automatic removal of misaligned subunits [44].

Recently, improvement in particle picking approaches such as 3D template matching (TM) [45, 46, 34] and deep-learning based methods [47, 48] started a shift from manual picking to automatic picking [5, 4, 34]. In the case of 3D-TM, particle lists are created from so-called score maps by specifying constrained cross-correlation (CCC) thresholds for each tomogram and picking positions above this threshold. In this step there is a trade off between having a clean particle list (reducing the number of false positive picks) and a complete particle list (minimizing the number of false negatives). The latter often results in more false positives in the data and requires further cleaning via classification, for example. The generated particle lists from TM also contain both positions and orientations, however they lack affiliation information associating individual particles to their respective assemblies. This prevents subsequent analyses of such objects. Here we demonstrate how TANGO can be used to tackle both of these issues.

3.1.1. Nuclear pore complexes

For the NPC structure, we worked with a data set from [4]. The input particle list was generated from GAPSTOP™ score maps coming from the original study. These score maps had been produced with a template containing a subunit of the cytoplasmic ring (CR). A manually curated particle list containing 1960 asymmetric subunits (SUs) from 322 NPCs was used as ground truth for the subsequent analysis. This list contained the affiliations of each SU to its respective NPC.

We analyzed a particle list from the original study [4] containing a total of 93,665 particles. Using cryoCAT's shape-based cleaning, this number was reduced to 63,187 particles, which were subsequently shifted in the intrinsic x-directions as indicated in Figure 3**A** by a distance corresponding to the NPC radius. This strategy was implemented as a pre-processing step to ensure that SUs of a given NPC cluster more closely, allowing for smaller spherical supports. The twist descriptor was computed on spherical supports with a radius of 15.5 nm. In the next step, a cylindrical support with a height of 9.7 nm, and a radius 15.5 nm was used to truncate the twist descriptor (Figure 3**A**). The choice of cylindrical supports was motivated by the spatial organization within the NPC and its 8-fold rotational symmetry (see [49] for more details on the NPC structure). Thus, neighboring SUs in the same NPC are expected to show a relative orientation that, up to noise, results from rotation around a particle's intrinsic z-axis (by $\sim 45°$), which can be considered a normal vector to the nuclear envelope. In the context of twist vectors, this is measured by ζ_z (5). The closer this value is to the overall angular distance, the closer is the relative orientation to a rotation around the z-axis. Only particle pairs for which the angular distance was within a tolerance of approximately $\pm 30°$ of $|\zeta_z|$ were kept. Additionally, the expected symmetry of the NPC was incorporated in the processing by only allowing angular distances that are around multiples of 45°. Since all of this information was already present in the twist descriptor, no extra computation was necessary. The data frame that resulted from this processing yielded a particle list made up of 2,341 members. For the performance of cleaning and classifying this data, consider the F_1-score [50]:

$$F_1 = \frac{2TP}{2TP + FP + FN}, \tag{13}$$

where TP is the number of true positives, FP refers to the number of false positives, and FN to the number of false negatives. Computing this score for each tomogram results in a median F_1-score of 0.967, pointing out a high precision in positive predictions. An example of the cleaning is shown in Figure 3**B**. Small examples for false positives, as well as false negatives can be observed. By visual inspection, two particles in this example have been correctly affiliated, despite not being part of the ground truth. The relevant particles are pointed out in Figure 3**B** (right).

3.1.2. Virus-like particles of immature HIV-1

In the case of virus-like particles (VLPs) of immature HIV-1, we focused on their CA-SP1 SUs from the EMPIAR-10277 data set. Unlike in the case of NPCs, this particle list was already clean in the sense that it did not contain any false

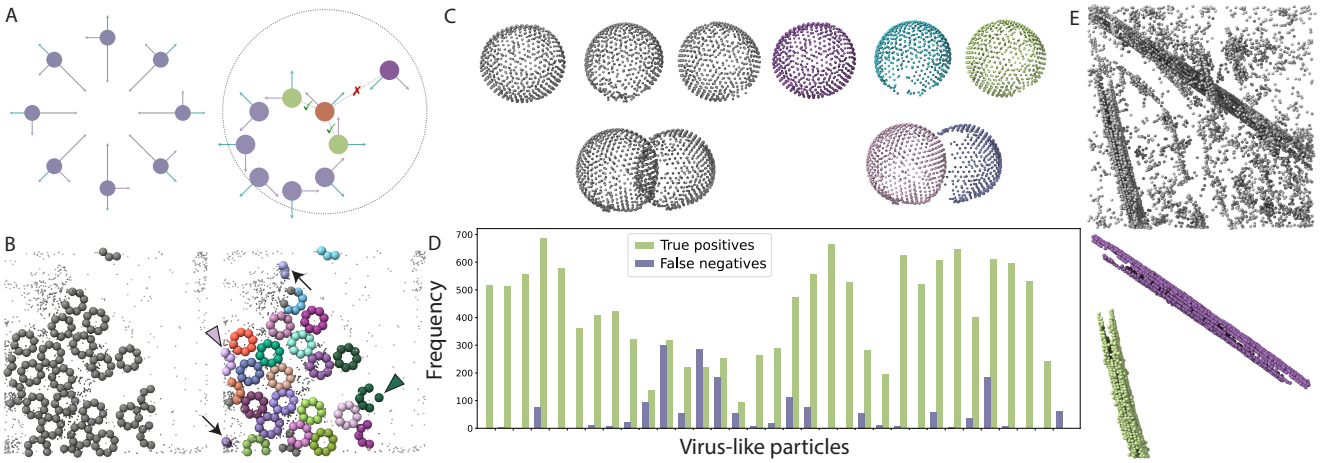

Figure 3: A. The ideal positions of the asymmetric units of the NPC, showing eight-fold symmetry, together with their orientation (left). The orientation is used to shift the particles along their intrinsic x-axes by the NPC radius (right). This reduces the initial spherical support for the twist descriptor that is further cropped to a cylindrical support (dashed line) followed by filtering out particles that deviate from the expected relative orientation. **B**. Particles from the uncleaned particle list placed back into a tomogram. Particles from the manually curated ground truth are represented with larger spheres (on the left). Results of cleaning and affiliation computation for NPCs are depicted on the right. Different colors are associated to different NPCs. Arrowheads point to the particles that were not in the ground truth but were correctly recovered by our method. Arrows point to wrongly spotted clusters. **C**. Particles representing CA-SP1 lattices of immature HIV-1 VLPs placed back into a tomogram (left). The assigned color-coded affiliations of subunits to their respective VLPs are shown on the right. **D**. The results of the affiliations for the complete HIV-1 dataset shown for each VLP separately. **E**. Particles from a particle list obtained from TM using a low CCC threshold to increase the completeness of the data (top). The result from clustering based on spatial proximity shows two distinct clusters corresponding to two MTs present in the data (bottom). The different colors correspond to the affiliation of each particle to its corresponding MT.

positives, yet the particles were lacking affiliation information. The latter was determined using the twist descriptor. The descriptor was computed on a support with radius 3.8 nm. The particles assemble on a spherical geometry, i.e., their intrinsic z-axes are perpendicular to the surface, pointing outwards. Thus $-z$ was chosen as the axis for a cylindrical support with height 0.25 nm and radius 3.8 nm. The twist descriptor was cropped accordingly. Since intrinsic z-normals can be considered surface normals, it follows that in cases where physically close particles have z-axes which deviate too much from each other, the particles in question belong to distinct VLPs. Naturally, these assumptions are within the context of appropriately chosen neighborhood radii and tolerance thresholds. This observation helps to distinguish between VLPs; we further cropped the twist descriptor by only keeping those particles within a support whose relative orientation was close to a rotation around z. Similar to the NPC case, the information necessary to compute object affiliations is already present in the twist descriptor. The nearest-neighbor information can be interpreted as a graph with nodes, where particles are nodes and two particles are connected via an edge if there is a twist vector associated to the pair, i.e., if one is a query particle and the other one lies within its support. In this sense, a twist descriptor decomposes into connected components and where, in the ideal case, one component corresponds to one VLP. In the case of sparsely populated VLPs, more than one component may be associated to the same VLP, and additional post-processing might be needed in such a case to correct for this. The results for the VLP data are shown in Figure 3. This affiliation reached a median F_1-score of 0.97 taken over all 32 VLPs in the data set. The lowest F_1-score of 0.48 was achieved for a VLP for which 139 true positives were found, versus 300 false negatives. This discrepancy was caused by data sparsity, as well as by considering a number of connected components that corresponded to the expected number of VLPs. The merging of connected components, as well as a larger support radius, could help improve individual cases, yet for this purpose, we decided for homogeneous, robust parameters that would benefit most VLPs.

3.1.3. Microtubules

A similar approach of computing affiliations based on proximity and particle orientation was applied to a tomogram used in the context of [46] containing two MTs. In this case, we obtained a score map from GAPSTOP™ using pieces of MTs as the template from the original study. We extracted a particle list of $\alpha\beta$-tubulin dimers using an absolute CCC-score threshold of 0.015 to ensure completeness of the data. In total, the particle list contained 6,997 particles. The twist descriptor was computed on spherical supports with a radius of 7.3 nm each. Given a particle (ω, p) from this data, the axis $p_x + p_y$ is approximately perpendicular to the MT. Thus, we again chose a cylindrical support around the intrinsic $x+y$-axes (with height 2.5 nm and radius 7.3 nm) to truncate the twist descriptor in a similar manner to the VLPs. The resulting data frame was decomposed into connected

components, the two biggest of which consisted of 1,174 and 974 particles respectively. Due to the lack of ground truth, the results were evaluated based on visual inspection, which confirmed that these particles lie close to the actual MTs, see Figure 3**E**.

3.2. Lattices

Pleomorphic assemblies such as protein filaments, coated vesicles, or some viral capsids are omnipresent in biological samples. They are built as non-covalent compounds out of numerous independently synthesized molecules serving as SUs of the final lattice covering the underlying geometry [51]. Analyzing regular, repeating patterns in such a structure is relatively straightforward. In contrast, spotting irregularities or deviations from this order can be more challenging, yet these are often the structures that are worth exploring more closely.

3.2.1. Mature HIV-1 capsids

Here, we show TANGO's ability to identify pentamers on the hexagonal lattice forming capsids of mature HIV-1 [4]. The mature HIV-1 capsid has a conical shape containing in total 12 pentamers located at regions of high curvature, hidden among hexameric SUs that form the rest of the lattice [52]. The input particle list contained positions of both hexamers (1,510 in total) and pentamers (57 in total) for 11 capsids. Note that all capsids were "incomplete", i.e. were missing a certain number of both hexameric as well as pentameric SUs.

To compute the twist descriptor for every capsid, the supports' radii were chosen to contain the neighboring hexamers (corresponding to the median nearest neighbor distance and 1.5·STD). On average, this resulted in a support radius of 10.8 nm with STD 0.37 nm. Since the majority of particles has C_6 symmetry, the angular score, tailored to compare the orientations of symmetric particles (σ_V, equation 12), was computed as well for the hexameric portion.

At first, a single lattice's regularity was investigated using the angular score. For this purpose, consider the concentric shells around any capsid protein, as visualized in Figure 4**B**. The average radii of the first three concentric shells are 9.8 nm, 18.2 nm, 25.8 nm. Over a given shell, this angular score was computed between the central particle and every particle in that current shell. Within that score, the value 1 hints at perfect orientational agreement, whereas 0 hints at maximum orientational displacement between particles of the same symmetry type, such as C_6 in the hexameric case. In the case of an unknown rotational symmetry type of the lattice particles, this descriptor can be applied to the data under the assumption of C_n symmetry for a range of values for n. As can be seen in 4**D**, the analyzed lattice shows the strongest tendency for C_6-symmetric particles. Knowing the symmetry type that is being handled, the evolution of the mean of the angular score for that specific type can be computed over shells of different radii. The results of this computation are shown in Figure 4**E**. The results were visualized using kernel density estimation, so as to point out peaks in the distribution. For the hexameric lattice, the highest degree of hexagonal regularity is observed in the first shell around each particle, decreasing over distance throughout the lattice, possibly due to the capsid's curvature.

Secondly, all eleven capsids were considered, for which the support was truncated to focus on the first shell only. For this support the α-complex descriptor was computed. That provided various statistics such as number of edges, the median and STD of the central angles (where the expected value is 60° for regular flat hexagons), Euler characteristics, and the link type. Principle component analysis (PCA) revealed that to explain more than 95% of variance, one needs to consider the median of central angles, their STD and the number of vertices. Using those three features we identified pentamers in 70.8% of cases. Figure 4**C** shows the results per capsid.

3.2.2. Immature HIV virus-like particles

A similar analysis was performed for the VLPs of immature HIV (same data as in section 3.1). The twist descriptor was computed, including the angular score, on spherical supports of radius $\sim$ 3.8 nm. This was followed by the analysis within separated shells. For a given VLP, the first shell has a radius of approximately 1.8 nm, followed by a shell with radius of $\sim$3.4 nm. There, the mean angular score of orientations within the first shell around each particle in every VLP shows that approximately 2/3 of the particles have immediate shells that show high hexagonal regularity (mean > 0.8). The information on angular similarity can be applied to subdivide the data into regions of different levels of angular agreement (see figure 4**G**), thus highlighting irregularities for further investigation. In this particular case, the particles showing the lowest degree of hexameric regularity were chosen for further investigation by STA. Indeed, already the initial average showed that those positions do not contain hexamers but can most likely be considered to be junk particles (see Supplementary Figure 2.

3.3. Pattern Recognition

The functional modules of the cell interact within its molecular sociology [53]. Among the challenges in understanding this cooperation is the study of interaction motifs. One approach to this can lie in spatial interactions as pronounced by relative orientations and positions. Within a given biological sample, the geometric interplay can form recurring patterns that can be uncovered using a descriptor based approach. On the one hand, this can be achieved in an exploratory manner, where dimensionality reduction and clustering methods can be applied to an array of locally computed features to scout for trends in the data. On the other hand, supports and features can be tailored to an initial assumption on what to find in the data.

3.3.1. Pattern recognition in synthetic chromatin data

To assess the ability of point descriptors to uncover patterns within data, we created synthetic data simulating an arrangement of nucleosomes within chromatin. Nucleosomes can be arranged in different structures and forms, for example, di- and trinucleosomes as described in [54]. Maps

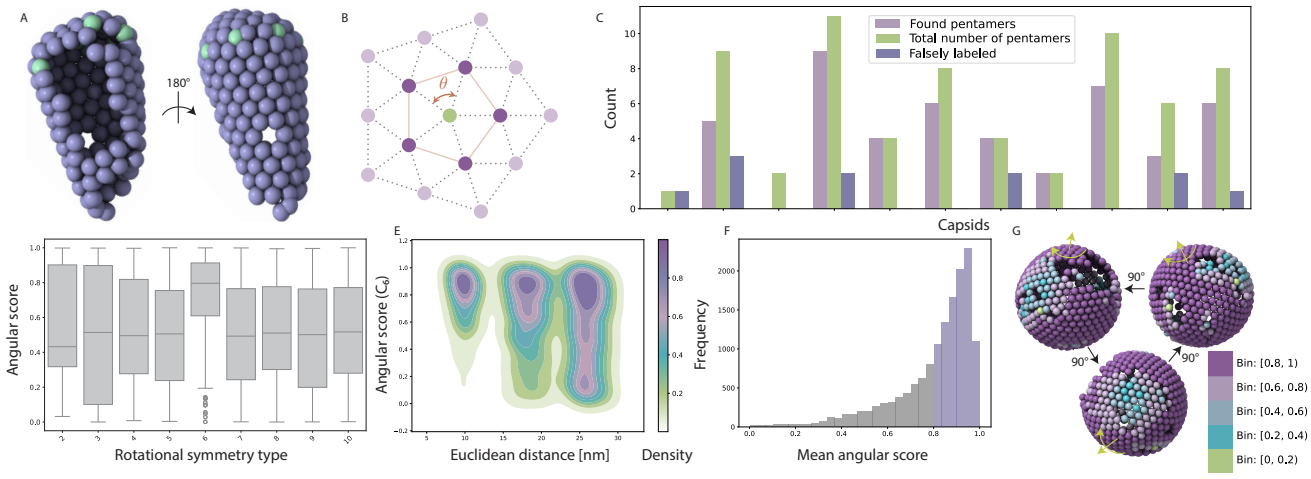

Figure 4: A. A representative capsid of the mature HIV-1 dataset, with pentameric and hexameric particles colored in two distinct colors. **B**. A two-dimensional depiction of a lattice structure. Given a query vertex in the lattice, the edges connecting its incident vertices form the link of that vertex. Alongside statistics describing the central angles θ, the link's topology and architecture (see equations 14 and 15 in the supplementary information) yield features to cluster with respect to the immediate shell of each particle. **C**. Results of labeling pentamers based on link topology- and geometry. The total number of pentamers refers to the annotated ground truth that was used. Among all the particles forming a given capsid, some were correctly identified as pentamers ("found pentamers"), others were labeled as such, despite being hexamers in the ground truth ("falsely labeled"). **D**. Symmetry analysis of the lattice shows a trend suggesting C_6-symmetry. Note that this result was achieved without imposing any preferred symmetry during the analysis. **E**. Analysis of the lattice regularity for C_6 symmetry based on angular similarity between the particles surrounding each query particle (i.e., all hexamers) in concentric lattice shells. The kernel density shows a trend towards high angular similarity ($\sigma_V \approx 1$) between C_6-symmetric particles, where regularity decreases over shells of larger radii, likely due to the capsid's curvature. **F**. The distribution of the mean angular score over the immediate shells for the hexamers of CA-SP1 forming lattices of immature HIV-1 VLPs. It shows a trend towards high angular similarity. **G**. A representative VLP of immature HIV-1 with its hexamers colored based on varying degrees of hexagonal regularity. Axes of rotation are indicated.

from EMDB corresponding to some of the aforementioned structures were used as bases for particle lists; helically arranged tetranucleosomes (EMD-2601), a pair of stacked nucleosomes (EMD-13365), and a trinucleosome (EMD-13363). For each structure, a list was generated by fitting the nucleosome structure (EMD-63079), the canonical orientation of which is indicated in Figure 5**A**, into the respective map in ChimeraX [55]. The resulting rigid motions were than translated into building blocks that were used for the synthetic data. A visualization of these building blocks is shown in Supplementary Figure 4. Copies of every such building block were rotated and translated at random (see equation 4), and combined into a single list, to which randomly generated particles representing mononucleosomes were added. To ensure that particles were not overlapping, cryoCAT's shape-based cleaning was applied to the synthetic data, using a binary mask that was created from the used nucleosome map. In the end, the list consisted of six helical arrangements (with 12 particles each), 350 stacked arrangements, and 340 trinucleosomes. As additional particles, a total of 3,200 were added in varying number densities, see Figure 5**B**. Afterwards, cleaning based on particle shape excluded one particle from the list.

This synthetic list was used as the unperturbed baseline for the analyses that follow. In order to investigate the robustness of the approach w.r.t. noise, the data was corrupted simultaneously with orientational noise and positional noise (see section 2.4 for implementation details). Five levels of positional noise were added, where the noise levels 1 through 5 refer to the STD of the normal distribution from which the noise was sampled. Noise level 1 results in a deviation of up to ∼0.8 nm, whereas a noise level of 5 can deviate the positions by up to 4 nm. Orientational noise levels were sampled up to 1, as indicated in Figure 5**D-F**.

An overview of angular distances occurring in the initial support (radius = 40.8 nm ≈ 4· median of nearest neighbor distances) suggests an investigation of short angular distances (Figure 5**C**). A high degree of angular similarity is to be expected in stacked nucleosomes for which linker DNA points in a similar direction. Therefore, the first pattern of interest to be detected was the one corresponding to stacked nucleosomes. Due to the canonical orientation of the model particle, the initial support was cropped to a cylindrical support around intrinsic y-axes (Figure 5**A**), in order to detect stacked particles. Once the twist descriptor was adapted, the angular distances occurring in the associated data frame were sorted into ten bins, the first one of which was used for the classification of particles. For the unperturbed case, this correctly classified 100% of the stacked particles. In total 83.3% of the particles from the helical arrangement and 66.8% of the particles belonging to trinucleosomes were classified as stacked. Note that 2/3 of every trinucleosome are in a stacked configuration, agreeing with our findings.

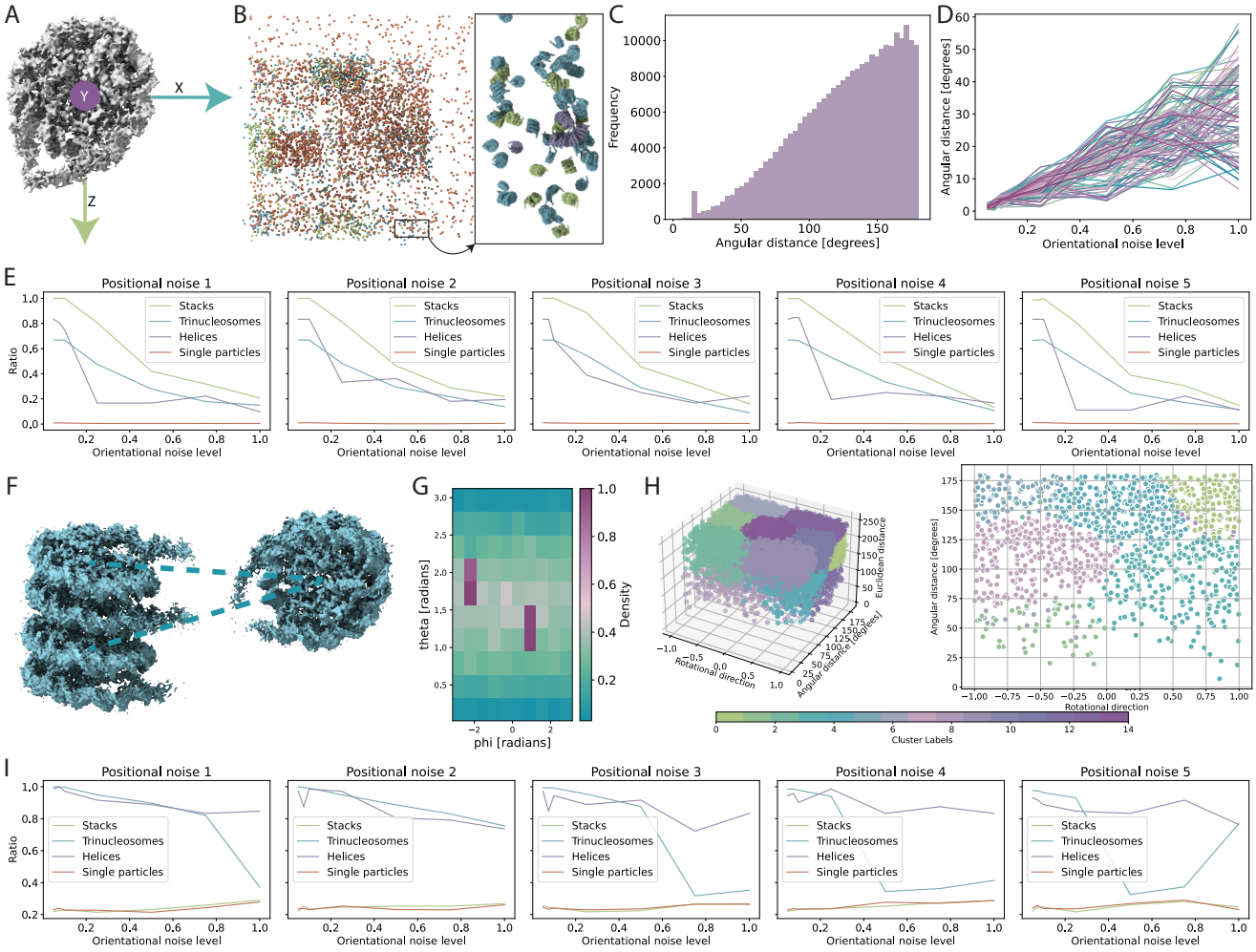

Figure 5: A. A nucleosome (EMD-63079) shown in its canonical orientation with indicated intrinsic coordinate frame. **B**. Overview of the synthetically generated particle list with varying local densities and inserted structures of stacked nucleosomes (EMD-13365), trinucleosomes (EMD-13363), helically arranged tetranucleosomes (EMD-2601), and mononucleosomes (EMD-63079) (left) and a close-up of the indicated region (right). Note that in the close-up, we removed the mononucleosomes to improve visualization. **C**. Histogram of angular distances between query particles and their neighbors within the initial spherical support showing a clear peak at the angular distance of 15°. **D**. The effect of different levels of orientational noise on angular distance. The graph presents a sample of 100 randomly generated rotation matrices. **E**. The performance of stack-classification with 5 levels of positional noise shown over different levels of orientational noise. **F**. A nucleosome model (EMD-63079) placed into the trinucleosome structure (EMD-13363). The dashed lines show dinucleosomes. **G**. A polar histogram with scores indicating relative occupation of bins by tangent directions in orientational space as computed from the twist vectors corresponding to cones extending in intrinsic z-directions. This analysis yielded the "rotational direction" feature. **H**. Overview of 15 clusters from k-means clustering applied to rotational direction, angular distance, and Euclidean distance (left). The right panel shows the section where Euclidean distance $\approx$ 24.5 nm. **I**. The performance of dinucleosome classification with 5 levels of positional noise shown over different levels of orientational noise.

1.03% of the additional particles were classified this way, some of which being part of stacked configurations.

Overall, the ratios concerning the structures of interest (stacks, trinucleosomes, helices) decreased when corrupting the list with increasing noise, while the ratio of classified additional particles remained low (Figure 5**E**). The highest fluctuation in quality can be observed for helix-particles, which might be attributed to their low number.

The other pattern of interest to be detected in this synthetic data set was that formed by dinucleosomes – pairs of nucleosomes that are connected via their linker DNA, such as can be detected in trinucleosomes (Figure 5**F**). For this purpose, the initial support was cropped to a conical one, with central axes corresponding to the intrinsic z-axes, since it is approximately where DNA leaves and enters the complex. To reduce the amount of possibilities to investigate potential connections between nucleosomes, common tangential directions in orientational space were probed. This meant considering the orientational portions ζ of the twist vectors from the conical support. When normalized, $\frac{1}{\|\zeta\|}\zeta$

lies in S^2, which was binned in order to select the most common orientational axis, $\hat{\zeta} \in S^2$, corresponding to the most common axis of (relative) rotation (Figure 5**G**). This finding motivated the definition of the descriptor used in this analysis: angular distance, Euclidean distance, and deviation from $\hat{\zeta}$ ("rotational direction", $\frac{\zeta \cdot \hat{\zeta}}{\|\zeta\|}$). These were computed on the conical support and subsequently clustered using the k-means algorithm (Figure 5**H**).

In the case of the unperturbed data, investigation of the outputs of the k-means algorithm uncovered a cluster containing both 100% of helix-labeled particles and all trinucleosome-labeled particles. The same cluster contained 22.7% of additional particles, as well as 21.7% of particles in stacks. The results of the data corrupted with noise are shown in Figure 5**I**, always depicting the cluster with the best ratio of correctly classified helix particles. It can be observed that, besides the falsely classified stacks and additional particles, the classification of helices is the most robust, while that of trinucleosomes decreases with noise levels above 4 and orientational noise levels above 0.75.

3.3.2. Pattern recognition in ribosome data

Ribosomes play a crucial role in protein synthesis. Improvements in the resolution that can be achieved using STA workflows has enabled the study of different ribosomal conformation within cells. Numerous studies investigated not only the different translational states but also their spatial distribution within cellular landscapes [1, 2, 56, 57]. Using our framework, we demonstrate how similar observations can be tackled with point descriptors.

The data we investigate here comes from the study by Xing et al. [2]. The study focused on eukaryotic ribosomes (80S) within human cells in their unperturbed condition as well as treated with an active site drug called Homoharringtonine. Correspondingly, the input for our analysis were two distinct particle lists encoding orientations and positions of ribosomes from the untreated and the treated cells. Similar to the original study, we shifted the central positions to correspond to the entry and exit points of mRNA. This resulted in two more particle lists per condition denoted as entry and exit lists.

The twist descriptor was computed based on the exit and entry particle lists. The exit list served as a pool for the query particles, while the neighbors were searched in the entry particle list. The initial search radius of 9 nm was chosen to remain consistent with the original study.

In the next step, we computed the SHOT descriptor (see Supplementary section 1 for more details). The spherical support was subdivided into six cones, each being divided into two concentric shells. In both conditions, four patterns – the SHOT footprints – occurred, representing common support occupancies relative to the query particle. These footprints were sorted by frequency, starting with the most common footprint. In both conditions, this was the cone revolving around the query particle's intrinsic y-axis. We focused on the three most common footprints for each condition.

In the untreated condition, the most common SHOT footprint was observed for 4968 query particles out of 38928 total particles. In the treated condition, it was 1554 query particles out of 38641. The orientational portions of the twist vectors that correspond to these footprints were then analyzed to investigate common orientational motifs in relation to the positional arrangements. To this end, orientational portions ζ were normalized and sorted into a 2D polar histogram. In order to focus on short entry-to-exit distances for mRNA, the subsequent analysis was restricted to pairs of ribosomes with an entry-to-exit distance below 5 nm. This cutoff was chosen after a review of the common distances per footprint.

The data corresponding to a given polar bin was then partitioned into five bins with respect to their angular distances. For all three SHOT footprints that are being discussed here, the fifth bin was always empty within the physical distance restriction that was imposed (Figure 6).

Note that for the first and second SHOT footprints, the polar bins of highest density aligned between the experimental conditions, whereas within the third footprint, the most common axes of rotations differed, as depicted in Figure 6**C**,**F**. 44 pairs fall into the spherical bin marked in **C**, versus 26 that fall into the bin that was otherwise the one with the highest density in all the other cases. In the treated condition (**F**), the marked bin contained 65 pairs, while the bin that corresponds to the marked bin in **C** only contained 3 pairs.

In comparison with Xing et al. [2], three configurations of pairs of ribosomes within polysomes were described: "top-top" (t-t), "top-down" (t-d), and "top-up" (t-u). Representatives of all three classes could be found using the exploratory approach described above in the untreated condition, as well as classes that were not conclusively categorized.

Interestingly, the mutual orientation among the pairs belonging to the bin highlighted in Figure 6 **F**, remarkably resembles the stalling disomes described in [56].

Representatives of all classes as detected using our framework are shown in Figure 6.

To validate the findings, STA was performed for all the cases as depicted in Figure 6. More precisely, we used the alignment from the original study to extract and directly average the pairs. Due to a low number of particles, the extraction was done on four times binned tomograms (with pixel size 2.446 Å) and no further alignment was performed. A selection of the outputs is shown in Supplementary Figure 5. In the untreated case, strong densities surrounding an averaged ribosome pair hint at the found pairs being parts of polysomes, whereas in the treated condition, the averaged maps hint at secluded pairs. These findings align with the results presented by Xing et. al [2]. The STA of the pairs from the most occupied bin of third footprint of the treated condition (Figure 6 **F**) was overlayed with the map of the stalling disome (EMDB-10398) (Supplementary Figure 6). While the overall mutual orientation of the pairs aligns well with the disome map, the resolution of the average is not sufficient to conclusively confirm that the pairs form

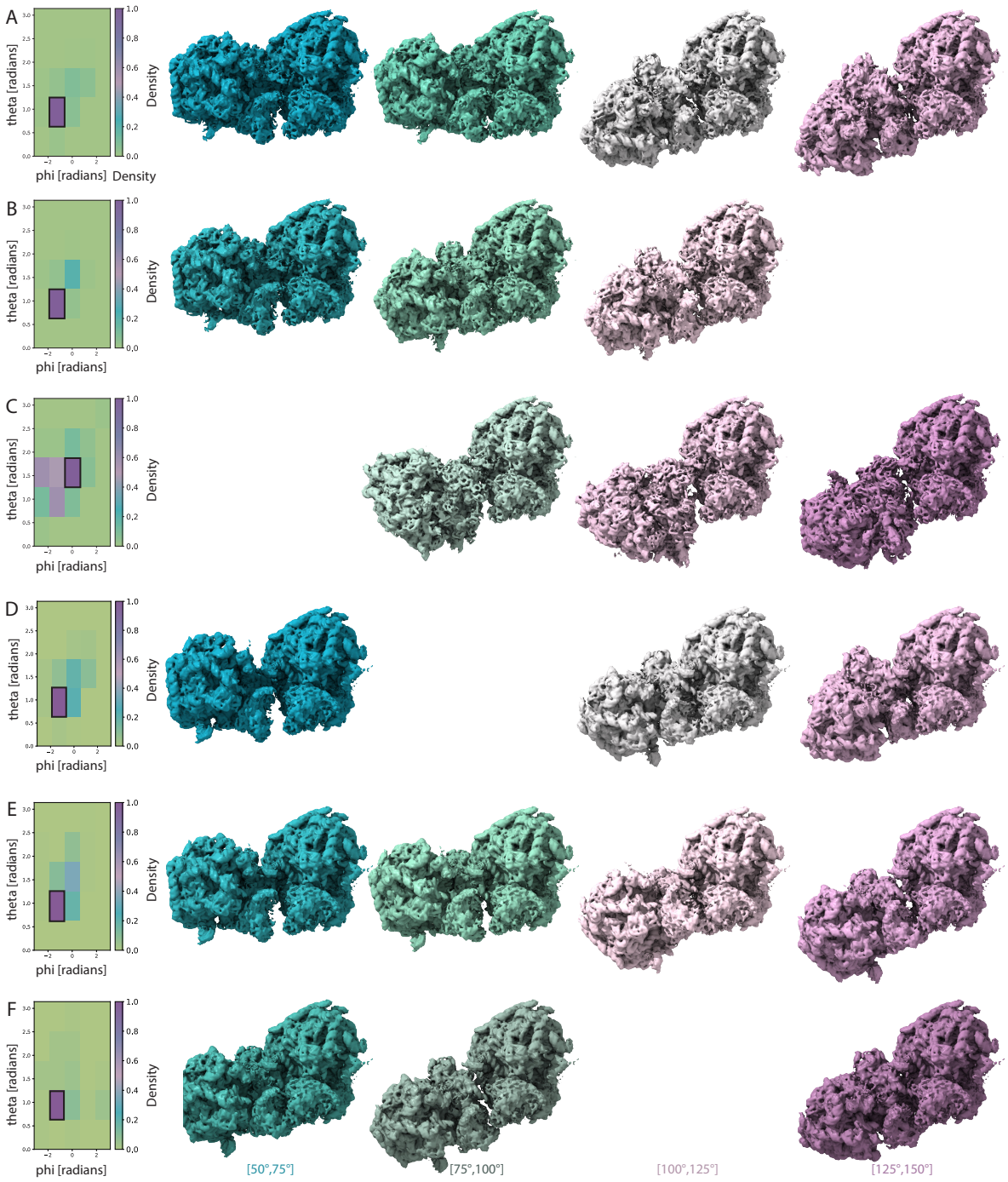

Figure 6: A-C. Untreated condition. Each row depicts another SHOT footprint, the most common one being in **A** and decreasing in frequency. The left columns show where the most common tangential directions in orientational space are within a polar histogram. The data from these polar bins was further separated with respect to angular distances as indicated on top. Representative pairs from each of these angular distance bins are shown in the columns to the right. The query particle of each pair is on the right. **D-F**. The same as in **A-C** but for the treated condition.

stalling ribosomes. Nevertheless, it shows the potential of our method to find rare occurrences within the data.

4. Limitations

While powerful, the proposed approach has its limitations. The most notable one in the context of cryo-ET concerns the possible incompleteness of the data. Due to many artifacts present in tomograms and imperfections in both particle picking approaches and STA, input particle lists rarely contain all particles that are present in tomograms. While this issue does not affect statistical analysis, finding patterns or reliably assigning affiliations might be negatively

affected. For instance, finding longer patterns of nucleosomes, such as the presented helically arranged tetranucleosomes, would fail in the case of missing particles and might skew the overall picture of the chromatin arrangement by suggesting the existence of only shorter patterns. One way of dealing with such scenarios is to implement a hierarchical approach in which one would first find the shorter motifs and in the second step tried to connect them into longer ones based on their proximity and orientation.

Another challenge comes in the form of noise caused either by particle misalignment or by inherent "imperfections" of cellular components – very rarely does it occur that the reality corresponds to geometrical expectations. For instance, the SUs of NPCs form a regular octagon and thus the central angle should be 45°. However, even for well aligned SUs, the central angle varies from this ideal value, making it difficult to distinguish between the physiology and the misalignment coming from data processing. One thus has to be careful when discarding data based on their geometry. It is advised to leave room for some deviations from the expected values.

For the point clouds in their original meaning, there are many descriptors solving both these issues by statistical means (for instance using histograms on the support, instead of direct values). Unfortunately, such approaches are less suited for cryo-ET data due to its sparsity. We propose to tackle noise-related issues by a suitable setup of filtering parameters and offer the GUI to make this optimization more straightforward. As for the data incompleteness, we currently do not offer any solution, however it should be possible to use the framework to actually identify more particles within the primary data and thereby reduce incompleteness. For instance, to search for trinucleosomes, one can identify footprints where one of the three nucleosomes is missing and search for the corresponding CCC peaks in TM score maps or use those positions and expected orientations as a starting point for STA.

Another possible future direction that could make this framework more robust to noise and incompleteness is the application of neural networks, as they have already proved themselves very useful in the classic point cloud descriptors field [58]. In cryo-ET, the use of deep learning approaches is still very limited in comparison to other areas. The main reason for that is the lack of annotated experimental data that is required for the reliable training of models. While it is difficult to generate data that would truthfully mimic noise properties in the case of image processing tasks, it should not be a problem to generate enough training data in the case of particle lists, as demonstrated in the synthetic nucleosome case.

5. Conclusions

The rich nature of particle data from cryo-ET comes in the form of positional and orientational information. Here, we applied a framework based on point cloud descriptors to analyze the geometric interplay between particles. The core of our framework lies within the twist descriptor, which provides an elegant way to describe these mutual relationships and offer a versatile basis for further analysis.

We showed that the twist descriptor itself is sufficient for geometrical cleaning of particle lists that contain either misaligned or falsely picked particles (the latter typically caused by automated methods for particle picking). Using our approach, we successfully removed false positives from lists stemming from TM of $\alpha\beta$-tubulin dimers as well as of the asymmetric unit of the NPC's cytoplasmic ring. In the latter case, our approach even recovered particles that were falsely removed during manual cleaning. Furthermore, focusing on orientational aspects yields built-in means of separating connected components, which allowed us to affiliate the particles with their underlying objects. Naturally, this approach to automated cleaning and affiliation computation is only suited for particles that assemble in a known geometry, such as large macromolecular complexes with known symmetry, or pleomorphic assemblies.

We further demonstrated the capabilities of features built on the basic twist descriptor to analyze highly structured particle lists – ones whose particles assembled in lattices. Different degrees of local incompleteness of these lattices can be investigated using descriptors of particle neighborhoods, as well as different lattice symmetries. We showed this on incomplete capsids of mature HIV-1, where a topological descriptor suited for regular lattices was used to identify pentamers within hexagonal arrangements, as well as local incompleteness within neighborhoods. A descriptor designed to account for orientational ambiguity among symmetric particles was used to identify irregularities within CA-SP1 lattices of immature HIV-1 VLPs.

Finally, we showed the potential of our method for the analysis of soluble particles, such as ribosomes. By applying an exploratory approach to ribosome data that can be partitioned into two experimental conditions, different common characteristics between those cases can be identified and statistically recorded. Such findings can be reinforced using STA to obtain structures from patterns or for fast classification. We thus demonstrated the potential of our framework in the context of geometric characterization, which may facilitate the comparison of experimental data exposed to differing conditions, such as wild types and mutants.

Although diverse in their character, all presented analyses were done by following the suggested workflow.

A key aspect of the proposed concept is its implementation, which delivers a tool designed to be both user-friendly and functionally tailored to the needs of the cryo-ET community. TANGO is a publicly available, fully open-source, Python-based tool designed to be accessible to experienced users as well as newcomers to the field, including those with limited programming skills, thereby supporting a wide range of levels of expertise. On the one hand, it is enhanced by a user-friendly GUI, which enables analyses without the need to write any code. This includes building custom descriptors from the features provided within the feature catalog. On the other hand, its modular design allows

more experienced users to extend the provided functionality by their own features, supports, or even entire descriptors. Overall, we showed that TANGO is a versatile framework for the analysis of particles stemming from cryo-ET and that it has the potential to become an essential tool for elucidating the complex molecular choreography within cells.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could appear to have influenced the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used chatGPT in order to improve phrasing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Code availability

The framework TANGO is part of the cryoCAT (Contextual Analysis Tool for cryo-electron tomography) public GitHub repository [29], specifically at `https://github.com/turonova/cryoCAT/blob/main/cryocat/tango.py`. Tutorials in the form of commented Jupyter notebooks are available under `https://github.com/turonova/cryoCAT/tree/main/docs/source/tutorials/tango_tutorials/`, guiding the user through two analysis cases. The required data is provided there as well.

Acknowledgements

We would like to thank Jan Philipp Kreysing for providing particle lists, score maps, and tomograms on mature HIV and NPCs as well as for helpful discussions. Huaipeng Xing for providing data for the analysis and visualization of ribosomes. Eunyoung Jeong and Jiasui Liu for providing data for further testing of our approaches. Desislava Glushkova, Giulia Tonon, Jan Philipp Kreysing, Maarten Tuijtel, Martin Beck, and Patrick Hoffmann for critically reading the manuscript. Finally, we would also like to thank all members of the Beck department for sharing with us their struggle with data analysis that ultimately lead to this study.

Funding: *The authors received funding from the Max Planck Institute of Biophysics, Frankfurt am Main, Germany.*

CRediT authorship contribution statement

Markus Schreiber: Conceptualization, Formal analysis, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Beata Turoňová:** Conceptualization, Formal analysis, Methodology, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing.

References

[1] Liang Xue, Swantje Lenz, Maria Zimmermann-Kogadeeva, Dimitry Tegunov, Patrick Cramer, Peer Bork, Juri Rappsilber, and Julia Mahamid. Visualizing translation dynamics at atomic detail inside a bacterial cell. *Nature*, 610(7930):205–211, 2022.

[2] Huaipeng Xing, Reiya Taniguchi, Iskander Khusainov, Jan Philipp Kreysing, Sonja Welsch, Beata Turoňová, and Martin Beck. Translation dynamics in human cells visualized at high resolution reveal cancer drug action. *Science*, 381(6653):70–75, 2023.

[3] Max Gemmer, Marten L Chaillet, Joyce van Loenhout, Rodrigo Cuevas Arenas, Dimitrios Vismpas, Mariska Gröllers-Mulderij, Fujiet A Koh, Pascal Albanese, Richard A Scheltema, Stuart C Howes, et al. Visualization of translation and protein biogenesis at the er membrane. *Nature*, 614(7946):160–167, 2023.

[4] Jan Philipp Kreysing, Maziar Heidari, Vojtech Zila, Sergio Cruz-Leon, Agnieszka Obarska-Kosinska, Vibor Laketa, Lara Rohleder, Sonja Welsch, Jürgen Köfinger, Beata Turoňová, et al. Passage of the hiv capsid cracks the nuclear pore. *Cell*, 188(4):930–943, 2025.

[5] Reiya Taniguchi, Clarisse Orniacki, Jan Philipp Kreysing, Vojtech Zila, Christian E Zimmerli, Stefanie Böhm, Beata Turoňová, Hans-Georg Kräusslich, Valérie Doye, and Martin Beck. Nuclear pores safeguard the integrity of the nuclear envelope. *Nature Cell Biology*, pages 1–14, 2025.

[6] Antonio Martinez-Sanchez, Wolfgang Baumeister, and Vladan Lučić. Statistical spatial analysis for cryo-electron tomography. *Computer methods and programs in biomedicine*, 218:106693, 2022.

[7] Wenhong Jiang, Jonathan Wagner, Wenjing Du, Juergen Plitzko, Wolfgang Baumeister, Florian Beck, and Qiang Guo. A transformation clustering algorithm and its application in polyribosomes structural profiling. *Nucleic Acids Research*, 50(16):9001–9011, 2022.

[8] Xian-Feng Han, Zhi-Ao Feng, Shi-Jie Sun, and Guo-Qiang Xiao. 3d point cloud descriptors: state-of-the-art. *Artificial Intelligence Review*, 56(10):12033–12083, 2023.

[9] George Vosselman, Ben GH Gorte, George Sithole, and Tahir Rabbani. Recognising structure in laser scanner point clouds. *International archives of photogrammetry, remote sensing and spatial information sciences*, 46(8):33–38, 2004.

[10] R Blomley, M Weinmann, J Leitloff, and B Jutzi. Shape distribution features for point cloud analysis–a geometric histogram approach on multiple scales. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2:9–16, 2014.

[11] Herbert Edelsbrunner, Michael Facello, and Jie Liang. On the definition and the construction of pockets in macromolecules. *Discrete Applied Mathematics*, 88(1-3):83–102, 1998.

[12] Zixuan Cang and Guo-Wei Wei. Topological fingerprints reveal protein-ligand binding mechanism, 2017.

[13] Michael Bleher, Lukas Hahn, Juan Angel Patino-Galindo, Mathieu Carriere, Ulrich Bauer, Raul Rabadan, and Andreas Ott. Topology identifies emerging adaptive mutations in sars-cov-2. preprint. *arXiv preprint arXiv:2106.07292*, 2021.

[14] Aaron S Keys, Christopher R Iacovella, and Sharon C Glotzer. Harmonic order parameters for characterizing complex particle morphologies. *arXiv preprint arXiv:1012.4527*, 2010.

[15] Torsten Fiolka, Jörg Stückler, Dominik A Klein, Dirk Schulz, and Sven Behnke. Sure: Surface entropy for distinctive 3d features. In *Spatial Cognition VIII: International Conference, Spatial Cognition 2012, Kloster Seeon, Germany, August 31–September 3, 2012. Proceedings 8*, pages 74–93. Springer, 2012.

[16] Joao Paulo Silva do Monte Lima and Veronica Teichrieb. An efficient global point cloud descriptor for object recognition and pose estimation. In *2016 29th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 56–63. IEEE, 2016.

[17] Aaron S Keys, Christopher R Iacovella, and Sharon C Glotzer. Characterizing complex particle morphologies through shape matching: Descriptors, applications, and algorithms. *Journal of Computational Physics*, 230(17):6438–6463, 2011.

[18] Jing Huang and Suya You. Point cloud matching based on 3d self-similarity. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 41–48. IEEE, 2012.

[19] Keke Tang, Peng Song, and Xiaoping Chen. Signature of geometric centroids for 3d local shape description and partial shape matching. In *Asian Conference on Computer Vision*, pages 311–326. Springer, 2016.

[20] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11*, pages 356–369. Springer, 2010.

[21] Michael Habeck, Andreas Kröpelin, and Nima Vakili. Matching biomolecular structures by registration of point clouds, 2024.

[22] Richard Murray, Zexiang Li, and Shankar Sastry. *A Mathematical Introduction to Robot Manipulation*. CRC Press, 1994.

[23] Roshan Kumar Hota and Cheruvu Siva Kumar. Derivation of the rotation matrix for an axis-angle rotation based on an intuitive interpretation of the rotation matrix. In Rajeev Kumar, Vishal S. Chauhan, Mohammad Talha, and Himanshu Pathak, editors, *Machines, Mechanism and Robotics*, pages 939–945, Singapore, 2022. Springer Singapore.

[24] Oene Bottema and Bernard Roth. *Theoretical kinematics*, volume 24. Courier Corporation, 1990.

[25] J.M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, 2003.

[26] José Luis Blanco-Claraco. A tutorial on **SE**(3) transformation parameterizations and on-manifold optimization, 2022.

[27] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *International journal of computer vision*, 103:267–305, 2013.

[28] Herbert Edelsbrunner. *A short course in computational geometry and topology*. Number Mathematical methods. Springer, 2014.

[29] Beata Turoňová. cryoCAT: Contextual Analysis Tools for cryo electron tomography, March 2025.

[30] Gilbert W Stewart. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993.

[31] Reiner Hegerl. The em program package: a platform for image processing in biological electron microscopy. *Journal of structural biology*, 116(1):30–34, 1996.

[32] Stephan Nickell, Friedrich Förster, Alexandros Linaroudis, William Del Net, Florian Beck, Reiner Hegerl, Wolfgang Baumeister, and Jürgen M Plitzko. Tom software toolbox: acquisition and analysis for electron tomography. *Journal of structural biology*, 149(3):227–234, 2005.

[33] Tanmay AM Bharat and Sjors HW Scheres. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in relion. *Nature protocols*, 11(11):2054–2065, 2016.

[34] William Wan, Sagar Khavnekar, and Jonathan Wagner. Stopgap: an open-source package for template matching, subtomogram alignment and classification. *Biological Crystallography*, 80(5):336–349, 2024.

[35] Daniel Castaño-Díez, Mikhail Kudryashev, Marcel Arheit, and Henning Stahlberg. Dynamo: a flexible, user-friendly development tool for subtomogram averaging of cryo-em data in high-performance computing environments. *Journal of structural biology*, 178(2):139–151, 2012.

[36] James R Kremer, David N Mastronarde, and J Richard McIntosh. Computer visualization of three-dimensional image data using imod. *Journal of structural biology*, 116(1):71–76, 1996.

[37] Jan Philipp Kreysing, Sergio Cruz-León, Johannes Betz, Carlotta Penzo, Tomáš Majtner, Markus Schreiber, Beata Turoňová, Marina Lusic, Gerhard Hummer, and Martin Beck. Molecular architecture of heterochromatin at the nuclear periphery of primary human cells. *bioRxiv*, 2025.

[38] Chris Parmer, Philippe Duval, and Alex Johnson. A data and analytics web app framework for Python, no JavaScript required., November 2024.

[39] Khanh Huy Bui, Alexander Von Appen, Amanda L DiGuilio, Alessandro Ori, Lenore Sparks, Marie-Therese Mackmull, Thomas Bock, Wim Hagen, Amparo Andrés-Pons, Joseph S Glavy, et al. Integrated structural analysis of the human nuclear pore complex scaffold. *Cell*, 155(6):1233–1243, 2013.

[40] Shyamal Mosalaganti, Agnieszka Obarska-Kosinska, Marc Siggel, Reiya Taniguchi, Beata Turoňová, Christian E Zimmerli, Katarzyna Buczak, Florian H Schmidt, Erica Margiotta, Marie-Therese Mackmull, et al. Ai-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science*, 376(6598):eabm9506, 2022.

[41] William Wan, Larissa Kolesnikova, Mairi Clarke, Alexander Koehler, Takeshi Noda, Stephan Becker, and John AG Briggs. Structure and assembly of the ebola virus nucleocapsid. *Nature*, 551(7680):394–397, 2017.

[42] Beata Turoňová, Mateusz Sikora, Christoph Schürmann, Wim JH Hagen, Sonja Welsch, Florian EC Blanc, Sören von Bülow, Michael Gecht, Katrin Bagola, Cindy Hörner, et al. In situ structural analysis of sars-cov-2 spike reveals flexibility mediated by three hinges. *Science*, 370(6513):203–208, 2020.

[43] Christian E Zimmerli, Matteo Allegretti, Vasileios Rantos, Sara K Goetz, Agnieszka Obarska-Kosinska, Ievgeniia Zagoriy, Aliaksandr Halavatyi, Gerhard Hummer, Julia Mahamid, Jan Kosinski, et al. Nuclear pores dilate and constrict in cellulo. *Science*, 374(6573):eabd9776, 2021.

[44] Beata Turoňová, Wim JH Hagen, Martin Obr, Shyamal Mosalaganti, J Wouter Beugelink, Christian E Zimmerli, Hans-Georg Kräusslich, and Martin Beck. Benchmarking tomographic acquisition schemes for high-resolution structural biology. *Nature Communications*, 11(1):876, 2020.

[45] Valentin J Maurer, Marc Siggel, and Jan Kosinski. Pytme (python template matching engine): A fast, flexible, and multi-purpose template matching library for cryogenic electron microscopy data. *SoftwareX*, 25:101636, 2024.

[46] Sergio Cruz-León, Tomáš Majtner, Patrick Hoffmann, Jan P Kreysing, Maarten W Tuijtel, Stefan L Schaefer, Katharina Geißler, Martin Beck, Beata Turoňová, and Gerhard Hummer. High-confidence 3d template matching for cryo-electron tomography. *Biophysical Journal*, 123(3):183a, 2024.

[47] Irene de Teresa-Trueba, Sara K Goetz, Alexander Mattausch, Frosina Stojanovska, Christian E Zimmerli, Mauricio Toro-Nahuelpan, Dorothy WC Cheng, Fergus Tollervey, Constantin Pape, Martin Beck, et al. Convolutional networks for supervised mining of molecular patterns within cellular context. *Nature Methods*, 20(2):284–294, 2023.

[48] Guole Liu, Tongxin Niu, Mengxuan Qiu, Yun Zhu, Fei Sun, and Ge Yang. Deepetpicker: Fast and accurate 3d particle picking for cryo-electron tomography using weakly supervised deep learning. *Nature Communications*, 15(1):2090, 2024.

[49] Martin Beck and Ed Hurt. The nuclear pore complex: understanding its function through structural insight. *Nature reviews Molecular cell biology*, 18(2):73–89, 2017.

[50] Yutaka Sasaki et al. The truth of the f-measure. *Teach tutor mater*, 1(5):1–5, 2007.

[51] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molekularbiologie der Zelle: 5. Auflage*. WILEY-VCH, 2011.

[52] Simone Mattei, Bärbel Glass, Wim JH Hagen, Hans-Georg Kräusslich, and John AG Briggs. The structure and flexibility of conical hiv-1 capsids determined within intact virions. *Science*, 354(6318):1434–1437, 2016.

[53] Carol V Robinson, Andrej Sali, and Wolfgang Baumeister. The molecular sociology of the cell. *Nature*, 450(7172):973–982, 2007.

[54] Shujun Cai, Désirée Böck, Martin Pilhofer, and Lu Gan. The in situ structures of mono-, di-, and trinucleosomes in human heterochromatin. *Molecular biology of the cell*, 29(20):2450–2457, 2018.

[55] Elaine C Meng, Thomas D Goddard, Eric F Pettersen, Greg S Couch, Zach J Pearson, John H Morris, and Thomas E Ferrin. Ucsf

chimerax: Tools for structure building and analysis. *Protein Science*, 32(11):e4792, 2023.

[56] Juliette Fedry, Joana Silva, Mihajlo Vanevic, Stanley Fronik, Yves Mechulam, Emmanuelle Schmitt, Amédée Des Georges, William James Faller, and Friedrich Förster. Visualization of translation reorganization upon persistent ribosome collision stress in mammalian cells. *Molecular cell*, 84(6):1078–1089, 2024.

[57] Iskander Khusainov, Natalie Romanov, Camille Goemans, Beata Turoňová, Christian E Zimmerli, Sonja Welsch, Julian D Langer, Athanasios Typas, and Martin Beck. Bactericidal effect of tetracycline in e. coli strain ed1a may be associated with ribosome dysfunction. *Nature Communications*, 15(1):4783, 2024.

[58] Lei Li, Siyu Zhu, Hongbo Fu, Ping Tan, and Chiew-Lan Tai. End-to-end learning local multi-view descriptors for 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1919–1928, 2020.

[59] Colin P Rourke and Brian Joseph Sanderson. *Introduction to piecewise-linear topology*. Springer Science & Business Media, 2012.

[60] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on information theory*, 29(4):551–559, 1983.

[61] Herbert Edelsbrunner. Smooth surfaces for multi-scale shape representation. In *International conference on foundations of software technology and theoretical computer science*, pages 391–412. Springer, 1995.

[62] Boris Delaunay. Sur la sphère vide. a la mémoire de georges voronoï. *Izvestia Akademia Nauk SSSR, VII Seria, Otdelenie Matematicheskii i Estestvennyka Nauk*, 7:793–800, 1934.

[63] Norman Levitt. The euler characteristic is the unique locally determined numerical homotopy invariant of finite complexes. *Discrete & computational geometry*, 7:59–67, 1992.

[64] turonova. turonova/novasta: Advanced particle analysis, January 2022.