UCLA

UCLA Previously Published Works

Title

Assessment of scoring functions to rank the quality of 3D subtomogram clusters from cryo-electron tomography

Permalink

https://escholarship.org/uc/item/6ch871wq

Journal

Journal of Structural Biology, 213(2)

ISSN

1047-8477

Authors

Singla, Jitin
White, Kate L
Stevens, Raymond C
et al.

Publication Date

2021-06-01

DOI

10.1016/j.jsb.2021.107727

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Struct Biol. Author manuscript; available in PMC 2022 June 01.

Published in final edited form as:

J Struct Biol. 2021 June ; 213(2): 107727. doi:10.1016/j.jsb.2021.107727.

Assessment of scoring functions to rank the quality of 3D subtomogram clusters from cryo-electron tomography

Jitin Singla[1,2,3], **Kate L. White**[3], **Raymond C. Stevens**[3], **Frank Alber**[1,2,*]

[1]Institute for Quantitative and Computational Biosciences, Department of Microbiology, Immunology, and Molecular Genetics, University of California Los Angeles, 520 Boyer Hall, Los Angeles, CA 90095.

[2]Quantitative and Computational Biology, Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA.

[3]Department of Biological Sciences, Bridge Institute, Michelson Center for Convergent Bioscience, University of Southern California, Los Angeles, CA 90089, USA

Abstract

Cryo-electron tomography provides the opportunity for unsupervised discovery of endogenous complexes in situ. This process usually requires particle picking, clustering and alignment of subtomograms to produce an average structure of the complex. When applied to heterogeneous samples, template-free clustering and alignment of subtomograms can potentially lead to the discovery of structures for unknown endogenous complexes. However, such methods require scoring functions to measure and accurately rank the quality of aligned subtomogram clusters, which can be compromised by contaminations from misclassified complexes and alignment errors. Here, we provide the first study to assess the effectiveness of more than 15 scoring functions for evaluating the quality of subtomogram clusters, which differ in the amount of structural misalignments and contaminations due to misclassified complexes. We assessed both experimental and simulated subtomograms as ground truth data sets. Our analysis showed that the robustness of scoring functions varies largely. Most scores were sensitive to the signal-to-noise ratio of subtomograms and often required Gaussian filtering as preprocessing for improved performance. Two scoring functions, Spectral SNR-based Fourier Shell Correlation and Pearson Correlation in the Fourier domain with missing wedge correction, showed a robust ranking of subtomogram clusters without any preprocessing and irrespective of SNR levels of subtomograms. Of these two

[*]Correspondence should be addressed to F.A. (falber@g.ucla.edu).

CRediT authorship contribution statement

Jitin Singla: Conceptualization, Investigation, Methodology, Formal analysis, Software, Validation, Writing - Original Draft, Writing – Review & Editing, Visualization. **Kate L. White**: Investigation, Discussion and Feedback, Writing – Review & Editing. **Raymond C. Stevens**: Investigation, Discussion and Feedback. **Frank Alber**: Funding acquisition, Supervision, Conceptualization, Methodology, Writing - Original Draft, Writing – Review & Editing.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

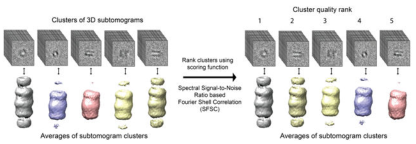
Declaration of Competing Interest

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

scoring functions, Spectral SNR-based Fourier Shell Correlation was fastest to compute and is a better choice for handling large numbers of subtomograms. Our results provide a guidance for choosing an accurate scoring function for template-free approaches to detect complexes from heterogeneous samples.

Graphical Abstract


Clusters of 3D subtomograms
Averages of subtomogram clusters
Rank clusters using scoring function
Spectral Signal-to-Noise Ratio based Fourier Shell Correlation (SFSC)
Cluster quality rank
1
2
3
4
5
Averages of subtomogram clusters

1. Introduction

Cryo-electron tomography (CryoET) has evolved as a promising tool to explore the world within a cell at molecular resolution (Beck and Baumeister, 2016; Oikonomou and Jensen, 2017; Schur, 2019; Zhang, 2019). With the advancement and increased automation of CryoET, it has become easier to collect a vast number of tomograms in a short period of time. Thus, we also require automated methods for the efficient analysis of these tomograms. Over the last few years, various efforts have been made to extract relevant information from tomograms by semi-automated and fully-automated methods. These include use of neural-networks (Che et al., 2018; Chen et al., 2017; Yu and Frangakis, 2011), template-based detection (Beck et al., 2009; Böhm et al., 2000; Lebbink et al., 2007) and template-free pattern mining (Frazier et al., 2017; Martinez-Sanchez et al., 2020; Xu et al., 2019, 2012, 2011). Template-based and neural-network-based methods are successful in detecting complexes in tomograms. However, they are limited to discover only those complexes for which structures are already known.

Template-free, unsupervised methods stand out as they are capable of identifying structures of unknown complexes in tomograms. We previously developed the Multi-Pattern Pursuit (MPP) (Xu et al., 2019), which allows large-scale template-free detection of macromolecular structures in tomograms of heterogeneous samples. The method performs unsupervised clustering of subtomograms into different structural classes and uses an iterative optimization process to select the best combination of alternative clustering results. The underlying structure is then retrieved by averaging the aligned subtomograms in each cluster. MPP, and all other methods based on unsupervised subtomogram clustering, require an effective scoring function for robust quality assessment of clusters and for filtering out unreliable results. Such a quality score can distinguish the homogeneous and well-aligned subtomogram clusters from contaminated and misaligned clusters.

A variety of scoring functions have been developed for image comparisons and cryo-Electron Microscopy (cryoEM) density fitting (Vasishtan and Topf, 2011). These scoring functions measure how well the atomic structure of a complex fits into its electron density maps. Similarly, scoring functions have been used to compare the alignments between 3D electron microscopy volumes (Joseph et al., 2017). However, currently, not much attention has been devoted to scoring functions for assessing subtomogram alignments and the overall

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

quality of a subtomogram cluster, a set of aligned 3D subtomograms that likely contain the same underlying complex. Averaging these subtomograms then produces the structure of the complex. The quality of subtomogram clusters depends on alignment errors between subtomograms and whether or not all the subtomograms in a cluster contain the same underlying complex. These clusters of subtomograms could have been generated by supervised classification and alignment methods or from unsupervised (i.e., reference-free) clustering methods from cryo-electron tomograms of purified complexes, cell lysates or native cellular landscapes containing heterogeneous set of complexes.

In contrast to template-based methods, clusters from unsupervised methods cannot be assessed by comparison to known template structures. So, they must be evaluated by cross-comparison of the similarity of aligned subtomograms. Here, we tested more than 15 scoring functions and compared their ability to rank the quality of subtomogram clusters without knowledge of template structures. The quality of a cluster is ranked higher when they; i) are homogenous in terms of their complex composition, and ii) constituent subtomograms are well-aligned to each other. Scoring functions were tested on sets of both simulated and experimental ground truth subtomograms. For simulated tomograms, we chose five complexes of varying size and shape from the Protein Data Bank (PDB) (Berman et al., 2000) to realistically simulate subtomograms in various different orientations and at three different SNRs (0.001, 0.01, 0.1 - Section 2.1.1). For the test on experimental subtomograms, we used a set of ~800kDa GroEL$_{14}$ and GroEL$_{14}$/GroES$_7$ subtomograms that have been used in other studies as quasi-standard in the field and a set of mammalian 80S ribosome collected with mixed defocus values (Section 2.1.2).

2. Methods

2.1 Data preparation

2.1.1 Simulated data—

As a test set, we used five protein complexes (Table 1) with varying sizes and shapes. Atomic structures of all the five complexes were converted into density maps using the pdb2vol program in the situs package (Wriggers et al., 1999) at 0.4 nm voxel spacing and bandpass filtered at 2 nm. The process takes into consideration the varying number of electrons in different atom types and limits the density map to desired resolution (2 nm). We generated ground truth data sets following a previously established approach for the realistic simulation of the tomographic image reconstruction process (Beck et al., 2009; Förster et al., 2008; Nickell et al., 2005; Pei et al., 2016; Xu et al., 2019, 2012, 2011). It allows the inclusion of noise, tomographic distortions due to missing wedge, and electron-optical factors such as Contrast Transfer Function (CTF) and Modulation Transfer Function (MTF). The density maps served as input for realistically simulating the cryo-electron imaging process with a noise-factor-SNR (SNR: Signal-to-Noise Ratio) of 0.001, 0.01, 0.1 and tilt angle range ±60°. Noise and CTF is added to simulated electron micrographs of each tilt projection. The noise-added and CTF distorted projections are then used in the back-propagation algorithm to reconstruct the final 3D subtomogram. The size of the box in real space is defined by the size of the tilt projection that confines the largest complex. Following a well-established procedure, subtomograms were simulated with voxel size = 0.4 nm, the spherical aberration = 2.2 mm, the defocus value = 7 μm, the voltage =

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

300 kV, the MTF corresponding to a realistic electron detector, defined as sinc(πω/2) where ω is the fraction of the Nyquist frequency. The mean intensity values of Fourier components for simulated subtomograms at different frequencies is very similar to the mean intensity of an experimental subtomogram as described in Section 2.1.2 (Supplementary Figure 1). Finally, we use a back-projection algorithm (Nickell et al., 2005) to generate a subtomogram from the individual 2D micrographs generated at the various tilt angles (Beck et al., 2009; Xu et al., 2011). For each protein complex, we generated 1000 subtomograms, each containing a randomly rotated complex. After simulation, the density values of each subtomogram were normalized to zero mean and unit variance.

We also generated two additional set of simulated subtomograms:

> **Defocus close to focus:** Subtomograms were also simulated for all PDB IDs in the Table 1 at SNR = 0.001, voxel size = 0.4 nm, spherical aberration = 2.2 mm, voltage = 300kV but at defocus value of 2 μm, which is closer to the focus compared to 7 μm. This dataset was used for the assessment of subtomograms with defocus values closer to focus.
>
> **Variable defocus:** Subtomograms were simulated for PDB IDs: 1FNT and 3DY4, at SNR = 0.001, voxel size = 0.4 nm, spherical aberration = 2.2 mm, voltage = 300kV but at five different defocus values (5 μm, 5.5 μm, 6 μm, 6.5 μm and 7 μm). This dataset was used for the assessment of subtomograms with variable defocus values.
>
> **Preferred angular distributions:** Another additional dataset was simulated using complexes GroEL_{14} (PDB ID: 1KP8) and $\text{GroEL}_{14}/\text{GroES}_7$ (PDB ID: 1AON) for assessment of biased angular distributions. The subtomograms were simulated at SNR = 0.001, voxel size = 0.4 nm, spherical aberration = 2.2 mm, voltage = 300kV and defocus value = 5 μm, but instead of randomly rotating the complex before simulation, the rotation was restricted to maximum of 10 degrees for each Euler angle from the starting orientation. The starting orientation was aligned such that the cylindrical axis for both complexes was along the electron beam. This cause the cap region of $\text{GroEL}_{14}/\text{GroES}_7$ to be distorted due to the missing wedge region for all subtomograms.

2.1.2 Experimental Data—We used two sets of experimental subtomograms, one of which has previously been established as a benchmark set in various studies of subtomogram alignment and classification (Förster et al., 2008; Heumann et al., 2011; Hrabe et al., 2012; Scheres et al., 2009; Xu and Alber, 2012; Yu and Frangakis, 2011) and another is a recent dataset collected with state-of-the-art experimental setup (Khoshouei et al., 2017).

GroEL dataset: Förster et al., (2008) collected 786 subtomograms, at voxel size 6Å, of the ~800 kDa GroEL_{14} and $\text{GroEL}_{14}/\text{GroES}_7$ complexes (GroEL_{14}: 214 subtomograms and $\text{GroEL}_{14}/\text{GroES}_7$: 572 subtomograms). Subtomograms were optimally aligned to a template by PyTom (Hrabe et al., 2012) using default parameters and imposed 7-fold symmetry. Out of 572 aligned $\text{GroEL}_{14}/\text{GroES}_7$ subtomograms, 500 were used to generate a primary cluster for computing quality scores. The primary $\text{GroEL}_{14}/\text{GroES}_7$ subtomogram cluster was contaminated, at varying levels, with GroEL_{14} subtomograms. Voxel densities of each

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

subtomogram were normalized to zero mean and unit variance. PDB structures of GroEL$_{14}$ and GroEL$_{14}$/GroES$_7$ are shown in Supplementary Figure 2.

Ribosome dataset: The pre-aligned set of ~3800 subtomograms of 80S mammalian ribosomes (Khoshouei et al., 2017) were provided by the Förster laboratory. Subtomograms were extracted from several tomograms, imaged under varying defocus values and a voxel size of 2.62Å. To trim surrounding empty regions, subtomograms were cropped from 192^3 size to 122^3 voxels. We ranked the aligned subtomograms based on their cross-correlation score with the subtomogram average and selected the top 500 subtomograms for our further cluster analysis. To create contaminations of ribosome clusters, we generated mirror images of the top 500 subtomograms along the x-y plane. The voxel density values of all 1000 subtomograms were normalized to zero mean and unit variance. The PDB structure of 80S ribosome is shown in Supplementary Figure 2.

2.2 Generation of Subtomogram Clusters

A subtomogram cluster is a set of 3D subtomograms, which contains, with exception of contaminations, the same complex. These subtomograms may not be perfectly aligned (Section 2.2.1 and 2.2.2). Subtomogram clusters are frequently produced by supervised or unsupervised clustering methods to identify and align target subtomograms. We created a large set of different subtomogram clusters of varying quality. The subtomogram cluster quality depends on the level of misalignments, i.e., the amount of alignment errors for subtomograms in a cluster and the level of contamination, i.e., the number of subtomograms in a cluster that does not contain the target complex. Contaminations are the result of misclassifications or clustering errors, especially when heterogeneous samples are involved. We generated benchmark sets of simulated and experimental subtomograms at different levels of SNRs. In the following section, we define how misalignment and contamination errors were emulated for subtomogram clusters.

2.2.1 Misalignment—Consider cluster $C(c_1) = \left\{S_i^{c1}\right\}_{i=1}^{N}$ of size N, where $S_i^{c1} \in \mathbb{R}^3$ is a 3D subtomogram of complex c_1 in the original orientation as extracted from a tomogram. Say, $\{\phi_i, \theta_i, \psi_i\}_{i=1}^{N}$ represents the set of Euler rotational angles with which subtomograms in cluster $C(c_1)$ can be rotated to achieve perfect alignment. $C^0(c_1)$ represents a perfectly aligned cluster with misalignment = 0:

$C^0(c_1) = \left\{Rot_{\phi_i, \theta_i, \psi_i}\left(S_i^{c1}\right)\right\}_{i=1}^{N}$, where $Rot_{\phi_i, \theta_i, \psi_i}$ denotes the rotation of the subtomogram S_i^{c1} using Euler angles (ϕ_i, θ_i, ψ_i).

To generate alignment errors in a subtomogram cluster, with misalignment = m, we rotated all the subtomograms in a cluster from their correctly aligned orientation with Euler angles (ϕ_i^m, θ_i^m, ψ_i^m) sampled from a normal distribution $\mathbb{N}\left(0, \frac{m}{3}\right)$ with zero-mean and standard deviation = $\frac{m}{3}$, i.e. $\phi_i^m, \theta_i^m, \psi_i^m \in N\left(0, \frac{m}{3}\right)$. At a standard deviation of $\frac{m}{3}$ approximately 99.7% of sampled Euler angles are within the range $[-m, m]$ degrees (Supplementary Figure

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

3). For example, a misalignment $m = 27$ means that subtomograms were rotated in each Euler direction with angles sampled from a normal distribution $\mathbb{N}\left(0, \frac{27}{3}\right)$, which selects ~99.7% angles between $[-27°, 27°]$.

So, $C^m(c_1)$ represents a cluster of subtomograms containing complex c_1 with misalignment $= m$:

$$C^m(c_1) = \left\{Rot_{\phi_i^m, \theta_i^m, \psi_i^m}\left(Rot_{\phi_i, \theta_i, \psi_i}\left(S_i^{c1}\right)\right)\right\}_{i=1}^{N} \qquad \text{Equation 1}$$

The rotational transformations are also applied to the missing-wedge mask of each subtomogram, and, $\left\{Rot_{\phi_i^m, \theta_i^m, \psi_i^m}\left(Rot_{\phi_i, \theta_i, \psi_i}(M_i)\right)\right\}_{i=1}^{N}$ represents the set of missing-wedge masks corresponding to the subtomograms in the cluster $C^m(c_1)$, where M_i is a binary missing-wedge mask of subtomogram S_i^{c1}.

A set of subtomogram clusters containing only complex c_1, and misalignment $m \in [0, 54]$, is represented as: $C^{m \in [0, 54]}(c_1)$

Here, we do not include shift misalignment, because it would have added another variable parameter, and would have dramatically increased the complexity of the study. Overall, the shift misalignment is easier to refine compared to rotational misalignment, because the rotation has a much larger search space for optimization.

2.2.2 Contamination—In both supervised classification and unsupervised clustering of subtomograms, complexes of different types but similar shapes or sizes may be falsely co-assigned to the same cluster. To assess scoring functions for their ability to detect contamination, we generated clusters $C_p(c_1,c_2)$ that mainly contain subtomograms of complex c_1 and subtomograms of contaminant complex c_2, of similar size or shape:

$$C_p(c_1, c_2) = \left\{S_i^{c1}\right\}_{i=1}^{N1} \oplus \left\{S_j^{c2}\right\}_{j=1}^{N2} \qquad \text{Equation 2}$$

where, $\oplus$ represents concatenation of two sets and S_i^{c1} and S_i^{c2} are subtomograms containing complex c_1 and c_2, respectively. N_1 and N_2 are the numbers of subtomograms for each complex. $N_1 + N_2 = N$ (cluster size). p is the amount of contamination error, defined as the percentage of the total number of subtomograms in the cluster that contain contaminant complex c_2, i.e., $N_2 = \frac{pN}{100}$.

A subtomogram cluster can have both alignment and contamination error. $C_p^m(c_1, c_2)$ represents a cluster of subtomograms containing complex c_1, contaminated with complex c_2 with p contamination percentage, as well as having m degrees of misalignment error:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$C_p^m(c_1, c_2) = \left\{ Rot_{\phi_i^m, \theta_i^m, \psi_i^m}\left(Rot_{\phi_i, \theta_i, \psi_i}\left(S_i^{c1}\right)\right)\right\}_{i=1}^{N_1} \oplus \left\{ Rot_{\phi_j^m, \theta_j^m, \psi_j^m}\left(Rot_{\phi_j, \theta_j, \psi_j}\left(S_j^{c2}\right)\right)\right\}_{j=1}^{N_2} \qquad \text{Equation 3}$$

A set of clusters having misalignment = m and different contamination error p, $p \in [0, 40]$, is represented as $C_{p \in [0, 40]}^m(c_1, c_2)$. We assessed scoring functions for their performance in ranking cluster quality with a set of clusters, containing varying levels of misalignments or contaminations or both. Clusters were generated using ten different pairs of complexes for simulated subtomograms (Table 1) and two pairs of complexes for experimental subtomograms.

2.2.3 Simulated benchmark set—

For each of the five complexes, clusters were generated with misalignment values m ranging from [0, 54] degrees with a step size of 5.4. Also, subtomogram clusters for each complex were contaminated with another complex, with a contamination percentage p ranging from [0, 40] with a step size of 10. For each subtomogram cluster, we tested the assessment for contamination with two different contamination complexes. Moreover, all clusters were simulated for three different SNR = {0.001, 0.01 and 0.1} (Table 1 and Supplementary Figure 2). In total, we generated a benchmark set of 1650 simulated subtomogram clusters with varying quality in terms of misalignment, and level of contamination and SNR. Each cluster contained a total of 500 subtomograms.

2.2.4 Experimental benchmark set—

Subtomogram clusters were generated for experimental subtomograms of GroEL$_{14}$/GroES$_7$ and ribosomes using the same misalignment and contamination range as applied for simulated subtomograms. GroEL$_{14}$/GroES$_7$ clusters were contaminated with GroEL$_{14}$ and ribosome clusters were contaminated with subtomograms containing the mirrored image of each ribosome. In total, a benchmark set of 55 subtomogram clusters were generated for each experimental dataset. Each cluster contained a total of 500 subtomograms.

2.3 Voxel Regions

We define three different regions of voxels in a subtomogram for computing the individual scores (Figure 1).

Global: The global score is computed from all the voxels in the subtomogram (Figure 1).

Contoured scores: Global scores can be affected by voxels that are not occupying the target complex and largely represent noise. To reduce the impact of noise-dominant voxels, we define a contoured region using only voxels that are likely to occupy any of the two target complexes. This procedure will prioritize complex containing voxels when computing the scores, therefore reducing the contribution of noise. To define contoured region, we select all the voxels with density values higher than one-and-half times the standard deviation of all voxel densities (> 1.5 σ). The score between two subtomograms is then

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

calculated from the union of selected voxels in both subtomograms. A threshold of 1.5 σ gives the highest value of the Jaccard index between the segmented masks and the complex containing voxels in the subtomogram (Supplementary Figure 4). The Jaccard index measures the similarity of two regions as the number of voxels in the intersection of both voxel sets divided by the number of voxels in the union of both sets. The sigma value can be customized based on the application and prior testing. For our analysis sigma = 1.5 is a good choice.

Overlap: A threshold is applied as in the contoured score definition. The score between two aligned subtomograms is calculated from the intersection of selected voxels in both subtomograms, focusing only on the overlapping complex-containing region.

2.4 Scoring Functions

In this section, we define the scoring functions for quality assessment of subtomogram clusters. The density values of each subtomogram image are normalized to zero mean and unit variance.

2.4.1 SFSC: Spectral SNR-based Fourier Shell Correlation—SFSC measures the SNR from the variance in the voxel intensities at all spatial frequencies, as previously introduced in the MPP method (Xu et al., 2019). SFSC uses all the subtomograms in the cluster and considers missing-wedge masks, one of the major distortions in cryoET, due to a limited range of angles to capture tilt series.

Say cluster C of size n contains the set of subtomograms $\{S_1, S_2 \ldots S_n\}$, with real component of Fourier Transforms $\{F_1, F_2 \ldots F_n\}$ and corresponding binary missing wedge masks $\{M_1, M_2 \ldots M_n\}$. The Spectral-Signal-to-Noise Ratio (Spectral-SNR or SSNR) η_r at frequency r is defined as:

$$\eta_r = \frac{\int_{\left|\,|\xi| - r\right| < \Delta r} \widehat{M}(\xi) \left|\mu(\xi)\right|^2}{\int_{\left|\,|\xi| - r\right| < \Delta r} \sigma^2(\xi)}$$

Equation 4

where $\Delta r = 1$, $\xi \in \mathbb{R}^3$ is location in Fourier space, $\widehat{M}$ is sum of missing wedge masks:

$$\widehat{M}(\xi) = \sum_i M_i(\xi)$$

Equation 5

$$\mu(\xi) = \frac{\sum_i M_i(\xi) F_i(\xi)}{\widehat{M}(\xi)}$$

Equation 6

and

$$\sigma^2(\xi) = \frac{\sum_i M_i(\xi) \left|M_i(\xi) F_i(\xi) - \mu(\xi)\right|^2}{\widehat{M}(\xi) - 1}$$

Equation 7

Given the η_r (i.e., SSNR, Eq. 4) at frequency r, ζ_r (i.e., FSC, Eq. 8) can be estimated as:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\zeta_r = \frac{\eta_r}{2 + \eta_r} \quad \text{Equation 8}$$

Then SFSC is defined as sum of FSC over all frequencies:

$$\hat{\zeta} = \sum_r \zeta_r \quad \text{Equation 9}$$

The higher the value of $\hat{\zeta}$ (i.e., SFSC, Eq. 9), the higher is quality of a subtomogram cluster.

The SFSC score is computed from the set of all individual subtomograms, while all other scores are calculated from pairwise comparisons of subtomograms in the same cluster.

2.4.2 gPC: Global Pearson Correlation—

gPC is the global Pearson correlation score and uses all the voxels in real space for both subtomograms to calculate the Pearson-correlation. The gPC between a pair of subtomograms (X, Y) is calculated as follows:

$$gPC(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_i (X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_i (X_i - \mu_X)^2}\sqrt{\sum_i (Y_i - \mu_Y)^2}} \quad \text{Equation 10}$$

where X_i and Y_i are density values for the i^{th} voxel of subtomograms X and Y, respectively. μ_X and μ_Y are mean density values over corresponding voxel region in each subtomogram.

Because each subtomogram is normalized to zero mean and unit variance ($\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$), gPC becomes directly proportional to the cross-correlation function (CCF).

$$gPC(X,Y) = \frac{\sum_i X_i Y_i}{N} \propto \sum_i X_i Y_i = CCF(X,Y) \quad \text{Equation 11}$$

The gPC score and all following scores are calculated by randomly picking 10% of all possible pairs of subtomograms in a cluster. The total score is then defined as the average over all the pairwise scores. We show separately that for the gPC and all following scores, a random selection of 10% of pairs is sufficient to capture the population mean by comparing 10% and 50% of all possible pairs. Due to increased time complexity for computing 50% pairs (62375 pairs), we show this test for only one structure (PDB ID: 2GHO), contaminated with structures (PDB IDs: 1QO1, 2H12) at SNR = 0.01, misalignment = 21.6 degrees and contamination range [0, 30] percentage. Supplementary Table 1 shows the resulting scoring value for few scoring functions for 10% and 50% pairs. We observed that 10% of subtomogram pairs are sufficient to capture the same amount of information as 50% subtomogram pairs.

2.4.3 cPC: Contoured Pearson Correlation—

cPC is calculated as defined in gPC. However, only the union of voxels in both subtomograms with density values larger than the threshold (X_i, $Y_i > 1.5\ \sigma$) are considered.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2.4.4 oPC: Overlap Pearson Correlation—oPC is calculated as defined in gPC. However, only the intersection of voxels from both subtomograms with density values larger than the threshold ($X_i, Y_i > 1.5\ \sigma$) are considered.

2.4.5 FPC: Pearson correlation in Fourier space—We computed the Pearson correlation in the Fourier Space as well. Say $F(X)$ and $F(Y)$ are real components of Fourier Transforms of subtomogram X and Y respectively. Then Pearson Correlation in Fourier space is computed as:

$$FPC(X, Y) = \frac{\sum_i \left(F_i(X) - \mu_{F(X)}\right)\left(F_i(Y) - \mu_{F(Y)}\right)}{\sqrt{\sum_i \left(F_i(X) - \mu_{F(X)}\right)^2}\sqrt{\sum_i \left(F_i(Y) - \mu_{F(Y)}\right)^2}} \qquad \text{Equation 12}$$

where $F_i(X)$ and $F_i(Y)$ are values at i^{th} voxel of Fourier Transforms of subtomograms X and Y, respectively. $\mu_{F(X)}$ and $\mu_{F(Y)}$ are mean intensity values of voxels in Fourier Transforms.

2.4.6 FPCmw: Pearson correlation in Fourier space with missing wedge correction—We also calculated the Pearson correlation in Fourier space with missing wedge correction. The overlap missing wedge mask in Fourier's space is defined as the intersection of missing wedge masks of both subtomograms. Say $F(X)$ and $F(Y)$ are real components of Fourier Transforms and $M(X)$ and $M(Y)$ are binary missing wedge masks of subtomogram X and Y respectively, then FPCmw score can be written as:

$$FPCmw(X, Y) = \frac{\sum_i M_i(X)M_i(Y)\left(F_i(X) - \mu_{F(X)}\right)\left(F_i(Y) - \mu_{F(Y)}\right)}{\sqrt{\sum_i M_i(X)M_i(Y)\left(F_i(X) - \mu_{F(X)}\right)^2}\sqrt{\sum_i M_i(X)M_i(Y)\left(F_i(Y) - \mu_{F(Y)}\right)^2}} \qquad \text{Equation 13}$$

where,

$$\mu_{F(X)} = \frac{\sum_i M_i(X)M_i(Y)F_i(X)}{\widehat{M}} \qquad \text{Equation 14}$$

$$\mu_{F(X)} = \frac{\sum_i M_i(X)M_i(Y)F_i(Y)}{\widehat{M}} \qquad \text{Equation 15}$$

$$\widehat{M} = \sum_i M_i(X)M_i(Y) \qquad \text{Equation 16}$$

FPCmw (Xu and Alber, 2012) is similar to the constrained cross-correlation (CCC), which is widely used for subtomogram alignment (Förster et al., 2008). Both FPCmw and CCC are constrained cross-correlation functions, which consider missing wedge corrections—the only difference is that CCC is computed in real space (after missing wedge correction), rather than Fourier space (after missing wedge correction) for FPCmw. Therefore, FPCmw is faster as it does not require the computation of the inverse Fourier Transform. We show the close similarity of both scores for simulated subtomogram clusters $C_p^m(1FNT, 1BXR)$ and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$C_p^m(1FNT, 3DY4)$ with $m \in [0, 54]$ and $p \in [0, 40]$ (Section 2.2.2) in Supplementary Figure 5.

Overall, we have five variants of the Pearson correlation scores defined, i.e. gPC, cPC, oPC, FPC and FPCmw.

2.4.7 gMI: Global Mutual Information—

Mutual information scores were previously used (i) to improve the alignment of class-averages in Single Particle Analysis (SPA) (Shatsky et al., 2009), (ii) to fit crystal structures in cryo-density maps and (iii) to assess structures determined by cryo-electron microscopy (Joseph et al., 2017; Vasishtan and Topf, 2011). Here we define a mutual information score to calculate the quality of a subtomogram cluster. The density values of all voxels in the desired voxel region were divided into k number of bins. The number of bins k was defined following the Sturges rule (Sturges, 1926) as:

$$k = int(1 + \log_2 n) \qquad \text{Equation 17}$$

where n is the total number of voxels.

Marginal entropies were then calculated for both the subtomograms X and Y as

$$H_X = -\sum_{i=1}^{k_X} p_i * \log_2 p_i \qquad \text{Equation 18}$$

$$H_Y = -\sum_{j=1}^{k_j} p_j * \log_2 p_j \qquad \text{Equation 19}$$

where p_i and p_j are the probabilities of finding a voxel density value in i^{th} and j^{th} bins in subtomogram X and Y respectively. k_X and k_Y are the number of bins into which subtomogram X and Y were divided.

$$p_i = \frac{Number\ of\ voxels\ in\ the\ voxel\ region\ of\ subtomorgam\ X with\ density\ value\ in\ i^{th}\ bin}{Total\ number\ of\ voxels\ in\ the\ voxel\ region\ of\ subtomogram\ X}, \quad i \in [1, k_X] \qquad \text{Equation 20}$$

Similarly, p_j is defined for subtomogram Y with $j \in [1, k_Y]$.

The joint entropy was computed as

$$H_{XY} = -\sum_{i=1}^{k_X}\sum_{j=1}^{k_Y} p_{ij} * \log_2 p_{ij} \qquad \text{Equation 21}$$

where, p_{ij} is the probability of finding the pair of bins i,j in the subtomogram pair. The joint entropy is minimum when there is no difference between subtomogram X and Y. Then gMI was calculated using all voxels in the subtomograms as:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$gMI(X, Y) = H_X + H_Y - H_{XY} \quad \text{Equation 22}$$

Also, if subtomograms X and Y are normalized to have zero means and unit standard deviations, H_X and H_Y are approximately equal and constant for any pair of subtomograms containing the same structure and SNR. Therefore, mutual information, in that case, is inversely proportional to joint entropy.

2.4.8 cMI: Contoured mutual Information—cMI score is calculated as defined in gMI. However, only the union of voxels in both subtomograms with density values larger than the threshold ($X_i, Y_i > 1.5\ \sigma$) are considered.

2.4.9 oMI: Overlap mutual Information—oMI score is calculated as defined in gMI. However, only the intersection of voxels in both subtomograms with density values larger than the threshold ($X_i, Y_i > 1.5\ \sigma$) are considered. oMI has also been used before but called Local Mutual Information (Joseph et al., 2017).

2.4.10 NMI: Normal Mutual Information—We also calculated a normalized version of the mutual information sore. The NMI score is calculated as:

$$NMI(X, Y) = \frac{H_X + H_Y}{H_{XY}} \quad \text{Equation 23}$$

where H_X and H_Y as the marginal entropies calculated from subtomograms X and Y and H_{XY} is the joint entropy. The statistical power of estimated probabilities decreases as the overlap between subtomograms decreases. But NMI (Studholme et al., 1999) make gMI more robust to overlap volume.

2.4.11 gNSD: Global Normalized Squared Deviation—Squared Deviation (SD) between two subtomograms is defined by the sum of squared difference between the density values of corresponding voxels in the two aligned subtomograms.

$$SD(X, Y) = \sum_i (X_i - Y_i)^2 \quad \text{Equation 24}$$

where X_i and Y_i are voxel densities at i^{th} voxel of subtomograms X and Y respectively. For global Squared Deviation (gSD), the score comes out to be directly proportional to cross correlation function.

$$\begin{aligned} & gSD(X, Y) = \sum_i X_i^2 - 2X_iY_i + Y_i^2 \\ & = \sum_i X_i^2 + Y_i^2 - 2\sum_i X_iY_i \\ & = constant - 2CCF(X, Y) \\ & \propto CCF(X, Y) \end{aligned}$$

The gNSD score is then defined by min-max normalization of gSD scores of set of clusters under analysis and by a subtraction from 1 to define a score that increases with quality.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$$gSD(X, Y) = 1 - minmax\ normalized\ gLSF \qquad \text{Equation 25}$$

2.4.12 cNSD: Contoured Normalized Squared Deviation—

The cSD score is calculated as

$$CSD = \frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{N} \qquad \text{Equation 26}$$

where X_i and Y_i are voxel densities at i^{th} voxel in the contoured voxel region of subtomograms X and Y respectively and N is the total number of voxels in the contoured region.

In contrast to gNSD, only the union of voxels in both contoured subtomograms with density values larger than the threshold ($X_i, Y_i > 1.5\ \sigma$) are considered for cNSD.

2.4.13 oNSD: Overlap Normalized Squared Deviation—

oNSD score is calculated as defined in cNSD. However, only the intersection of voxels in both subtomograms with density values larger than the threshold ($X_i, Y_i > 1.5\ \sigma$) are considered.

2.4.14 DSD: Difference Squared Deviation—

The DSD score is similar to SD. However, instead of using density values directly, it uses the difference of density values between the pairs of corresponding voxels in the two subtomograms.

$$DSD(X, Y) = \sum_{i,j}\left(\left(X_i - X_j\right) - \left(Y_i - Y_j\right)\right)^2 \qquad \text{Equation 27}$$

where (i,j) is the pair of voxels, X_i, X_j, Y_i, Y_j are density values at voxel indices i and j for subtomograms X and Y. As the number of all possible voxel pairs can be very expensive to compute, we only used 10,000 randomly selected voxel pairs that have density values higher than a particular threshold. Here we chose that threshold to be the standard deviation of voxel densities in a subtomogram. Similar to SD, DSD also represents the difference between the subtomograms, so after min-max normalization of the score, we subtract it from 1. DSD we mention throughout Results section is:

$$DSD = 1 - minmax\ normalized\ DLSF \qquad \text{Equation 28}$$

2.4.15 OS: Overlap Score—

The overlap score is defined as the fraction of contoured voxel regions that are part of the intersection of both subtomograms.

$$OS(X, Y) = \frac{vol_{overlap}(X, Y)}{\min(vol_{contoured}(X), vol_{contoured}(Y))} \qquad \text{Equation 29}$$

where $vol_{contoured}$ is the volume of contoured regions in a subtomogram and $vol_{overlap}(X, Y)$ is the volume for overlap regions in subtomograms X and Y (contour and overlap regions are defined as previously described).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2 summarizes the scoring functions compared in this study categorized based on different voxel regions used for computing the score value. Supplementary Table 2 summarizes previously published literature for scoring functions in Table 2. The implementation of scoring functions is also available on GitHub (https://github.com/alberlab/cryoET_ScoringFunctions).

2.5 Estimation of *effective-SNR*

We estimated the *effective-SNR* of subtomograms as previously described (Frank and Al-Ali, 1975; Xu et al., 2019). By calculating effective-SNR levels for both experimental and simulated tomograms, we validated that *SNR* levels of simulated subtomograms are at a comparable range to those in experiment.

2.5.1 Simulated Data—

At each SNR level, we sampled 10,000 pairs of aligned subtomograms for each of the five complexes. For each pair of subtomograms, we calculated the Pearson correlation of their voxel densities and then estimated the corresponding *effective-SNR* according to (Frank and Al-Ali, 1975):

$$effective-SNR = \frac{\sum_{p=1}^{N} \frac{c_p}{1 - c_p}}{N} \qquad \text{Equation 30}$$

where, N is the number of pairs of aligned subtomograms and c_p is the Pearson correlation between subtomograms in pair p. To estimate the *effective-SNR* for a given simulated SNR level, we average the *effective-SNR* for each of the five complexes. This procedure calculates an *effective-SNR* of ~0.002, ~0.01 and ~0.08 for subtomograms simulated at SNR levels of 0.001, 0.01 and 0.1, respectively (Supplementary Table 3).

We also tested the *effective-SNR* when only a subset of voxels is selected in real and Fourier space. We tested these three additional criteria: i) using only voxels of the actual complex in real-space, ii) using all frequencies in Fourier space, and iii) using only frequencies excluded from the missing-wedge mask in Fourier space. We show that using the complex containing voxels in real-space or considering missing wedge mask in Fourier space still estimates the SNR with the same order of magnitude (i.e., decimal range) as the simulated SNR (Supplementary Table 4).

2.5.2 Experimental Data—

We also calculated the *effective-SNR* for experimental tomograms from 10,000 aligned subtomogram pairs. The *effective-SNR* for GroEL$_{14}$/GroES$_7$ was ~0.115, for GroEL$_{14}$ ~0.113 and for ribosomes ~0.0001 in real space and ~0.0003 in Fourier space.

2.6 Gaussian Filtering of subtomograms

We also evaluated scoring functions when, as a preprocessing step, a Gaussian filter, with two kernel values (σ = 1 and σ = 2), is applied to subtomograms. Gaussian filtering in real-space is equivalent to Gaussian low-pass filtering in Fourier space, and the filtered resolution can be directly computed from the standard deviation of the Gaussian function used for blurring the subtomogram (Supplementary Note 1). It removes high frequency components

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and subsequently density variance from noise-dominated voxels, which improves the segmentation of contour and overlap mask segmentation in single subtomograms (Supplementary Figure 6). We used python package Scipy to filter the 3D subtomograms (Virtanen et al., 2020). The standard deviation of Gaussian filter used in real space directly relates to frequencies low bandpass filtered in Fourier space (Supplementary Note 1).

3. Results

The quality of a subtomogram cluster depends on various factors, including (i) subtomogram misalignments and (ii) cluster contaminations. Subtomogram misalignments (i.e., alignment errors) are non-optimal alignments of two subtomograms, which result from low accuracy of alignment programs, in particular for subtomograms of low resolution and with high noise levels. Cluster contaminations (i.e., assignment errors) occur when subtomograms with structures other than the target complex are classified into the same cluster. This can be the result of errors in classification programs due to subtomograms with low resolution and higher noise levels.

To assess each scoring function for correctly ranking the quality of subtomogram clusters based on misalignment and contamination errors we compute the Spearman's rank correlation coefficient (ρ) between the predicted subtomogram cluster quality and the amount of actual error in the clusters. A Spearman's correlation of ρ = 1 indicates a strictly monotonic behavior of the quality score so that the scoring function values decrease with increasing errors in the subtomogram clusters. The main criteria to categorize the scoring function as useful will be its ability to correctly rank the clusters in the order of their actual quality, i.e., a monotonic decrease in the determined cluster quality score will then agree with an increase in the amount of alignment or contamination errors.

3.1 Assessment against Misalignment

We first assess the scoring function performance when only alignment errors are introduced in clusters, i.e., contamination = 0 for perfectly homogeneous clusters. Each cluster contains a total of 500 subtomograms. We generated 11 clusters for each of the five benchmark complexes, and each sampled with an increasing range of misalignments from 0 to 54° (step size = 5.4 degrees, Section 2.2). Because the angles for misalignments are sampled randomly from a normal distribution, we repeated the process three times and averaged the scores over the three replicates.

Figure 2A shows average structures from subtomogram clusters with increasing misalignment levels and the performance of all scoring functions to rank the quality of clusters for an example complex (PDB ID: 1FNT, using subtomograms at SNR = 0.001, i.e., $C_{p=0}^{m \in [0,54]}(1FNT)$ (Section 2.2.1). To allow comparison between scoring functions, scores were min-max normalized to the range [0, 1]. Also, to compute Spearman's correlation (ρ), we ranked zero misalignment as the top rank among error amounts.

Supplementary Table 5 lists the Spearman's correlations (ρ) for all scoring functions averaged over all benchmark complexes. The scoring functions differed greatly in their performance, with Spearman's correlations ρ ranging from 1.0 to −0.93. Five scoring

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

functions, **SFSC**, **gPC, gNSD, FPC** and **FPCmw** (Section 2.4) stand out as they showed excellent performance with averaged Spearman’s correlations ρ > 0.95 over the entire benchmark set, indicating that clusters can be well ranked by their ground truth quality. We noticed that all scoring functions that depend on segmented subtomogram regions (i.e., contoured and overlap regions) did not perform well for subtomograms at such low SNR value (SNR = 0.001). That is because thresholding for selecting candidate voxel regions cannot always correctly identify the volume containing the actual structure of the complex (Supplementary Figure 6A). Preprocessing can improve the thresholding for segmenting regions of the actual target complex even for very low SNR subtomograms (Section 3.3). Global and Overlap Mutual Information failed to rank clusters with subtomograms at such high noise levels. Mutual Information is inversely proportional to the joint entropy of two subtomograms containing the same underlying structure (Section 2.4.7). If subtomograms were perfectly aligned, their joint entropy is lower compared to misaligned subtomograms, i.e., the Mutual Information is higher for aligned subtomograms. This holds true only when bins with voxel intensity values of the target complex have higher probabilities than those of other regions in the subtomogram. But at very high noise levels, probabilities are more widespread across intensity bins. The performance of the mutual information score will improve by increasing the SNR of subtomograms or by preprocessing individual subtomograms. The ρ values of **gMI** and **oMI** improved when subtomograms were Gaussian filtered or were generated at higher SNR (Sections 3.3, 3.5).

3.2 Assessment of Cluster Contamination

We then assessed scoring functions with respect to cluster contamination, which can result from assignment errors. Clusters of a benchmark complex were contaminated with subtomograms containing other structures (Section 2.2.2, Table 1). We generated 5 clusters per benchmark complex, which varied in the level of contamination ranging from 0 to 40%. We first assessed these clusters without containing any alignment errors. Figure 3 shows the cluster averages and depicts the min-max normalized scores for an example complex (PDB ID: 1FNT) contaminated with another complex (PDB ID: 1BXR), i.e., $C_{p \in [0,40]}^{m=0}(1FNT, 1BXR)$ (Eq. 3). Also, here, the scores **SFSC**, **gPC**, **gNSD**, **FPC** and **FPCmw** showed the best performance in predicting the quality of the contaminated clusters (Supplementary Table 6). Similar to the assessment against misalignments, most scoring functions that depend on segmented subtomogram regions and scores based on mutual information failed to rank the quality of clusters accurately.

3.3 Effect of Gaussian Filtering

Next, we tested if preprocessing of subtomograms with Gaussian filtering improves the performance of scoring functions, in particular for subtomograms with low SNR values. Gaussian filtering in real-space is equivalent to Gaussian low-pass filtering in Fourier space, which means damping higher frequencies to reduce the amount of noise. We tested Gaussian filtering with two different kernels (σ = 1 and 2, Section 2.6). Applying a Gaussian kernel enhanced the global structural features of the complex against background noise for subtomograms with low SNR of 0.001 (Supplementary Figure 6B). However, with an increase in σ, naturally, the structures also lose their high-resolution features. At very low

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

SNR (SNR = 0.001), Gaussian filtering improved the automatic thresholding of subtomograms to detect contoured and overlap regions (Supplementary Figure 6A). Gaussian filtering, therefore, improved the performances for some of the scoring functions (Table 3).

The scores **gPC**, **gNSD**, **FPC** and **FPCmw**, which performed well without applying Gaussian filtering, maintain their good performance. The scores **cPC**, **oMI**, **cNSD**, **DSD** and **OS**, which failed to rank the quality of subtomogram clusters without Gaussian filter preprocessing, now show sufficiently improved Spearman's correlation with ρ > 0.95 for assessment against misalignment and ρ > 0.85 for assessment of cluster contamination (Table 3). Therefore, these scores can rank clusters in the desired order of quality based on subtomogram misalignments and cluster contamination. In general, scores based on Mutual Information (**gMI**, **cMI**) fail across all Gaussian kernel settings for both misalignment and contamination tests, except for the overlap based Mutual information score (**oMI**), which shows reasonable improvements when applying a Gaussian filter (Table 3). Some scores only perform well with a narrow window of Gaussian kernel value. For instance, **oPC** (overlap Pearson correlation) performs best using Gaussian kernels with an intermediate value (σ = 1) and lose their performance with larger kernel values (Table 3). This holds true for both misalignment and contamination tests. **SFSC** decreases in performance when applying a Gaussian kernel with relatively high σ values because **SFSC** measures the variance of voxel intensities between the constituent subtomograms of the cluster (Table 3). With an increase in σ, the variation in high-frequency structural features is lost. So, **SFSC** works well when subtomograms are not preprocessed using a Gaussian filter.

3.4 Varying Misalignment and Contamination at the same time

In our analysis so far, we tested scores separately either with respect to misalignments or contamination error. Now, we want to assess how scoring functions perform when misalignments and contaminations are introduced simultaneously. We assessed the performance by calculating the average Spearman's correlation ρ for a given score across all ten target-contaminant pairs (each benchmark complex is tested with two different contaminant, Table 1 and Supplementary Figure 2). We first tested the scoring function's ability to rank clusters with varying levels of misalignments at each level of contamination (from 0 to 40%). These tests were performed with subtomograms simulated with relatively low SNR level (SNR=0.001), i.e., $C_{p=0}^{m \in [0,54]}$, $C_{p=10}^{m \in [0,54]} \ldots C_{p=40}^{m \in [0,54]}$ (see Section 2.2, Eq. 3). The **SFSC** score showed excellent performance for ranking clusters against misalignment errors across all contamination levels (Figure 4A). Scoring functions based on global Pearson correlation (**gPC**), its Fourier-based variants with (**FPCmw**) and without missing wedge corrections (**FPC**), and global Normalized Squared Deviation (**gNSD**) also showed excellent performance against misalignments (with ρ > 0.95), except for the highest contamination level of 40% (Figure 4A). All other scoring functions, in comparison, perform very poorly. Also, some scoring functions (**cPC**, **gMI, oMI, NMI and OS)** had high negative correlations. However, the negative correlation is not consistent across all SNR levels and datasets (Section 3.5), which implies that these scoring functions are not reliable for estimating subtomogram cluster quality.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Next, we assessed the scores for their ability to rank clusters with varying levels of contamination for each level of misalignment, ranging from 0 to 54 degrees (i.e., $C_{p \in [0,40]}^{m=0}$, $C_{p \in [0,40]}^{m=5.4} \ldots C_{p \in [0,40]}^{m=54}$, see Section 2.2, Eq. 3). Also here, **SFSC**, **gPC**, **FPCmw**, **FPC** and **gNSD** showed excellent performance to rank contamination across all levels of misalignments, except for the highest misalignment error of 54 degrees, at which **FPC** and **FPCmw** dropped performance below our threshold level of $\rho = 0.85$. All other scores performed very poorly across all misalignment ranges and, therefore, cannot rank correctly cluster quality (Figure 5A).

In practical applications, both assessment against misalignment and contamination will be useful. For example, if we have two clusters A and B, say, with different sets of subtomograms containing the same target complexes containing possibly different levels of contaminations, then we can first minimize alignment errors for clusters A and B independently, by finding alignments that produce a maximum score (Section 3.1). Then the score difference between cluster A and B will primarily be due to contamination, i.e., the cluster with the higher score is more homogeneous (Section 3.2).

Preprocessing with Gaussian filters: Preprocessing of subtomograms with Gaussian filters (σ= 2) improved the performance for those scoring functions that rely on segmented subtomograms. Particularly, **cPC** and **oMI** showed dramatic improvements with Gaussian filters (at $\sigma = 2$) for ranking misalignments across all levels of contamination even at SNR=0.001 (Supplementary Figure 7B). However, these scores performed much poorer for the ranking of contamination errors, especially when larger levels of misalignment errors were present (Supplementary Figure 7D). **cNSD** and **OS** scores performed better in their ability to rank clusters with varying levels of contamination only for lower levels of misalignment errors (Supplementary Figure 7D). Global scores based on Pearson correlations, in real and Fourier space, (**gPC**, **FPC** and **FPCmw)** retained their good performance with applied Gaussian filtering at high misalignment and contamination levels. Also, as seen earlier, **SFSC**'s performance decreased with increasing σ in Gaussian filtering, due to loss in voxel-density variations (Supplementary Figure 7). These observations confirm that **SFSC** scores work best without subtomogram preprocessing with Gaussian filters.

Average masked scoring functions: So far, contoured scoring functions (**cMI**, **cNSD**, **cPC**) considered target regions from the union of separately detected contour regions in each subtomogram. We also tested the performance of contoured scoring functions when the target regions were detected by the average of all subtomograms in the cluster. To do so, we selected the target region by defining a mask from the cluster average, instead of the density values of each subtomogram. We therefore introduced the contour average-masked scoring functions: i) average-masked Mutual Information (**amMI**), ii) average-masked Normalized Squared Deviation (**amNSD**), iii) average-masked Pearson Correlation (**amPC**), and iv) average-masked Constrained Cross-Correlation (**amCCC**). The **amCCC** is a widely used scoring function for subtomogram alignments. We tested the performance of all these additional scoring functions both against misalignment and contamination for simulated subtomograms at SNR = 0.001. We observed that the cluster average-mask scoring

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

functions, except **amCCC** performed poorly and below the required threshold for simulated subtomograms at SNR = 0.001 (Supplementary Figure 8). **amCCC** is the only scoring function among these four that have performed above the required threshold.

Assessment for subtomograms with defocus level close to focus: We repeat the analysis conducted for SNR = 0.001 and 7 μm defocus, but for defocus value of 2 μm, which is closer to focus value. We observed that SFSC performs far better than any other scoring function for smaller defocus value, both against misalignment (Supplementary Figure 9A) and contamination (Supplementary Figure 9B).

Assessment for subtomograms with variable defocus levels: When particles are extracted from different tomograms, it is possible that they were imaged with different defocus levels. These differences can affect the assessment of cluster quality. We, therefore, also tested the performance of scoring functions with subtomograms collected with different defocus levels, i.e., with subtomograms distorted with different CTF. We simulated sets of subtomograms for complex pair (1FNT, 3DY4), with 5 different defocus values ranging between 5 μm and 7 μm (Supplementary Figure 10AB, Section 2.1.1). Each cluster contained a mix of subtomograms simulated with different defocus. Scoring functions were then tested against cluster misalignment and contamination errors. We observed similar results to our analysis with subtomograms generated from a single defocus value. The same set of scoring functions (**SFSC**, **gPC**, **FPC**, **FPCmw** and **gNSD)** performed well and above the required threshold, while all others failed to provide robust quality ranking (Supplementary Figure 10CD). Also, out of the four average-masked scoring functions only **amCCC** showed good performance.

3.5 Assessment against SNR

Signal-to-Noise-Ratio (SNR) is one of the important factors that affect the performance of scoring functions. So, we simulated subtomograms at three different SNRs [0.001, 0.01 and 0.1]. For these simulated subtomograms, we also computed the *effective-SNR* as described previously (Xu et al., 2019). The *effective-SNR* levels (for target SNR 0.001, 0.01 and 0.1) were 0.002, 0.01 and 0.08, respectively (Section 2.5.1), which indicates that the simulation process adds the required amount of noise to the subtomograms. At low SNR levels (SNR=0.001), only 5 out of 15 scoring functions were capable to rank clusters based on misalignment and contamination errors (Figure 4A, 5A). With increasing SNR levels, we observed improved performances even without Gaussian filtering, for those scoring functions that rely on threshold-based segmentation of contoured and overlap voxel regions. At the highest SNR = 0.1, almost all scoring functions (**SFSC**, **gPC**, **cPC**, **FPC**, **FPCmw**, **oMI**, **gNSD**, **cNSD**, **DSD** and **OS**) showed excellent performance and were all equally competent to distinguish the amount of misalignment in the clusters across all contamination levels (Figure 4C). However, for the ranking of contamination levels, only **SFSC**, **gPC**, **FPC**, **FPCmw** and **gNSD** performed above the threshold, except for very high misalignment levels, for which **SFSC**’s performance dropped below the threshold to rank cluster contamination (Figure 5BC). **oPC**, **cMI** and **oNSD** still performed very poorly across all contamination and error levels (Figure 5C).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

As expected, **gMI**, **oMI** and **NMI** scores, which are based on mutual information, increased in performance with increasing SNR (Figure 4). However, at intermediate and low SNR, all three scores still failed to rank clusters reliably. At the highest SNR=0.1, only the **oMI** score reached an acceptable Spearman's correlation $\rho > 0.95$ threshold for ranking misalignments across all levels of contamination (Figure 4C). **gMI, oMI** and **NMI** were able to rank clusters based on contamination errors only if low levels of misalignment errors are present (Figure 5C). The **cMI** score failed to rank clusters even at the highest SNR levels. We also observed that the Overlap score (**OS**), performed well at the highest SNR level and ranked well misalignments across all contamination levels (Figure 4C). With an improved signal component in the subtomograms, the thresholding for selecting accurate overlap voxel regions improves. So, misalignment among subtomograms were easily recognized by the Overlap score (**OS**). Also, our complexes have non-spherical shapes, and complexes with a more spherical distribution of electron density will remain indistinguishable for overlap scores across different alignment errors. However, contaminations can only be ranked by the **OS** score when relatively low levels of misalignment errors are present (Figure 5BC).

The **SFSC**, **gPC**, **FPC**, **FPCmw** and **gNSD** scores still outperformed all other scoring functions even at high SNR levels (Figure 4 and Figure 5). Gaussian filtering with $\sigma = 2$ for subtomograms at SNR = 0.1, improved the Spearman's correlation against contamination for many scores but only **cPC** and **oPC** showed performance above our cut-off of $\rho > 0.85$.

3.6 Assessment of biased angular distributions

It is possible that complexes have a preferred orientation in the sample, leading to non-uniform angular distributions of complexes with respect to the tomographic tilt axis. Such behavior can lead to undersampling of structure factors in missing wedge regions that are not sampled by any subtomograms (Supplementary Figure 11A). To study the effect of biased angular distributions, we simulated subtomograms of $GroEL_{14}$ and $GroEL_{14}/GroES_7$ complexes with orientations sampled from a biased, instead of a uniform, distribution—the orientations were restricted to rotations of maximum 10 degrees from the perfectly aligned position (Section 2.1.1).

Almost all scoring functions, except **cPC**, **gMI**, **oMI**, **NMI** and **OS**, performed very well for detecting misalignment errors across all contamination levels, despite the biased angular distributions of the samples (Supplementary Figure 11B). However, ranking contamination levels with biased angular distributions is more challenging for this particular complex. The most distinct differences between $GroEL_{14}$ and $GroEL_{14}/GroES_7$ structures (at the cap region) were affected most by the under sampled structure factors due to a preferred orientation of the missing-wedge region. Despite these challenges, **SFSC, gPC**, **gNSD**, **FPC**, **FPCmw** and **amCCC** performed well for ranking contamination levels, but only for clusters with relatively low misalignment errors (Supplementary Figure 11C).

3.7 Assessment for Experimental Subtomograms

We further assessed the scoring functions with experimental subtomograms, namely clusters of $GroEL_{14}/GroES_7$ (Förster et al., 2008), contaminated with $GroEL_{14}$ as well as clusters of ribosomes contaminated with mirror-imaged subtomograms of ribosomes (Section 2.1.2). As

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

it is challenging to know the exact SNR of experimental subtomograms, we estimated the *effective-SNR* of all subtomograms following a procedure previously described in (Xu et al., 2019). The aligned experimental subtomograms have an *effective-SNR* of ~0.11 for the GroEL dataset and ~0.0001 for the ribosomal dataset (Section 2.5, Eq. 30). Notably, the GroEL$_{14}$/GroES$_7$ complexes showed also a restricted angular distribution (i.e., a preferred orientation with respect to the tilt axes,) evident by the missing region in the averaged missing-wedge mask of aligned subtomograms (Supplementary Figure 11A).

Experimental GroEL dataset: All scoring functions, except **oNSD**, performed well for ranking misalignments in experimental subtomograms across all contamination levels, improved from performance seen in simulated subtomograms (at SNR = 0.1), which also showed a poor performance of **oNSD** (Figure 6A). The ranking of clusters based on contamination (across different levels of misalignments) was more challenging, again similar to the results observed for simulated subtomograms. **gPC**, **FPC**, **gMI**, **oMI** and **gNSD** ranked the contamination of clusters well, but only for clusters with relatively low misalignment errors and failed with increasing levels of misalignments. This behavior was similar to the assessment of biased angular distributions using simulated subtomograms in Section 3.6.

Ribosomal dataset: We observed that **SFSC**, **cMI**, **cNSD**, **oNSD** and **DSD** rank clusters well against misalignment (Figure 6B). However, when considering cluster contamination, **SFSC** was the only score that could robustly rank clusters in increasing order of contamination, but even then, only for clusters with lower misalignment errors (Figure 6D). We also tested an additional ‘cluster average-masked constrained cross-correlation’ (**amCCC**) scoring function. **amCCC** performed below required threshold against both misalignment and contamination.

3.8 Assessment with respect to cluster Size

We also tested performance of scoring functions for different cluster sizes. Tests were performed for simulated subtomograms (complexes 1FNT and 3DY4 at misalignment = 12) at two different contamination error levels (p=0 and p=10). With increasing cluster size **SFSC** scores increase (i.e., indicating better quality) for clusters at the same error level. **SFSC** is based on the spectral signal-to-noise ratio, which improves with a larger number of subtomograms at the same error level. All other scoring functions, based on pairwise subtomogram comparisons, maintain score values with increasing cluster size at similar error levels (except for small variations due to random sampling of subtomogram pairs) (Supplementary Figure 12). At comparable error level, a cluster with a larger number of particles will naturally produce an average density map with a higher resolution. And so, for comparable cluster quality, the larger clusters should preferentially be selected. We also note that for some scoring functions (**gPC**, **FPC**, **FPCmw** and **gNSD**) score values for clusters at lower contamination level are always higher than those for clusters with higher contamination, irrespective of cluster sizes (Supplementary Figure 12).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3.9 Time Complexity

The time complexity of scoring functions varies based on the type of computations required. Fourier Space-based scores (**SFSC**, **FPC** and **FPCmw**) require computing the Fourier Transform of each subtomogram, whereas all the mutual information variants need to bin the voxel densities first. The time complexity reported here is calculated on a single-core machine with all 500 subtomograms and 500 corresponding missing-wedge regions loaded in the memory. I/O operations are, therefore, not included in the time complexity measurements. Gaussian filtering, or any other preprocessing step, increases the time complexity further. Table 4 shows the time required to compute each score, without Gaussian filtering, for a cluster with 500 subtomograms. **SFSC** shows the best computational efficiency and is computationally more efficient by orders of magnitudes compared to almost all other scores. Also, **SFSC** scales linearly with the cluster size and all other scores scale quadratically, which makes **SFSC** computationally favorable. **gPC** and **gNSD** are linearly proportional to one another and, without Gaussian filtering, produce similar performance in ranking the cluster quality. Because **gNSD** takes only one-fifth of the time required by **gPC**, **gNSD** is a better choice than **gPC** for increasing computational efficiency. Calculations of all scores are parallelizable on multi-core machines, including the **SFSC** score (Xu et al., 2019).

4. Discussions

We compared more than fifteen scoring functions to test their ability to rank the quality of subtomogram clusters, which can vary in the amount of misalignment errors—non-optimal alignments of subtomograms—and contaminations errors—false assignments of non-target complexes to the subtomogram clusters. Such clusters can readily be generated, by unsupervised clustering methods, from tomograms, containing a heterogeneous set of complexes. A scoring function, if applicable as accurate similarity metric, will facilitate such clustering efforts. An accurate, well-performing, scoring function should be able to rank or distinguish clusters in their order of quality, according to the amount of misalignment and contamination errors. Here, we assessed a variety of scoring functions for their ability to act as accurate similarity metrics, using simulated subtomograms and experimental datasets over a wide range of SNR levels.

Overall, we observe a large variation in the performance of scoring functions. **Spectral SNR-based Fourier Shell Correlation (SFSC)** showed the best performance to rank alignment as well as contamination errors across all conditions without the need for subtomogram preprocessing. **Pearson correlation in Fourier space with missing wedge constraint (FPCmw)** was also a robust scoring function performing well across all datasets. The **FPCmw** score is similar to the constrained cross-correlation score (**CCC**) (a commonly used score in many alignment methods), except that it is directly calculated in Fourier Space, and therefore saves considerable computation time, because it does not need to compute the inverse Fourier Transform thousand to millions of times for each cluster evaluation. It is important to note, that both **SFSC** and **FPCmw** utilize missing wedge information (Section 2.4.1 and 2.4.6), which gives these scores an advantage by constricting their computation to only valid, rather than unreliable, frequency regions. **SFSC** showed several other

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

advantages. Its computation was the fastest, in some cases by several orders of magnitudes, among all the scoring functions. **SFSC** is calculated from all subtomograms in the cluster and does not require, unlike all other scoring functions, computation of scores for randomly selected pairs of subtomograms. Therefore, the **SFSC** score is free from potential biases in cluster quality assessment, introduced by the limited sampling of randomly selected subtomogram pairs in large clusters. Moreover, **SFSC** performs well for subtomograms with low SNR levels, even without Gaussian blurring of subtomogram.

Pearson correlation scores that are based on contoured and overlap voxel regions (**cPC**, **oPC**) fail for subtomograms at low SNR levels, in particular for ranking clusters based on contaminations when larger levels of misalignments are present. Preprocessing with Gaussian filtering (Gaussian low-pass filtering) can improve their performance, but not to a sufficient level for robustly ranking these clusters.

Scores based on mutual information are highly sensitive to SNR levels and fail to rank clusters at low SNR levels. Among all mutual information-based scores, only the **overlap Mutual Information (oMI)** performs above our threshold ($\rho > 0.95$) for ranking misalignments, but only at the highest SNR level of SNR = 0.1 or after Gaussian filtering with $\sigma = 2$ for simulated subtomograms with lower SNR. In previous studies (Joseph et al., 2017; Vasishtan and Topf, 2011), mutual information-based scores performed much better when applied to the ranking of atomic structures fitted into density maps from cryo-electron microscopy. This is because the 3D volumes used in previous studies have density values concentrated on the target complex regions, i.e., almost no noise component in the 3D-EM volumes. The probability of density values, used by the mutual information scores, are only accurately reproduced without high noise levels. When high noise levels and missing wedge effects are present, mutual information scores are only accurate when the score calculation is restricted to the overlap regions, which ensures density values are considered only from voxels of the target complex. This is only reliably possible either at high SNR level or after Gaussian filtering. **NMI** was able to rank the experimental subtomograms only for subtomograms with relatively high SNR levels.

We also conclude that preprocessing of subtomograms with Gaussian low-pass filters improves the performance of some scoring functions that depend on contoured and overlap voxel regions and decreases the performance of scores like **SFSC** that are dependent on the global variation of voxel intensities. Applying Gaussian filters to all subtomograms adds further to the time complexity. Moreover, scoring functions like **oPC** perform well only in a certain window of Gaussian filtering, which introduces uncertainty in determining the optimal **σ** value when performing a quality assessment of subtomogram clusters. Using scores that perform well without Gaussian filtering seems to be a better, more robust, choice.

Conclusion

Scoring functions, as metrics of subtomogram similarities, are of fundamental importance to assess the quality of subtomogram alignments, and subsequently relevant to maximize the accuracy of subtomogram classifications and averaging to determine structures of macromolecular structures. Here, we perform a comprehensive analysis of strengths and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

weaknesses for a series of relevant scoring functions, commonly used as image similarity metrics in the image analysis field. This comparison will be highly relevant in the field of subtomogram averaging and complex detection from heterogeneous samples. Our analysis narrows down a set of scoring functions for accurate detection of cluster quality. These scoring functions vary in their performance, depending on the specific context and goal of the problem statement. We believe the analysis done in this paper, will help users to choose a relevant function for their problems, as we move towards unsupervised methods in cryo-electron tomography. Overall, the **SFSC** and **FPCmw** scores have the most robust performance to assess the quality of subtomogram clusters over a large range of conditions. Accurate assessment of clusters opens up new, currently unexplored, avenues of cluster optimization, for instance, through ensemble methods, which leverage multiple complementary alignment methods to identify clusters of highest quality.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank High Performance Computing cluster resources available at Quantitative and Computational Biology department at University of Southern California and Institute for Digital Research and Education at University of California, Los Angeles, so we could run massive number of computations in parallel over 500 CPU cores. We thank Martin Beck for providing the original tomogram simulation code. We thank Prof. Friedrich Förster and Marten Chaillet for providing experimental ribosome and GroEL dataset. We also thank Dr. Stephen Carter, Alister Burt and Benjamin A. Himes for useful discussions. We thank the reviewers for useful comments and suggestions. This work was supported by NIH R01GM096089, Arnold and Mabel Beckman Foundation (BYI), NSF career 1150287 to F.A. F.A. is Pew Innovation Fund Investigator, supported by The Pew Charitable Trusts.

References

Beck M, Baumeister W, 2016. Cryo-Electron Tomography: Can it Reveal the Molecular Sociology of Cells in Atomic Detail? Trends Cell Biol. 10.1016/j.tcb.2016.08.006

Beck M, Malmström JA, Lange V, Schmidt A, Deutsch EW, Aebersold R, 2009. Visual proteomics of the human pathogen Leptospira interrogans. Nat. Methods 6, 817–823. 10.1038/nmeth.1390 [PubMed: 19838170]

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242. 10.1093/nar/28.1.235 [PubMed: 10592235]

Böhm J, Frangakis AS, Hegerl R, Nickell S, Typke D, Baumeister W, 2000. Toward detecting and identifying macromolecules in a cellular context: Template matching applied to electron tomograms. Proc. Natl. Acad. Sci. U. S. A 97, 14245–14250. 10.1073/pnas.230282097 [PubMed: 11087814]

Castaño-Díez D, Kudryashev M, Arheit M, Stahlberg H, 2012. Dynamo: A flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments. J. Struct. Biol 178, 139–151. 10.1016/j.jsb.2011.12.017 [PubMed: 22245546]

Che C, Lin R, Zeng X, Elmaaroufi K, Galeotti J, Xu M, 2018. Improved deep learning-based macromolecules structure classification from electron cryo-tomograms, in: Machine Vision and Applications. pp. 1227–1236. 10.1007/s00138-018-0949-4 [PubMed: 31511756]

Chen M, Dai W, Sun SY, Jonasch D, He CY, Schmid MF, Chiu W, Ludtke SJ, 2017. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. Nat. Methods 14, 983–985. 10.1038/nmeth.4405 [PubMed: 28846087]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Förster F, Pruggnaller S, Seybert A, Frangakis AS, 2008. Classification of cryo-electron sub-tomograms using constrained correlation. J. Struct. Biol 161, 276–286. 10.1016/j.jsb.2007.07.006 [PubMed: 17720536]

Frank J, Al-Ali L, 1975. Signal-to-noise ratio of electron micrographs obtained by cross correlation. Nature 256, 376–379. 10.1038/256376a0 [PubMed: 1095934]

Frazier Z, Xu M, Alber F, 2017. TomoMiner and TomoMinerCloud: A Software Platform for Large-Scale Subtomogram Structural Analysis. Structure 25, 951–961.e2. 10.1016/j.str.2017.04.016 [PubMed: 28552576]

Heumann JM, Hoenger A, Mastronarde DN, 2011. Clustering and variance maps for cryo-electron tomography using wedge-masked differences. J. Struct. Biol 175, 288–299. 10.1016/j.jsb.2011.05.011 [PubMed: 21616153]

Himes BA, Zhang P, 2018. emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging. Nat. Methods 15, 955–961. 10.1038/s41592-018-0167-z [PubMed: 30349041]

Hrabe T, Chen Y, Pfeffer S, Kuhn Cuellar L, Mangold AV, Förster F, 2012. PyTom: A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. J. Struct. Biol 178, 177–188. 10.1016/j.jsb.2011.12.003 [PubMed: 22193517]

Joseph AP, Lagerstedt I, Patwardhan A, Topf M, Winn M, 2017. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. J. Struct. Biol 199, 12–26. 10.1016/j.jsb.2017.05.007 [PubMed: 28552721]

Khoshouei M, Pfeffer S, Baumeister W, Förster F, Danev R, 2017. Subtomogram analysis using the Volta phase plate. J. Struct. Biol 197, 94–101. 10.1016/j.jsb.2016.05.009 [PubMed: 27235783]

Lebbink MN, Geerts WJC, van der Krift TP, Bouwhuis M, Hertzberger LO, Verkleij AJ, Koster AJ, 2007. Template matching as a tool for annotation of tomograms of stained biological structures. J. Struct. Biol 158, 327–335. 10.1016/j.jsb.2006.12.001 [PubMed: 17270464]

Martinez-Sanchez A, Kochovski Z, Laugks U, Meyer zum Alten Borgloh J, Chakraborty S, Pfeffer S, Baumeister W, Lučić V, 2020. Template-free detection and classification of membrane-bound complexes in cryo-electron tomograms. Nat. Methods 10.1038/s41592-019-0675-5

Nickell S, Förster F, Linaroudis A, Del Net W, Beck F, Hegerl R, Baumeister W, Plitzko JM, 2005. TOM software toolbox: Acquisition and analysis for electron tomography. J. Struct. Biol 149, 227–234. 10.1016/j.jsb.2004.10.006 [PubMed: 15721576]

Oikonomou CM, Jensen GJ, 2017. Cellular electron cryotomography: Toward structural biology in situ. Annu. Rev. Biochem 10.1146/annurev-biochem-061516-044741

Pei L, Xu M, Frazier Z, Alber F, 2016. Simulating cryo electron tomograms of crowded cell cytoplasm for assessment of automated particle picking. BMC Bioinformatics 17. 10.1186/s12859-016-1283-3

Scheres SHW, Melero R, Valle M, Carazo JM, 2009. Averaging of Electron Subtomograms and Random Conical Tilt Reconstructions through Likelihood Optimization. Structure 17, 1563–1572. 10.1016/j.str.2009.10.009 [PubMed: 20004160]

Schur FK, 2019. Toward high-resolution in situ structural biology with cryo-electron tomography and subtomogram averaging. Curr. Opin. Struct. Biol 10.1016/j.sbi.2019.03.018

Shatsky M, Hall RJ, Brenner SE, Glaeser RM, 2009. A method for the alignment of heterogeneous macromolecules from electron microscopy. J. Struct. Biol 166, 67–78. 10.1016/j.jsb.2008.12.008 [PubMed: 19166941]

Studholme C, Hill DLG, Hawkes DJ, 1999. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognit 32, 71–86. 10.1016/S0031-3203(98)00091-0

Sturges HA, 1926. The Choice of a Class Interval. J. Am. Stat. Assoc 10.1080/01621459.1926.10502161

Vasishtan D, Topf M, 2011. Scoring functions for cryoEM density fitting. J. Struct. Biol 174, 333–343. 10.1016/j.jsb.2011.01.012 [PubMed: 21296161]

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli A, Pietro, Rothberg A,Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik DV, Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold GL, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso JV, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO, Vázquez-Baeza Y, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. 10.1038/s41592-019-0686-2 [PubMed: 32015543]

Wriggers W, Milligan RA, McCammon JA, 1999. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. J. Struct. Biol 125, 185–195. 10.1006/jsbi.1998.4080 [PubMed: 10222274]

Xu M, Alber F, 2012. High precision alignment of cryo-electron subtomograms through gradient-based parallel optimization. BMC Syst. Biol 6. 10.1186/1752-0509-6-S1-S18

Xu M, Beck M, Alber F, 2012. High-throughput subtomogram alignment and classification by Fourier space constrained fast volumetric matching. J. Struct. Biol 178, 152–164. 10.1016/j.jsb.2012.02.014 [PubMed: 22420977]

Xu M, Beck M, Alber F, 2011. Template-free detection of macromolecular complexes in cryo electron tomograms. Bioinformatics 27. 10.1093/bioinformatics/btr207

Xu M, Singla J, Tocheva EI, Chang YW, Stevens RC, Jensen GJ, Alber F, 2019. De Novo Structural Pattern Mining in Cellular Electron Cryotomograms. Structure 27, 679–691.e14. 10.1016/j.str.2019.01.005 [PubMed: 30744995]

Yu Z, Frangakis AS, 2011. Classification of electron sub-tomograms with neural networks and its application to template-matching. J. Struct. Biol 174, 494–504. 10.1016/j.jsb.2011.02.009 [PubMed: 21382496]

Zhang P, 2019. Advances in cryo-electron tomography and subtomogram averaging and classification. Curr. Opin. Struct. Biol 10.1016/j.sbi.2019.05.021

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Highlights

- Scoring functions to assess the quality of CryoET 3D subtomogram clusters
- Assessment for alignment errors and contamination from other complexes in a cluster
- SFSC is most robust to rank clusters based on alignment and contamination errors

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

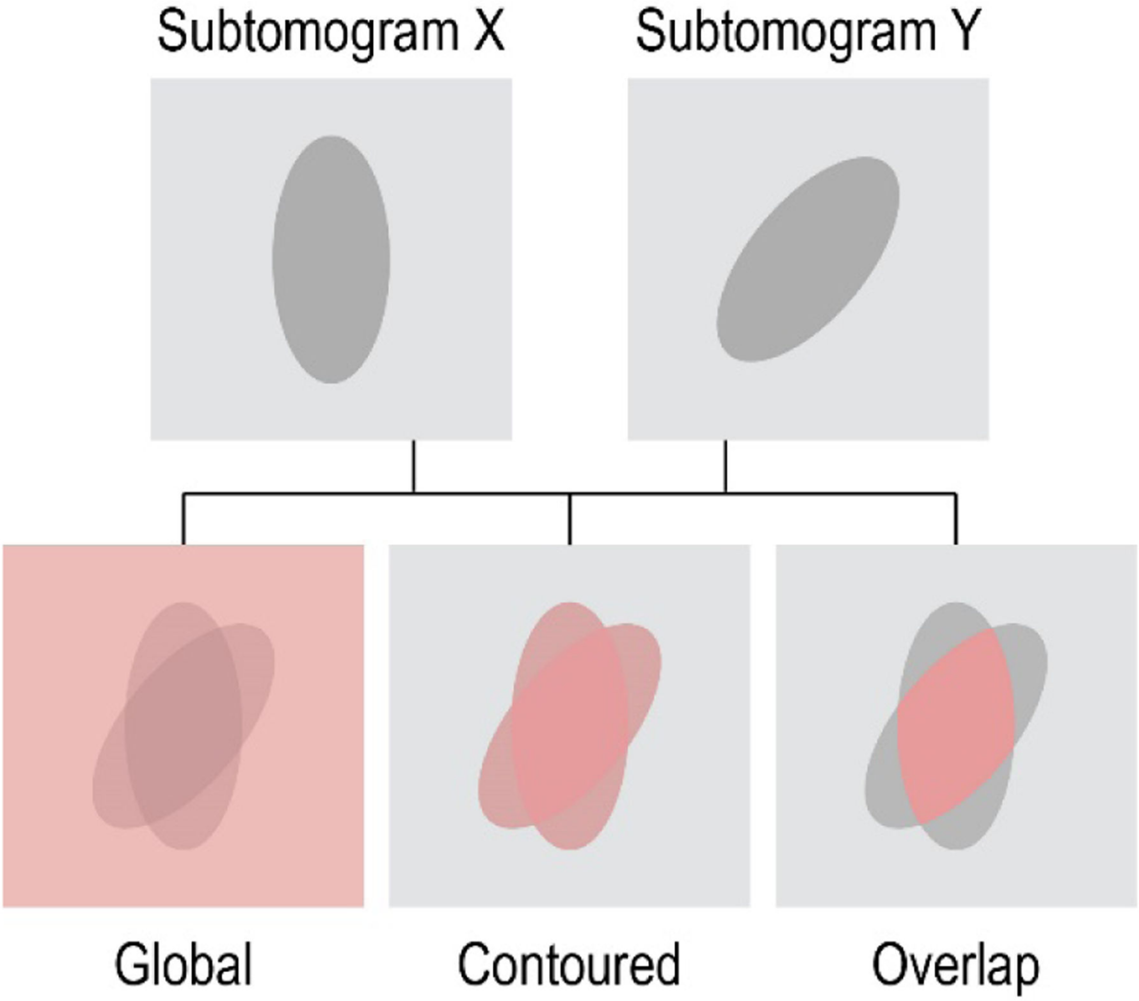


Figure 1: Voxel regions.

Schematic representation of global, contoured and overlap regions (highlighted in red) used for computing scores between two subtomograms.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

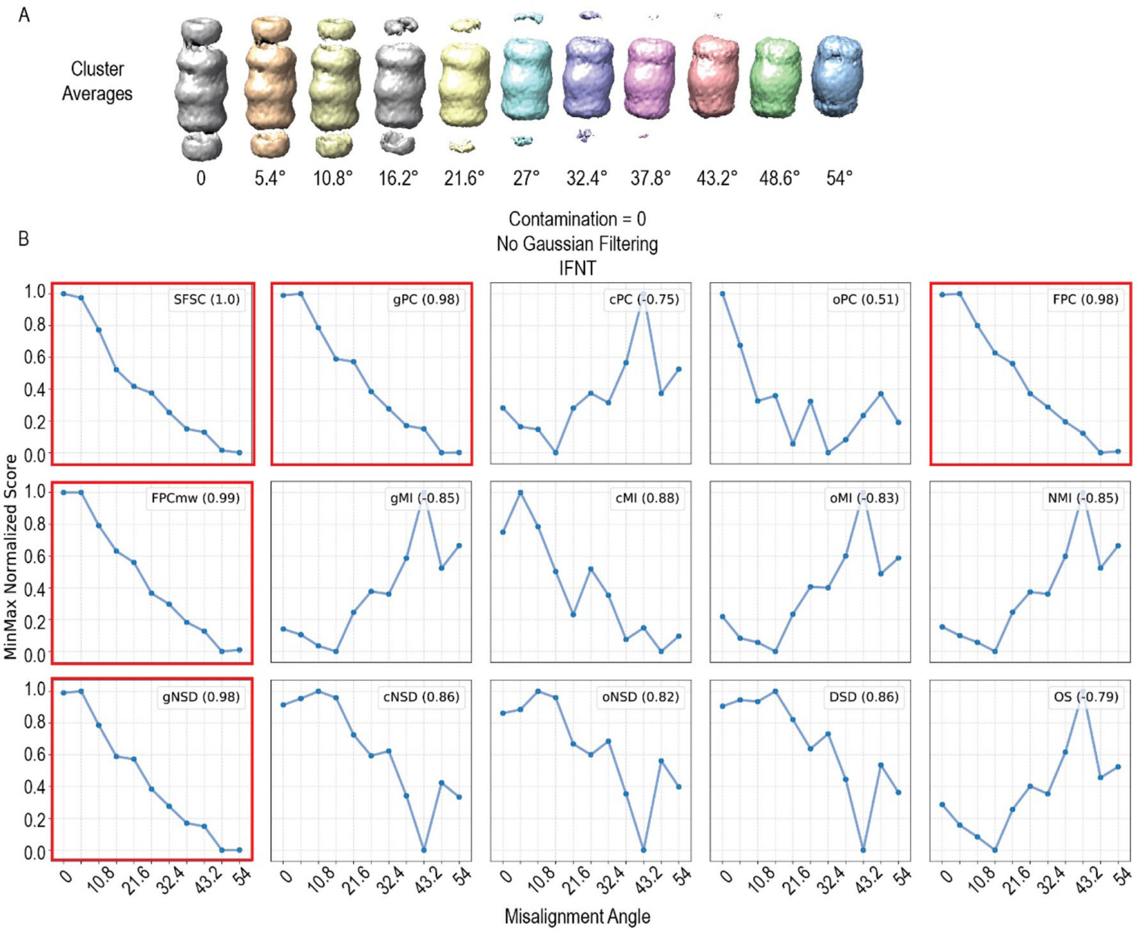


Figure 2: Assessment against misalignment for example complex 1FNT.

A) Averages of subtomogram clusters containing complex 1FNT. Clusters have zero contamination and vary in misalignment error increasing from 0 (Far left) to 54 degrees (Far right). **B)** Line plots showing min-max normalized score values on y-axis varying with misalignment on x-axis for clusters constituting 1FNT subtomograms. Legend in each subplot mentions the scoring function and its performance in Spearman’s correlation to rank clusters based on misalignment. Scores that have Spearman’s correlation above the cutoff of 0.95 are outlined with red line.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

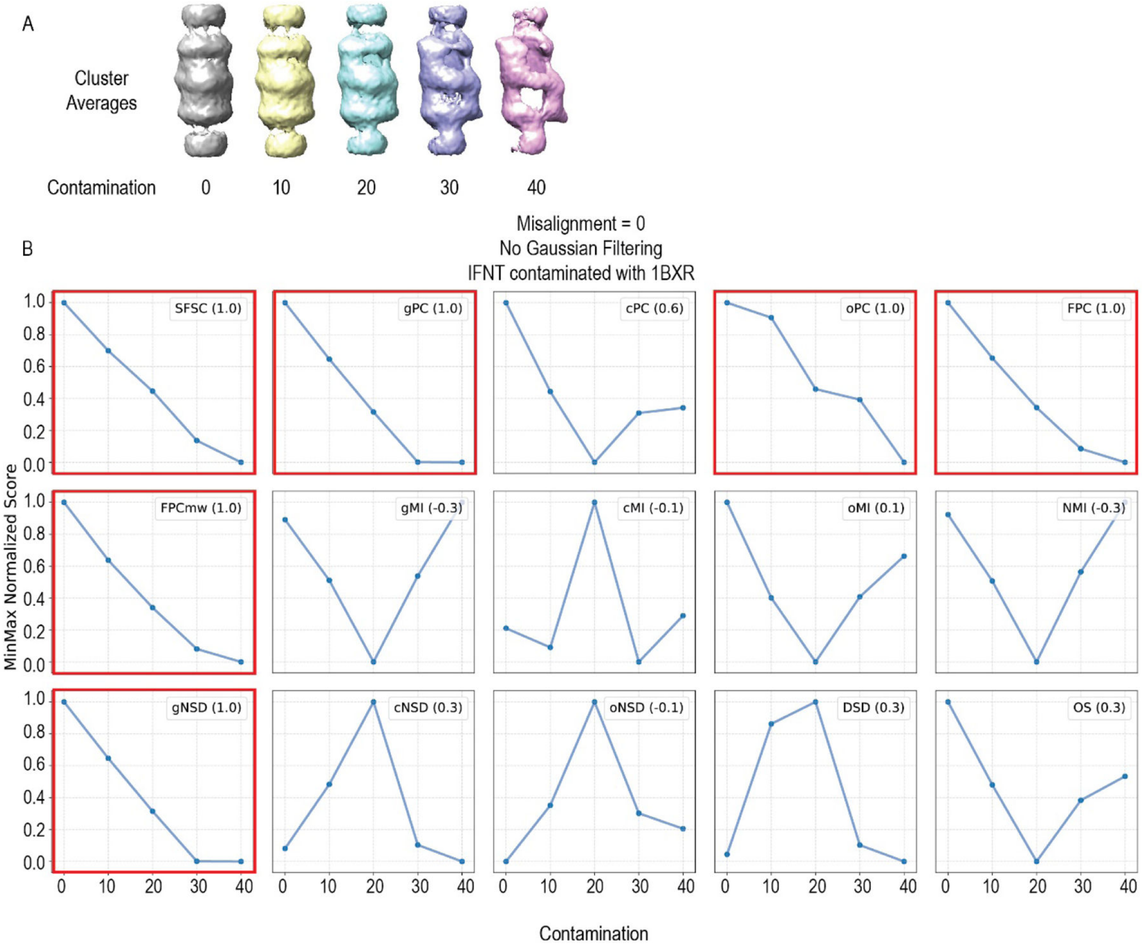


Figure 3: Assessment for cluster contamination for example complex 1FNT.

A) Averages of subtomogram clusters containing complex 1FNT with contamination of complex 1BXR. Clusters have no misalignment error but vary in contamination percentage, $p \in [0, 40]$, with contamination levels increasing from 0 (Far left) to 40% (Far right). **B)** Line plots shows min-max normalized score values on y-axis varying with contamination on x-axis for clusters constituting target complex (PDB ID: 1FNT) and contaminated with contaminant complex (PDB ID: 1BXR). Legend in each subplot mentions the scoring function and its performance in Spearman's correlation to rank clusters based on contamination. Because we have only five sample points to compute ρ, we lower the threshold and select those functions as well-performing that have ρ > 0.85. Scores that have Spearman's correlation above the cutoff of 0.85 are outlined with red line.

Author Manuscript

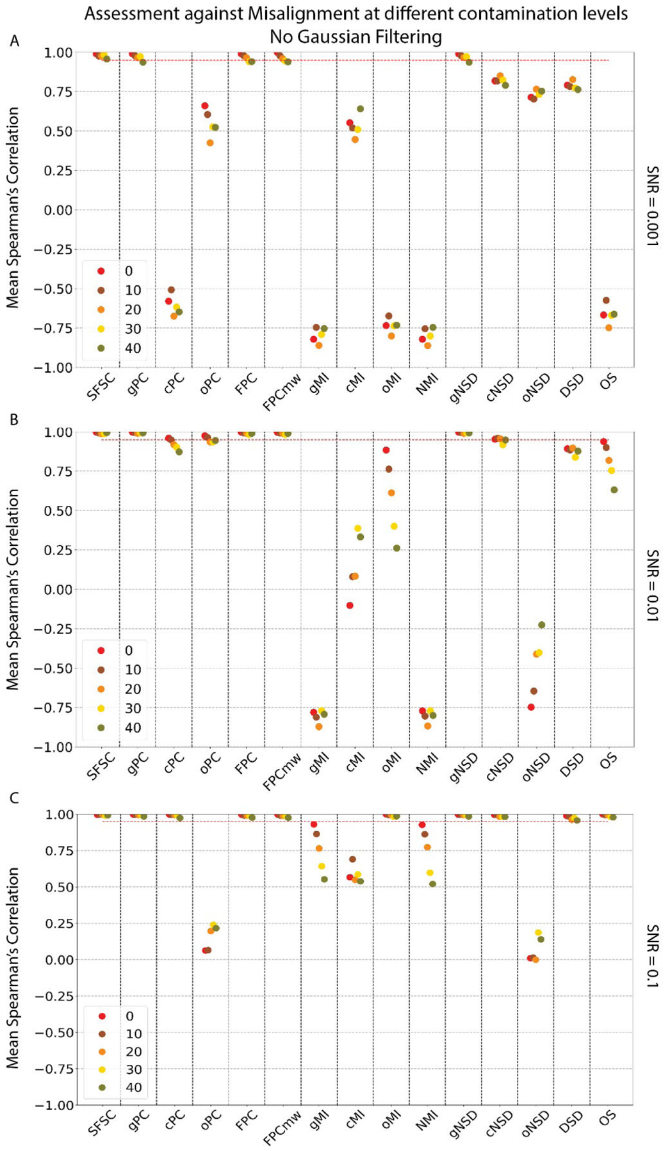


Figure 4: Assessment against Misalignment at different contamination levels.

Spearman's correlation ρ (y-axis) of scoring functions (x-axis) on simulated subtomograms without Gaussian filtering. Each panel is a scatter plot of Spearman's correlation (ρ) of scoring functions vs. misalignment for clusters at different contamination levels, i.e., $C_{p=0}^{m \in [0,54]}$, $C_{p=10}^{m \in [0,54]} \ldots C_{p=40}^{m \in [0,54]}$ (Section 2.2, Eq. 3). Clusters generated with the target complex can be contaminated with other complexes (Table 1, Section 2.2.2). So, each point is average ρ across all the ten target-contaminant complex pairs, except for contamination = 0, where it is averaged over only five target complexes. Red dashed line shows a cutoff value of 0.95. Subtomograms simulated at different SNR levels are shown in separate panels: **(A)** SNR = 0.001 **(B)** SNR=0.01 **(C)** SNR = 0.1.

Author Manuscript

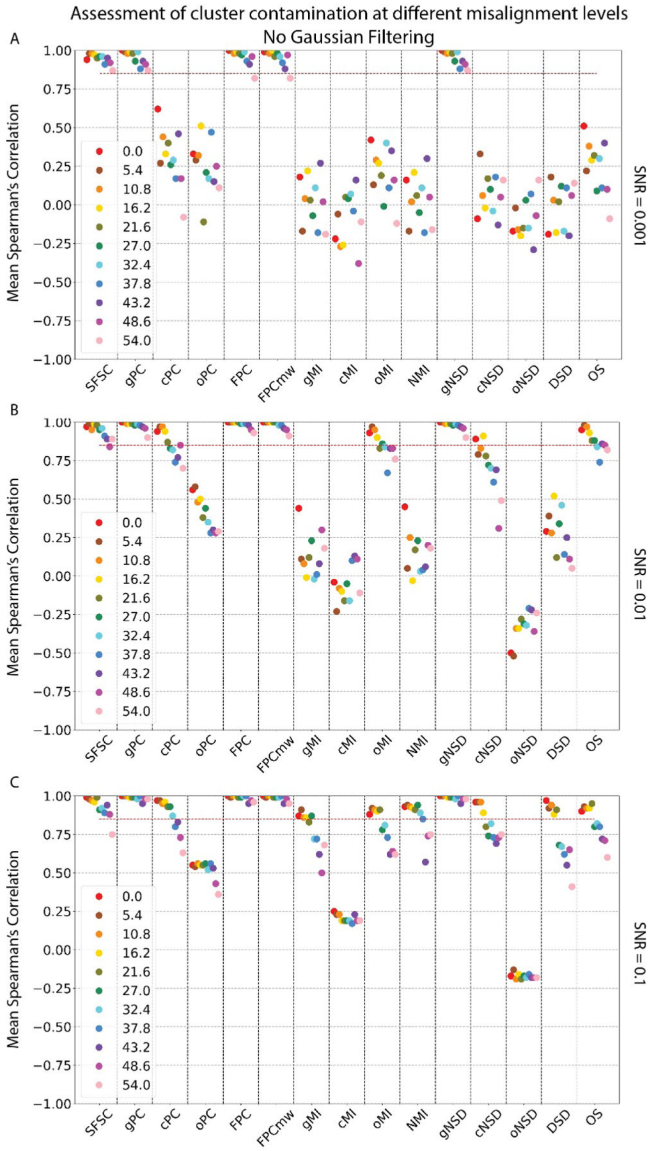


Figure 5: Assessment of cluster Contamination at different misalignment levels.

Spearman's correlation ρ (y-axis) of scoring functions (x-axis) on simulated subtomograms without Gaussian filtering. Each panel is a scatter plot of Spearman's correlation (ρ) of scoring functions vs. contamination for clusters with different misalignment errors, i.e., $C_{p \in [0, 40]}^{m = 0}, C_{p \in [0, 40]}^{m = 5.4} \dots C_{p \in [0, 40]}^{m = 54}$ (Section 2.2, Eq. 3). Clusters generated with the target complex and contaminated with other complexes can still have a varying amount of misalignment within the subtomograms. Each point is average ρ across all the ten target-contaminant complex pairs. Red dashed line shows a cutoff value of 0.85. Subtomograms simulated at different SNR levels are shown in separate panels: **(A)** SNR = 0.001 **(B)** SNR=0.01 **(C)** SNR = 0.1.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

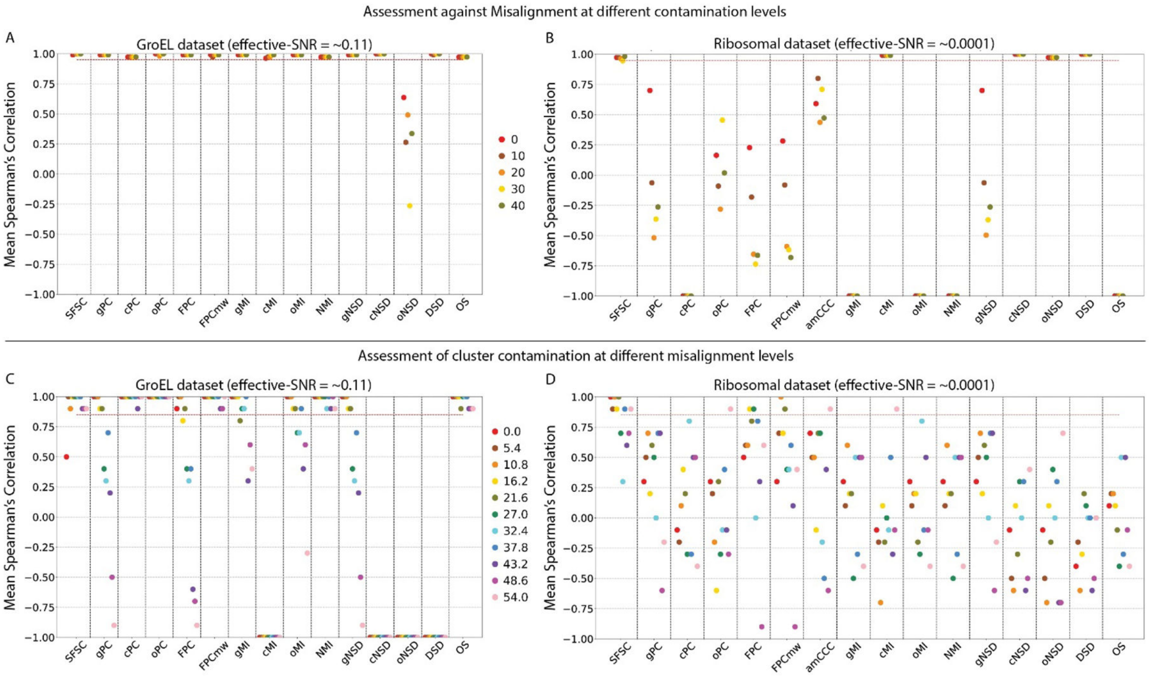


Figure 6: Assessment on experimental subtomograms:

Scatter plot showing Spearman's correlation (ρ) (y-axis) of scoring functions (x-axis) on experimental subtomograms without Gaussian filtering. Each scatter point is a ρ value. **(A, B)** Spearman's correlation of Scoring functions vs. Misalignment at different contamination levels. Red dashed line shows threshold value 0.95 (A: GroEL, B: Ribosomal). **(C, D)** Spearman's correlation of Scoring functions vs. Contamination at different misalignment levels. Red dashed line shows threshold value 0.85 (C: GroEL, D: Ribosomal). $GroEL_{14}$/ $GroES_7$, subtomogram clusters were contaminated with $GroEL_{14}$ subtomograms and ribosome subtomogram clusters were contaminated with mirrored subtomograms of ribosomes.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Table 1:
Complexes for simulated studies.

PDB IDs of complexes used to generate clusters. First column shows PDB IDs of the target complex in the cluster and second and third column contains PDB ID of complexes with which target complex is contaminated with. The structures of these PDB IDs are shown in Supplementary Figure 2.

Target complex PDB ID	Contaminant complex PDB IDs	
1F1B	2BO9	1A1S
1FNT	1BXR	3DY4
2GHO	1QO1	2H12
2GLS	1KP8	1VPX
2REC	1VPX	1VRG

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2:

Scoring Functions.

Acronyms of all the scoring functions and their variations based on voxel regions used for computing scores. Scoring functions marked with * are only discussed in section 3.4, 3.6 and 3.7.

Scoring Function	Global	Contoured	Overlap	Average-masked	Significant Voxels
Spectral SNR-based Fourier Shell Correlation	SFSC (Xu et al., 2019)				
Pearson Correlation	gPC FPC FPCmw (Xu and Alber, 2012)	cPC	oPC	amPC*	
Mutual Information	gMI NMI (Joseph et al., 2017)	cMI	oMI (Joseph et al., 2017; Shatsky et al., 2009)	amMI*	
Squared Deviation	gNSD	cNSD	oNSD	amNSD*	DSD (Joseph et al., 2017)
Overlap Score			OS		
Constrained cross-correlation	CCC* (Castaño-Díez et al., 2012; Himes and Zhang, 2018; Hrabe et al., 2012)			amCCC* (Castaño-Díez et al., 2012; Himes and Zhang, 2018; Hrabe et al., 2012)	

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3:

Effect of Gaussian Filtering:

Column 2–4: Spearman’s correlation (ρ) of Scoring functions vs. Misalignment for homogeneous clusters (i.e., contamination = 0). Column 5–7: Spearman’s correlation (ρ) of Scoring functions vs. Contamination for perfectly aligned clusters (i.e., misalignment = 0). ρ values are averaged over all the 10 target-contaminant complex pairs (Table 1). Cells with bold text shows average ρ values that are above the cut-off, average ρ > 0.95 against misalignment and average ρ > 0.85 against contamination. All values of ρ are rounded to 2 decimal places. Subtomograms were simulated at SNR = 0.001 and Gaussian filtered with σ = 1 and 2.

Scoring Functions	**Against Misalignment**			**Against Contamination**		
	No Gaussian filtering	**Gaussian kernel σ = 1**	**Gaussian kernel σ = 2**	**No Gaussian filtering**	**Gaussian kernel σ = 1**	**Gaussian kernel σ = 2**
SFSC	**0.99**	**0.98**	0.55	**0.94**	**0.92**	0.39
gPC	**0.99**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**
cPC	−0.58	**0.99**	**0.99**	0.62	**0.97**	**0.93**
oPC	0.66	**0.95**	0.07	0.33	0.50	0.38
FPC	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
FPCmw	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**
gMI	−0.82	−0.77	0.76	0.18	0.39	0.78
cMI	0.55	−0.03	0.01	−0.22	−0.46	0.11
oMI	−0.73	**0.95**	**1.00**	0.42	**0.97**	**0.92**
NMI	−0.82	−0.75	0.72	0.16	0.44	**0.86**
gNSD	**0.99**	**0.95**	**0.99**	**1.00**	0.84	**0.95**
cNSD	0.82	**0.97**	**0.98**	−0.09	0.76	**0.87**
oNSD	0.71	−0.79	−0.18	−0.17	−0.47	−0.10
DSD	0.79	0.91	**0.97**	−0.19	0.39	**0.86**
OS	−0.67	**0.97**	**1.00**	0.51	**0.97**	**0.93**

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4:

Time complexity:

Time required to compute score values on cluster size of 500 subtomograms and with subtomogram and mask of size 91^3 voxels. Time was computed without Gaussian filtering, on single core computer and with all the files already loaded in the memory.

Score	Time (in seconds)
SFSC	28.70
gPC	261.84
cPC	497.96
oPC	426.82
FPC	1047.05
FPCmw	985.62
gMI	1353.70
cMI	1584.96
oMI	1689.41
NMI	1369.82
gNSD	52.91
cNSD	476.84
oNSD	408.29
DSD	13190.03
OS	380.24