Template-free detection and classification of heterogeneous membrane-bound complexes in cryo-electron tomograms

Antonio Martinez-Sanchez*, Zdravko Kochovski, Ulrike Laugks, Johannes Meyer zum Alten Borgloh, Stefan Pfeffer, Wolfgang Baumeister and Vladan Lucic*

Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
*Corresponding authors: AM-S martinez@biochem.mpg.de and VL vladan@biochem.mpg.de

Abstract

With faithful sample preservation and direct imaging of fully hydrated biological material, cryo-electron tomography (cryo-ET) provides an accurate representation of heterogeneous cellular constituents in their native environment. However, detection and precise localization of complexes within a cellular environment is aggravated by their large number and heterogeneous nature. We developed a template-free image processing procedure that allows an accurate tracing of complex networks of biological densities, as well as a comprehensive and automated detection, and an unsupervised classification of heterogeneous membrane-bound complexes in cryo-electron tomograms. Applying this procedure to tomograms of membrane vesicles from rough endoplasmic reticulum (ER), we detected and classified small protein complexes like the ER protein translocons, which were not detected by other methods. This classification provided sufficiently homogeneous particle sets for further processing by the currently available subtomogram averaging methods. Furthermore, we present structural evidence that different ribosome-free translocon species are present at the ER membrane, determine their 3D structures, and show that they have different localization patterns and form nanodomains.

Introduction

The cellular environment is characterized by a large number of heterogeneous molecular components. Stable and transient complexes form basic units that perform distinct steps comprising cellular functions. Many biochemical cascades like signaling and protein synthesis depend on the composition and the precise location of the relevant complexes.

In cryo-electron tomography (cryo-ET), biological samples are faithfully preserved by rapid freezing, which prevents water crystallization and rearrangements of the biological material. Importantly, samples are imaged in transmission electron microscopy in the same vitrified, fully hydrated state [1]. Therefore, cryo-ET is uniquely suited for high resolution, direct three-dimensional (3D) imaging of fully hydrated, unperturbed cellular complexes within their native environment [2, 3].

The potential of cryo-ET to yield a cellular map of molecular complexes is hampered by the noisy and heterogeneous cellular environment. Because visual detection is limited to large complexes of characteristic shapes [4], image processing methods gained prominence. In template matching a high resolution structure of a protein or complex of interest is used to computationally search for similar structures in a tomogram or on projection images [5, 6, 7, 8, 9]. This approach is particularly suited for complexes that are not embedded in larger assemblies and critically depends on the existence of higher resolution 3D structures of complexes large enough to be detected in cellular cryo-tomograms. Similarly, automated methods were developed for segmentation of membranous structures [10, 11]. In a different approach, pleomorphic complexes are detected by an automated procedure, based on their membrane attachment [12]. Their molecular identification is complicated because it requires a functional characterization of the detected complexes [13] and may involve an array of different genetic

Algorithm 1 Complete procedure.

1. Density tracing and particle picking
   - Tracing of biological densities by the Discrete Morse theory based algorithm (DisPerSe)
   - Simplification by topological persistence
   - Spatially embedded graph representations of the biological density
   - Selection of complexes - particle picking
2. General classification
   - Determination of membrane normal vectors
   - Constrained refinement (Relion)
   - Unsupervised classification of rotationally averaged complexes by Affinity propagation
3. Spatial analysis and averaging
   - Standard 3D classification and constrained refinement (Relion)
   - Spatial distribution analysis within or between classes

conditions, particularly for higher eukaryotic systems [14]. Subtomogram averaging can yield 3D structures at high resolution, but requires biological systems containing a high number of homogeneous proteins or complexes of interest [15, 16, 17].

To allow a more comprehensive processing of complex cryo-tomograms, we developed a software procedure for template-free detection and classification of heterogeneous membrane-bound molecular complexes imaged by cryo-ET of cellular systems. It is based on methods from other fields that we adapted and further developed, such as the discrete Morse theory based segmentation, affinity propagation and spatial point processes, and includes custom-made software. The classes obtained are sufficiently homogeneous to allow further processing by standard subtomogram averaging methods [18, 17]. Validations were performed on both phantom and real datasets.

Results

Procedure overview

Our procedure consists of three major parts (Algorithm 1). First, complexes are detected in a comprehensive, template-free manner. Then, they are classified into classes containing structurally similar complexes, rendering them suitable for further processing. Finally, the spatial distribution of complexes and their average densities are determined.

Density tracing and simplification

Detection of complexes in a tomogram, that is the determination of their location, is analogous to particle picking in single particle analysis and subtomogram averaging. However, our goal is to detect heterogeneous complexes in a comprehensive manner.

For an automated tracing of the biological material density in cryo-tomograms, we adapted DisPerSE, a software package based on the discrete Morse theory, which was developed for the identification of astrophysical structures in 3D images of the large-scale matter distribution in the Universe [19]. In general, Morse theory is used to calculate topological indices (invariants) of a given manifold, while the discrete Morse theory is applied to simplicial complexes [20, 21].

Here, we used tomogram greyscale values to define a Morse function. Using DisPerSE software, the critical points of the Morse function (where its gradient is 0), some higher dimensional manifolds and their inter-relations were determined. Together, these form Morse complexes. In order to allow tracking of biological density in cryo-electron tomograms, we selected the following manifolds for further processing: (i) Minimum points (termed 0-critical points), (ii) Saddle points that have minima

Algorithm 2 Simplification by topological persistence

For each pair of connected minima and saddle points (p_i, s_a) whose values differ by less than a specified persistence value:

1. Find the other minimum (p_k) connected to the saddle point s_a and connect it to all saddle points connected to p_i
2. Remove p_i, $s_a$and all their arcs
3. Add ascending manifold associated with minimum p_i to the one associated with minimum p_k
4. Remove arcs associated with saddle points of low density

in two and a maximum in one direction (1-critical points), (iii) Arcs that connect minima and saddle points defined as maximum gradient curves between these points (descending 1-manifolds), (iv) 3-manifolds associated with minima (ascending 3-manifolds). To present the detection in a more intuitive way, we processed a (2D) tomographic slice of a neuronal synapse (Figure 1 A, C). The minima and the arcs visually corresponded well to the distribution of the bio-material.

A high level of noise present in cryo-tomograms causes the detection of many closely spaced local minima that have only slightly smaller values than their neighborhood, resulting in an overly complex structure of the calculated Morse complexes. Because of a highly complex network of densities present in cellular cryo-tomograms, we could not use the Morse complex simplification procedure implemented in DisPerSE. To solve this problem, we implemented a modified version of the simplification by topological persistence [19]. In essence, pairs of a minimum and its connected saddle point that had similar grayscale values were removed, and the affected Morse complex elements were reassigned. (Algorithm 2, Figure 2). Alternatively, one can consider this method as introducing small changes in the greyscale values so that some pairs of minima and saddle points disappear, effectively reducing the contribution of noise.

The simplification by topological persistence resulted in a greatly simplified Morse complex, and a faithful tracing of biological material by minima, saddle points and the connecting arcs (Figure 1D). These form a skeleton that provides an intuitive representation of the biological density and its constituents, protein complexes. All together, the choice of manifolds provided by the discrete Morse theory, combined with the custom-made implementation of the simplification procedure, made it possible to accurately trace biological density.

Graph embedding and detection of complexes (particle picking)

We implemented a graph representation of Morse complexes, where minima are assigned to graph vertices and edges to saddle points (Figure 1B). Because a saddle point is connected to two minima, the edge corresponding to the saddle point connects the corresponding vertices. Vertices keep the information about their spatial location, while edges contain the full information about the underlying arcs, thus forming a spatially embedded graph. Additionally, greyscale values of minima are associated with vertices, while greyscale values of saddle points, euclidean distances between connected minima and geodesic length of arcs are associated with edges.

These graphs may also contain external information provided by segmentation of large cellular structures, such as lipid membranes, organelles or cytoskeleton. The Morse complexes and their corresponding graphs represent the distribution of the biological material visualized in a tomogram. The spatially embedded graphs occupy the central part in the software we developed, because they combine precise geometrical, topological and biological information and allow computationally efficient queries that can extract specific information used to detect individual complexes (particles). For example, the yellow and the orange star-bound paths in Figure 1D represents an extracellular presynaptic membrane-bound and a transcleft complex, respectively, while the brown path shows a complex composed of transcleft and pre- and postsynaptic intracellular components. Therefore, subgraphs can be selected starting from vertices belonging to a previously defined membrane or another cellular structure and extending by a specified length. Greyscale values and geometrical information associated with vertices and edges can be used as further constraints. Subgraphs containing few vertices and

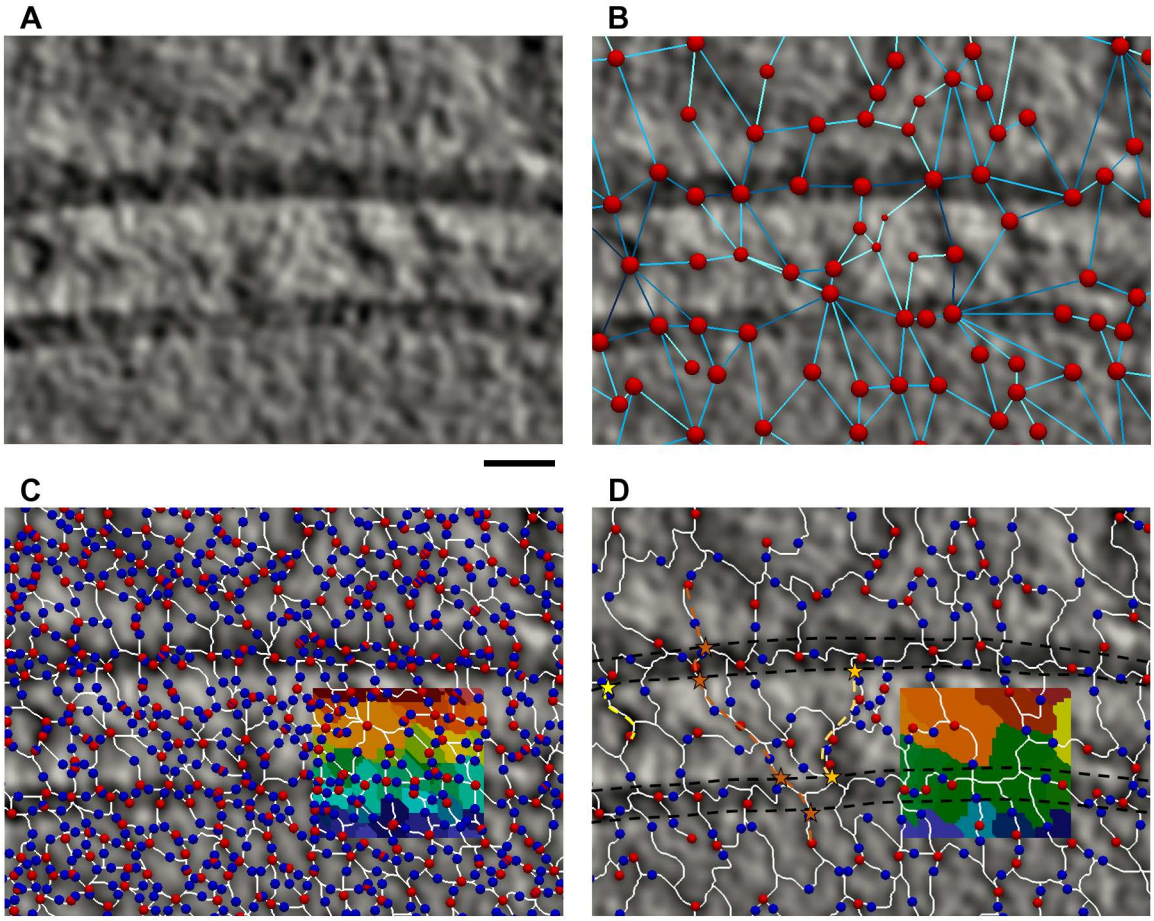

Figure 1: Tracing biological material at a synapse. A Tomographic slice of 1.37 nm thickness. B Graph representation of the Morse complex shown in D. C Morse complex obtained by application of DisPerSE on the slice shown in A. D Morse complex obtained from C after simplification by topological persistence. B-D Images superimposed on the slice from A. C, D Color insets show ascending 2-manifolds, labeled by different colors. Red circles represent greyscale minima (graph vertices, larger size denotes minima of higher density) and the blue ones the saddle points. White lines are arcs and blue lines graph edges (darker shade denote saddle points of higher density). Yellow, orange and brown paths represent possible extracellular presynaptic, trans-cleft and extended trans-cleft complexes. Stars are intersection points between the selected paths and membrane faces. Black dashed lines outline synaptic membranes. Scale bar 10 nm.

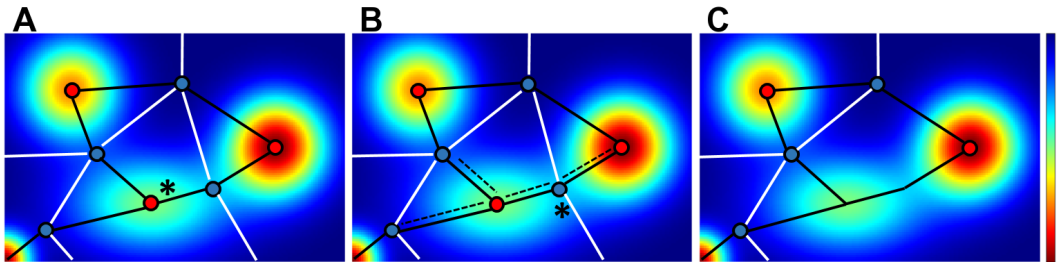

Figure 2: Simplification by topological persistence in 2D. A The initial Morse complex is shown on the density map. The minimum to be removed by the simplification is labeled by an asterisk. B Arcs that are to be modified are indicated by dashed lines and the saddle point that is to be removed is indicated by an asterisk. C Morse complex after the simplification. On all panels minima are shown by red and saddle points as blue points, arcs are shown as black lines and the white lines denote borders between ascending 2-manifolds.

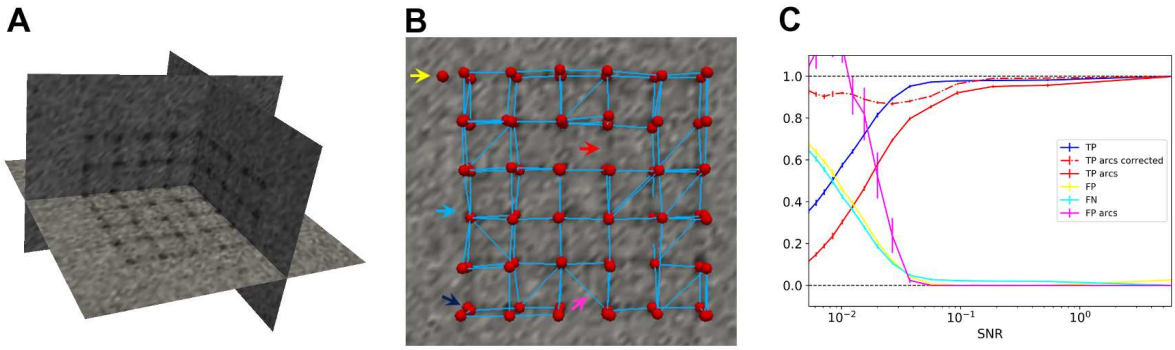

Figure 3: Validation of the density tracing and the simplification on phantom data. A Tomographic slices of the phantom data set. B Detected minima and arcs at SNR = 0.025 in the graph representation superposed on a slice of the simulated tomogram. Red spheres denote detected minima (vertices) and blue lines detected arcs (edges). The yellow arrow points to a FP minimum, blue to a FN minimum, dark blue to a double minimum, pink to a FP arc and red to a FN arc. C Normalized number of TP, FP and FN minima (labeled as TP, FP and FN), and TP and FP for arcs (mean ± std, N=10 simulations per SNR).

edges may represent molecular complexes and define their positions. Hence, the procedure described so far corresponds to particle picking in the single particle analysis.

Detection of density in phantom data

To validate the density tracing and the simplification procedures, we created a phantom dataset comprising a rectangular grid having higher density at the intersections, and added variable amounts of Gaussian noise (Figure 3A). Densities in all phantom datasets were detected by applying the discrete Morse theory and the topological simplification as outlined above. Grid intersections and grid bars were taken as the ground truth features for the detection of minima and arcs, respectively. We detected the minima and arcs that matched the ground truth (true positives, TP), did not match the ground truth (false positives, FP), as well as the unmatched ground truth features (false negatives, FN) (Figure 3B). We did not consider multiple minima occurring at the same grid intersection, because these are eliminated by imposing an exclusion distance between particles during particle picking.

Numbers of TPs, FPs and TNs were normalized to the total number of the corresponding ground truth features. For SNR above 0.05, the FPs and FNs were below 10% and TP minima was above 90%, however for SNR between 0.05 and 0.1 TP arcs was between 80% and 90% (Figure 3C). To a large extent, this failure to detect some of the ground truth arcs (these constitute FN arcs) was caused by the minima that were not detected (FN minima). This was confirmed by normalizing TP arcs to the total number of ground truth arcs that could be formed given the detected minima (TP arcs corrected in Figure 3C).

General classification

The application of the procedure described in the previous section is expected to yield a set of membrane-bound complexes possessing high compositional and conformational heterogeneity. Therefore, it was essential to develop a general classification procedure capable of separating highly heterogeneous complexes into groups (classes) of similar complexes.

Particle (complex) positions were used to generate particle subtomograms and calculate the direction of vectors perpendicular to the membrane. These normal vectors specify two of the angles that determine the orientation of the particles, while the third angle is left undetermined. To optimize particle positions and normal vectors, we performed particle refinement where during alignment the angles defined by the normal vectors were allowed only small changes around the initial values, while the third angle was not constrained (here termed constrained refinement). In addition, a high symmetry was imposed on the third angle (around the normals), diminishing its importance for the

alignment. To reduce the influence of the missing wedge, the initial model for this refinement was obtained by randomizing the third angle and averaging all particles (without alignment).

We classified particles using the affinity propagation clustering [22], whereby nodes (particles) exchange information between each other to reach the optimal partitioning. Compared to the standard clustering methods, this algorithm has the advantage that it is unsupervised, the number of classes is not specified in advance but obtained from the clustering and the algorithm can handle cases where classes have a very different number of particles.

The success of a clustering procedure critically depends on the manner the clustering distance (similarity) is defined. Here, we represented particles as 2D images obtained by computing rotational averages (of particle subtomograms) around their normal vectors and defined the distance between two particles as the dot product of their rotational averages. In this way, the 2D averages used for clustering were aligned to each other, and there were no degrees of freedom that could hinder the clustering. This is in contrast to clustering based on the 3D particle subtomograms, where the angle around the normal vector is not known.

Detection and classification on microsomal membranes

For validation of the particle picking and general classification methods, we used a subset (26%) of previously analyzed cryo-ET data depicting canine pancreatic microsomes [23]. This work established the basic architecture of the translocon complex and structure of its constituents: the Sec61 protein-conducting channel, the translocon-associated protein complex (TRAP) and the Oligosaccharyltransferase complex (OST) [24, 25].

Biological material was detected using our procedure, as explained above (Figure 4A, Video 1). The persistence simplification allowed tomogram normalization, by setting the persistence threshold to obtain a fixed density of greyscale minima on the microsomal membranes on all tomograms. Particles at the cytoplasmic and lumenal faces of the ER membrane were picked independently of each other, based on minima that satisfied certain geometrical constraints (see the Methods).

The cytosolic and lumenal particle positions and the membrane normal vectors were optimized by the constrained refinement using C10 symmetrization. A well-positioned density and a resolved lipid bilayer obtained by the refinement (Figure 4B) argue that particle positions and membrane normals were determined precisely.

Classification of cytosolic particles by affinity propagation yielded more than 100 classes (Figure S2A). Constrained refinement of these classes, using internal initial references, showed different species of ribosome-translocon complexes (Figure 4C, D). Constrained refinement of the best class that contained both cytosolic and lumenal density was used to generate an initial reference for further processing. All cytosolic particles were subjected to three rounds of 3D classification. The first round was applied to each class separately, to remove suboptimal particles. The other two 3D classification rounds focused on the lumenal segments and the small ribosomal subunit, respectively, resulting in structures comparable to those previously reported (Figure 4E) [24, 23]. These included fully assembled ribosomes bound to the fully assembled and partial translocon complexes, resolved to 18 Å and 22 Å, respectively, as well as ribosomal large subunits resolved to 21 Å (Figure S2C). This confirms that sufficiently homogeneous particle sets were generated by our procedure, which could be further processed by standard external reference-free subtomogram classification and averaging.

Alternatively, the first 3D classification was performed on all particles together (termed "bulk cleaning", as opposed to the above "AP cleaning" variant). Bulk cleaning variant yielded similar 3D averages. However, using the AP cleaning variant, TRAP was better resolved in the partial translocon complex class, the number of particles was increased and the resolution obtained was slightly higher (Figure S3). In addition to providing an internal initial reference, the affinity propagation classification thus contains information that can be exploited by subsequent processing.

The best of over 100 affinity propagation classes of lumenal particles (Figure S2B) was refined to yield a 3D density of the translocon. The bulk cleaning variant, using the translocon density as the reference, yielded a well resolved ribosome-translocon class and classes representing two different ribosome-free translocon states (Figure 4F). Among the particles that contained a defined lumenal density, 15% had an associated ribosome and thus corresponded to the ribosome-translocon complex. Among the ribosome-free complexes, 68% corresponded to fully assembled translocon (TRAP, OST and Sec61) and 17% likely represented individual OST complexes. The resolution was determined to

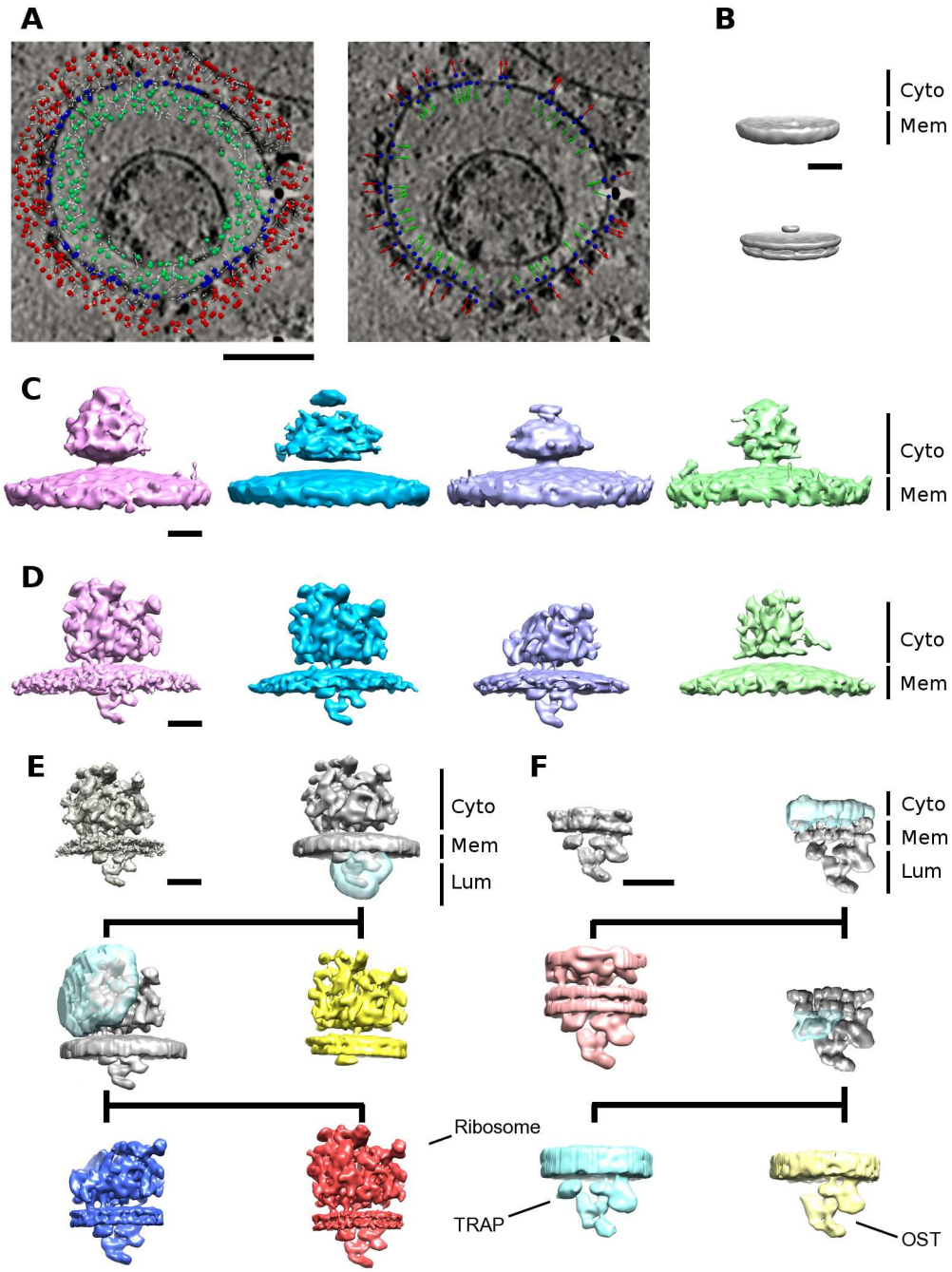

Figure 4: Processing of ER membrane-associated complexes. A Tracing of biological density. Density minima are shown as small spheres (left) and particles as spheres with arrows representing the associated membrane normal vectors (right). Position of the spheres is color-coded: Red cytosol; blue membrane; green lumen. B Average of all cytosolic particles obtained without alignment (above) and with constrained refinement with C10 symmetry (below). C, 3D class averages of the representative affinity propagation classes of cytosolic particles (ribosomes), obtained without alignment. D Refinement of the same classes shown in (C) obtained using the averages shown in (C) as initial models. E 3D classification and refinement of cytosolic particles (ribosomes). Densities of ribosomes bound to the fully assembled (red) and partial translocon complex (yellow), as well as the large ribosomal subunit bound to the fully assembled translocon (blue) are shown. F 3D classification and refinement of lumenal particles (translocon). Densities of the ribosome-translocon complex (light red), the ribosome-free fully assembled translocon (light blue) and the non-translocon associated individual OST complex (light yellow) are shown. In both (E) and (F) initial references obtained from the affinity propagation classes (top row, left) and densities obtained by the first classification of all particles (top row, right) are shown together with the classes obtained by the second (middle row) and the third round of classification (bottom row). Transparent blue regions correspond to the masks used for classification. Refined and post-processed densities are shown in color. 'Cyto', 'Mem' and 'Lum' denote cytosolic, membrane and lumenal regions, respectively. Scale bars A 100 nm, B-F 10 nm.

22 Å, 14 Å and 16 Å, respectively (Figure S2D). The same 3D classification procedure failed without the initial translocon density.

Therefore, our procedure is capable of picking small complexes, like the translocon or even smaller individual OST complexes (≈200 kDa lumenal mass). The unsupervised classification by affinity propagation was instrumental to carry the processing to a level where the standard 3D classification and refinement procedures could be used.

Spatial distribution analysis

Methods available for the analysis of spatial point processes can provide further information about the biological system of interest and assist with classification. Specifically, monovariate distribution functions analyze the distribution and clustering of particles within a class. Among these, the first order functions are based on the distance to the closest point, either from other points (nearest neighbor distribution) or from arbitrary locations (spherical contact function), or both (J-function) [26]. A more detailed description is obtained by Ripley's second order functions, which evaluate the distribution at different length scales, by considering distances between all pairs of points [27, 28]. Furthermore, bivariate versions of the nearest neighbor and Ripley's functions characterize colocalization and co-clustering of particles between two classes.

To assess the statistical significance, the above functions are evaluated with respect to the random distribution (null hypothesis). Due to the restricted and irregular shape of the region where the particles are located, analytical models cannot be used. Instead, many random point distributions need to be generated within the particle region. To this end, we implemented a Monte-Carlo method that generates random distributions of a specified number of particles within an arbitrary space (see Methods).

Spatial organization of microsomal complexes

As an example of using spatial point distribution methods to address biological questions, we investigated the spatial organization of the microsomal particle classes obtained above. Upon visual inspection, some classes showed distinct distributions (Figure 5A, Video 2). In order to quantitatively determine whether the complexes were clustered and at which length scales, we calculated the univariate Ripley's function for the particle classes and compared them with results obtained for simulated particles.

The ribosome-free translocon complex showed a significant clustering at length scales from about 8 nm to more than 50 nm, while the non-translocon associated OST was borderline significant (at p=0.05 level) at 10-20 nm in respect to the random distribution (Figure 5B,C). We confirmed these results by using the first order functions: the nearest neighborhood, spherical contact distribution and J-functions all showed significant clustering distribution of the ribosome-free translocon complexes (Figure S4).

As expected, the distribution of ribosomes, comprising all three final classes of the cytosolic particles (Figure 4E), also showed significant clustering (Figure 5D) likely induced by polyribosome formation. In addition, using the bivariate Ripley's L function, we did not detected a significant colocalization between ribosomes and ribosome-free translocon complexes (Figure 5E, F). These examples show that spatial point process methods allow a quantitative characterization of the organization of molecular complexes.

Discussion

Cellular cryo-electron tomograms contain large amounts of unexplored information due to the heterogeneity and complex spatial organization of their molecular constituents. To solve this problem, we used and modified existing computational methods and developed new software to create a template-free procedure that can classify pleomorphic membrane-bound molecular complexes, so that the resulting classes are sufficiently homogeneous for further processing by the currently available subtomogram averaging methods.

In order to trace densities in biological cryo-electron tomograms, it was necessary to adapt a discrete Morse theory-based procedure and modify topological persistence simplification [19]. Applications on

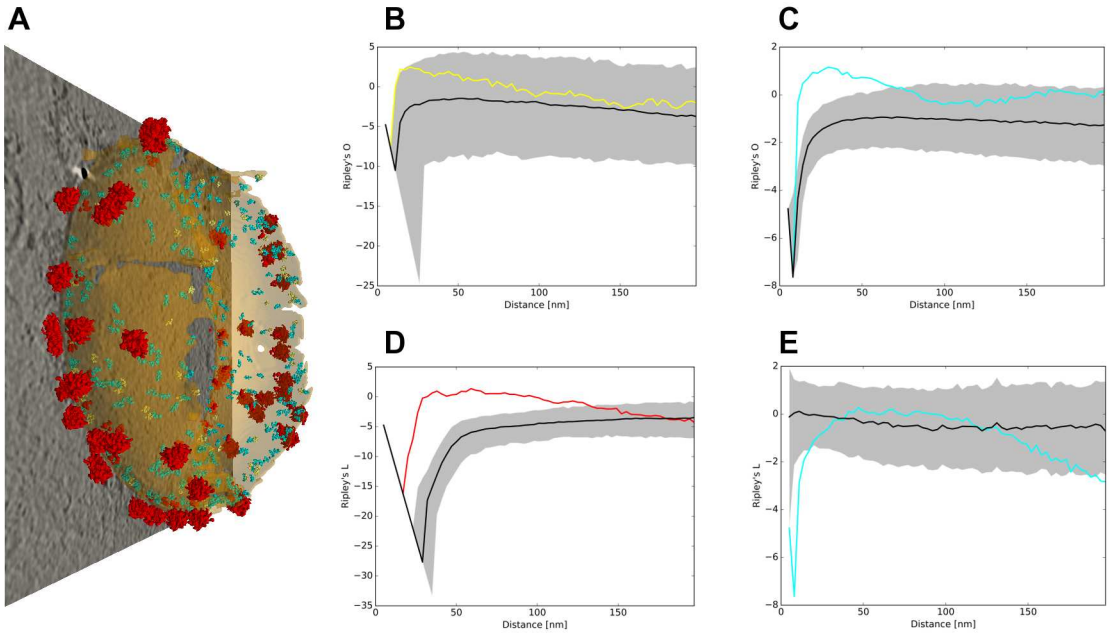

Figure 5: Spatial distribution of microsomal particles. A Distribution of the particles is shown on one microsome (all three cytosolic ribosome classes red, ribosome-free full translocon: cyan, non-translocon associated OST: yellow). B-D Univariate Ripley's L function of the three classes, the same color code as in A. E Bivariate Ripley's L function between the cytosolic ribosome classes and the ribosome-free translocon complex. In B-E Black lines show the median of the Ripley's L function for a set of random particle distributions (≈1200) and the gray areas represent regions of p>0.05 confidence.

both phantom and biological data showed accurate tracing of densities at low signal-to-noise conditions. We developed software to convert the tracing data into spatially embedded graphs that allowed us to use geometrical, connectivity, and external information to extract complexes (pick particles) and determine their orientation in respect to local membranes. This rather comprehensive particle picking yields highly heterogeneous particles, thus making their structural classification a challenging task. We found that 2D particle rotation averages around an axis perpendicular to the membrane provide a representation that allows an efficient, unsupervised classification by affinity propagation [22]. Together, these steps provide a template-free procedure that can accurately trace complex networks of biological densities, and localize and classify heterogeneous membrane-bound molecular complexes.

The application of our template-free procedure on a previously reported dataset depicting microsomes [23] resulted in a direct, ribosome-independent detection of translocons, small ER membrane-resident complexes with domains projecting into the lumenal side of microsomes (≈260 kDa total lumenal mass). We obtained 3D densities of ribosome-free translocon and even smaller individual OST (≈200 kDa lumenal mass) complexes, which were not previously detected by template-matching. Focusing on the cytosolic side, we detected 3D structures of ribosomes associated with different translocon species, consistent with previous template-matching ribosome localization approaches [24, 23]. In both cases, the unsupervised classification by affinity propagation was instrumental to generate initial references and particle sets that were sufficiently homogeneous for the subsequent standard 3D classification and refinement procedures.

These results demonstrate that the procedure presented here can be applied to small complexes that were beyond the reach of template-matching. Furthermore, being template-free, our procedure can be applied to localize heterogeneous complexes, or complexes in different compositional and conformational states. Obviously, a template-free approach is not limited by the availability of higher resolution structures and, unlike template-based approaches, does not introduce an initial bias that might affect subtomogram classification and averaging.

The rotational averaging of membrane-bound complexes eliminates the only unknown rotational orientation, thus removing the problem of incorrect alignment that can hinder classification. Never-

theless, it keeps sufficient information available for classification, in the form of 2D rotational averages. The combination of these may be the reason why we successfully applied the affinity propagation classification on large datasets (more than 60 000 particles). The previous template-free approaches, based on 3D rotation-invariant properties, automated pattern mining or difference of Gaussians picking were successful only on large complexes, and did not reach a resolution comparable to ours, neither on simulated nor on real datasets [29, 30, 31]. Because the ability to determine normal vectors is the only membrane-related requirement, our procedure can be applied to complexes attached to any cellular membrane and also larger structures such as the cytoskeleton.

The results obtained from the microsomal data provide structural proof for the presence of ribosome-free translocon complexes in the ER membrane that either await binding of ribosome-nascent chain complexes for co-translational protein transport and membrane insertion, or are engaged in post-translational processes. Notably, the majority of these ribosome-free translocon complexes already comprise all constituents known to be present in the ribosome-associated translocon, arguing against a step-wise assembly of the translocon complex on the ribosome. In metazoans, the STT3A type OST complex is stably integrated into the translocon complex for co-translational glycosylation of nascent proteins, while the STT3B type OST complex is excluded from the translocon and takes care of glycosylation sites skipped by STT3A [32, 33]. Thus, the individual, not translocon-associated OST complexes we localized in the ER membrane likely correspond to STT3B type OST complexes.

Finally, we implemented and adapted functions required to characterize spatial distribution of point-particles for application to spatial regions of arbitrary shape. These can be used to provide further information about the biological system under study and to assist particle processing. Using these functions, we observed a significant clustering of ribosome-free translocon complexes. This suggests the presence of nanodomains in the ER membrane for post-translational protein transport and membrane insertion, established by direct or indirect interactions between these complexes.

In conclusion, the procedure described here, based on the template-free detection and unsupervised classification of pleomorphic macromolecular complexes, extends the applicability of cryo-ET to small and heterogeneous membrane-bound molecular complexes and therefore makes possible a large-scale, non-invasive detection, localization and averaging of molecular complexes in-situ.

Methods

Synaptosomal preparation

Cerebrocortical synaptosomes were extracted from 6–8 week old male Wistar rats as described previously [34, 35, 36] in accordance with the procedures accepted by the Max Planck Institute for Biochemistry. In brief, anesthetized animals were sacrificed, and the cortex was extracted and homogenized in homogenization buffer (HB; 0.32 M sucrose, 50 mM EDTA, 20 mM DTT, and one tablet of Complete mini EDTA-free protease inhibitor cocktail (Roche; 10 ml, pH 7.4) with up to seven strokes at 700 rpm in a Teflon glass homogenizer. The homogenate was centrifuged for 2 min at 2000 g, and the pellet was resuspended in HB and centrifuged for another 2 min at 2 000 g. Supernatants from both centrifugations were combined and centrifuged for 12 min at 9500 g. The pellet was resuspended in HB and loaded onto a three-step Percoll gradient (3%, 10%, and 23%; Sigma-Aldrich) in HB without protease inhibitor cocktail. The gradients were spun for 6 min at 25 000 g, and the material accumulated at the 10/23% interface was recovered and diluted to a final volume of 100 ml in Hepes-buffered medium (HBM; 140 mM NaCl, 5 mM KCl, 5 mM NaHCO3, 1.2 mM Na2HPO4, 1 mM MgCl2, 10 mM glucose, and 10 mM Hepes, pH 7.4). Percoll was removed by an additional washing step with HBM by centrifugation for 10 min at 22 000 g, and the pellet was resuspended in HBM and immediately used in the experiments. All steps were performed at 4°C.

Cryo-ET of synaptosomes

For vitrification, a 3-µl drop of 10-nm colloidal gold (Sigma-Aldrich) was deposited on plasma-cleaned, holey carbon copper EM grids (Quantifoil) and allowed to dry. A 3-µl drop of synaptosomes was placed onto the grid, blotted with filter paper (GE Healthcare), and plunged into liquid ethane.

Tilt series were collected under a low dose acquisition scheme [37] on Titan Krios [FEI] equipped with a field emission gun operated at 300 kV, with a post-GIF energy filter (Gatan) operated in the

zero-loss mode and with a computerized cryostage designed to maintain the specimen temperature at <-150°C. Images were recorded on a direct electron detector device (K2 Summit operated in the counting mode). Tilt series were typically recorded from -60° to 60° with a 2° angular increment. Pixel sizes was 0.34 nm at the specimen level. Volta phase-plate with nominal defocus of -1 μm [38] was used. The total dose was kept <100 $e^-/Å^2$. Tilt series were aligned using gold beads as fiducial markers, and 3D reconstructions were obtained by weighted back projection (WBP) using Imod [39]. During reconstruction, the projections were binned once (final voxel size of 0.68 nm) and low pass filtered at the post-binning Nyquist frequency.

Computational methods

DisPerSE software package was used for the tracing of biological densities [19]. Tomogram greyscale values were used to define a Morse function, while the simplicial complex was defined as the 3D voxel-based Cartesian grid. The processing of a Morse function on a simplicial complex by DisPerSE yields manifolds that form a Morse complex and describe critical points and their inter-relations. Importantly, 1-critical points are always connected by arcs to two minima.

We modified the topological persistence simplification method and implemented it in PySeg package (Algorithm 2). The procedure first removes the pairs consisting of a minimum and a connected saddle point whose greyscale values differ by an amount smaller than a specified persistence threshold. Then, the arcs and ascending manifolds related to removed points are reassigned. Because this procedure may leave multiple arcs linking the same pair of minima, arcs associated with low-density saddle points are removed.

Creation of spatially embedded graphs from Morse complexes and their manipulation was embedded in PySeg. For some of the standard graph tasks, the graph-tool library was used [40]. Methods to query properties associated with graph vertices, and edges and methods to extract particles were also implemented in PySeg. All together, representing biological density by spatially embedded graphs significantly increased the computational efficiency.

Radial averaging of subtomograms around the membrane normal vectors and an interface for the scikit-learn [41] implementation of the affinity propagation algorithm [22] are provided in PySeg. We also implemented contrast limited adaptive histogram equalization in Pyseg.

The software developed for the work presented here (as mentioned above, together with the spatial distribution functions, below) was developed in the object-oriented manner in Python as PySeg package. The package contains installation and usage instructions, examples on real biological data and more than 66 000 lines (instructions and examples excluded). It is open-source and available upon demand.

PySeg uses Numpy package, surface meshes were stored using VTK [42] and graphs are plotted using *matplotlib* library [43]. We parallelized the most intensive operations in order to provide a software package able to effectively process big datasets with many particles and tomograms,

For visualization, Paraview [44] and the UCSF Chimera package from the Computer Graphics Laboratory, University of California, San Francisco (supported by NIH P41 RR-01081) [**?**] software packages were used.

All computations were done on Linux clusters at the computer center of the Max Planck Institute of Biochemistry.

Processing of phantom data

The phantom dataset contained a 6x6x3 grid with variable amount of Gaussian noise (SNR between 0.005 and 5). For each SNR, 10 datasets were generated. The size of intersections was 2x2x2 voxels and of grid bars 8x2x2 voxels. These datasets were processed in 3D by DisPerSE and simplified by topological persistence. The parameters were set using our standard procedure. That is, the persistence threshold was set so that the number of minima was 20% higher than the number of grid intersections. The low-density saddle points were removed to obtain 20% more arcs than grid bars, resulting in a higher ratio of arcs to minima (2.3) than the default (2.0), which better captured the high connectivity of the phantom grid. TPs, FPs and FNs were normalized to the total number of ground truth features (grid intersections and arcs). In order to remove the influence of the detection

of minima on arc detection, we also normalized the TP arcs to the corrected number of ground truth arcs, that is the number of arcs that could be formed given only the detected minima.

Processing of microsomal ribosomes

The processing work-flow is schematically shown in Figure S1.

55 out of 210 the tomograms previously obtained from canine pancreatic microsomes [23] were used in this study. From these tomograms (1.048 nm pixel size), 122 microsomal membranes were segmented using an automated procedure [45]. For tracing of densities, persistence simplification and particle picking, tomograms were smoothed by Gaussian filtering at σ= 2 / 0.8 pixels (for the cytoplasmic / lumenal sides). Density tracing and the initial topological simplification were performed in 3D by DisPerSE [19], and the resulting skeleton was simplified by the topological persistence procedure choosing the persistence threshold automatically in a way that the density of vertices (minima) on all microsomal membrane was 0.0025 nm^{-3} and the number of arcs was two times higher than the number of vertices.

To localize cytoplasmic / lumenal particles, we selected vertices that were located at 25-75 nm / 3-15 nm Euclidean and 25-50 nm / 3-30 nm geodesic distance (length of the shortest path composed of arcs) from the membrane and had up to 2.5 sinuosity (ratio of geodesic and euclidean distances). For each selected vertex, the closest membrane vertex was detected, and these membrane vertices were chosen to represent particles, resulting in 64 000 and 62 000 cytosolic and lumenal particles, respectively. A particle exclusion distance of 5 nm was imposed. These particle positions were used to reconstruct particles in Imod [39] at a pixel size 0.524 / 0.262 nm and with a box size 120 / 160 pixels (for the cytoplasmic / lumenal sides).

For classification by affinity propagation, particles were rotationally averaged by computing mean greyscale values of 2 pixel-wide rings around the particle normal vectors and the resulting rotational averages were normalized to a density mean of 0 and standard deviation of 1. Unless noted otherwise, the masks used here and for the following classification and refinement steps were cylindrical, for cytoplasmic particles 40x120 pixels (radius x height) containing both cytosolic and lumenal space, and for lumenal particles 25x110 pixels, containing lumenal and only little cytoplasmic space just above the membrane.

The other 3D classification steps, as well as all 3D particle refinement steps were performed in Relion [18]. During the refinement, particle half-datasets were processed independently according to the "gold-standard" procedure, as implemented in Relion. The resolution was determined by Fourier shell correlation at the FSC = 0.143 criterion. The constrained refinement was carried out by initially aligning particles according to the direction of their normal vectors. The alignment was then optimized by allowing small changes in the two normal vector angles and small spatial displacements. The alignment around the third angle (around the normal vector) was not constrained to explore the entire angular range, except when a high symmetry is used (typically C10). Specifically, we used the two angles defining particle normals to the membrane to set the prior values for angles tilt and psi in Relion particle star files and specified small values (3.66) for the standard deviations of these two angles as the refine command options. Unless noted otherwise, the initial reference was obtained by aligning all particles according to the two angles determined from normals and randomizing the third angle (around the normal direction), that is no external reference was used.

Further processing was done essentially in the same way for the cytoplasmic and lumenal particles, as follows. We obtained 2D averages of the affinity propagation classes (by averaging the rotational averages of individual particles belonging to a class) and visually inspected them to select the cytosolic class showing the best cytoplasmic and lumenal densities and the lumenal class showing the best lumenal density. These classes were averaged by constrained refinement to yield densities to be used as initial references for the further processing.

We then performed three rounds of 3D classification by Relion. False positive particles were removed by the first 3D classification round, with constrained alignment, using all cytosolic / lumenal particles and starting from the previously obtained initial references. In the bulk cleaning variant, all particles were classified together and the best class (resembling the initial reference the most) was selected for further processing (2600 cytosolic and 21 000 lumenal particles). In the AP cleaning variant, each affinity propagation class was subjected to 3D classification and the best classes were selected (7100 cytosolic particles from 15 affinity propagation classes). The second 3D classification

round was focused on the opposite sides of the particles (lumenal / cytosolic) and the third round on the smaller regions. Both classification steps were performed without alignment, using the alignment parameters from the first 3D classification round (the masks used are shown in Figure 4 E, F). Specifically, for the cytosolic particles, the second 3D classification was focused on the ER lumen and generated ribosome class bound to different translocon species (fully assembled translocon: 3400 particles; partially assembled: 1800 particles), while the third round of 3D classification focused on the cytosolic face of the ER membrane to separate translocon-bound ribosome (1064 particles) from translocon-bound large ribosomal subunits (873 particles) (Figure 4 E). For the lumenal particles, ribosome-bound (1800 particles) and the ribosome-free translocon (11 000 particles) classes were generated during the second 3D classification step focused on the cytosolic side, while the third 3D classification round focused on the ER lumen yielded two classes representing different ribosome-free states (separate OST complexes and full translocon complexes, 2200 and 8600 particles, respectively) (Figure 4F). An exclusion distance of 15 nm was imposed in order to remove overlapping particles, likely originating from translation shifts during the alignment steps. The final classes were averaged by the constrained refinement, post-processed and the FSC curves were generated.

Spatial distribution functions

We implemented the following spatial distribution functions. The nearest neighbor distribution function *G(r)* of a particle set is defined as a probability that the nearest neighbor of a particle is found at a distance $\leq r$. The spherical contact distribution *F(r)* is a probability that the closest particle from an arbitrary point is found at a distance $\leq r$. Consequently, *G(r)* primarily describes the organization within particle clusters, while *F(r)* the empty space. The J-function is a combination of the two:

$$J(r) = \frac{1 - F(r)}{1 - G(r)}$$

Ripley's L function is calculated considering the region within which the particles are detected (particle region), which can have an arbitrary shape, as follows:

$$L(r) = r \left(\sqrt{\frac{\sum_k N_k(r)}{\lambda \sum_k V_k(r)}} - 1 \right)$$

where λ is the overall concentration of particles, $N_k(r)$ is number of points (particles) within a distance $\leq r$ to point k, and $V_k(r)$ is the volume of the intersection of the particle region and a sphere of radius r centered on the point k [27, 28]. The bivariate versions of functions *G(r)* and *L(r)* characterize the colocalization of two particle sets. They differ from the univariate functions specified above in that the distances are calculated from particles of one set to the particles of the other set.

We calculated Ripley's L function for particles on each microsome and obtained the mean. For the determination of the statistical significance of Ripley's L function, we generated multiple random distributions (10 for each microsome, that is $\approx$1200 total) that had the same number of particles and were located within the same spatial region as the particle set. Random points were given real particle size and were not allowed to overlap among the same class, effectively imposing an exclusion distance within a class. The envelope within which 95% of the curves were located was then used to determine whether the distribution of the particle set was significantly different from the random distribution (at the $p<0.05$ significance level).

Acknowledgements

We would like to thank Florian Beck for useful discussions and Gabriela J. Greif for critical reading of the manuscript. A.M.-S. is the recipient of a postdoctoral fellowship from the Séneca Foundation.

Author Contributions

AM-S and VL conceived and designed the research. AM-S designed and implemented the software; ZK and UL acquired original tomograms; SP provided expertize related to the previously recorded tomograms; AM-S, JMzAB and VL analyzed the data; WB provided resources and acquired funding; VL supervised research; AM-S and VL wrote the manuscript. All authors edited the manuscript

Competing interests

The authors declare no competing interests.

Additional information

Accession codes: EM densities have been deposited in the EMDataBank obtained from cytosolic particles: ribosome bound to the fully assembled (EMD-0074) and partial translocon complex (EMD-0084), the large ribosomal subunit (EMD-0075), and obtained from lumenal particles: the ribosome-translocon complex (EMD-0085), the ribosome-free fully assembled translocon (EMD-0086) and the non-translocon associated OST complex (EMD-0087).

References

[1] Dubochet, J. *et al.* Cryo-electron microscopy of vitrified specimens. *Q Rev Biophys* **21**, 129–228 (1988).

[2] Lucic, V., Rigort, A. & Baumeister, W. Cryo-electron tomography: the challenge of doing structural biology in situ. *J Cell Biol* **202**, 407–419 (2013).

[3] Oikonomou, C. M. & Jensen, G. J. Cellular electron cryotomography: Toward structural biology in situ. *Annual Review of Biochemistry* **86**, 873–896 (2017).

[4] Medalia, O. *et al.* Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* **298**, 1209–13 (2002).

[5] Ortiz, J. O., Forster, F., Kurner, J., Linaroudis, A. A. & Baumeister, W. Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition. *J Struct Biol* **156**, 334–41 (2006).

[6] Beck, M. *et al.* Visual proteomics of the human pathogen Leptospira interrogans. *Nat Methods* **6**, 817–23 (2009).

[7] Rigort, A. *et al.* Automated segmentation of electron tomograms for a quantitative description of actin filament networks. *J Struct Biol* **177**, 135–144 (2012).

[8] Asano, S. *et al.* Proteasomes. a molecular census of 26s proteasomes in intact neurons. *Science* **347**, 439–442 (2015).

[9] Rickgauer, J. P., Grigorieff, N. & Denk, W. Single-protein detection in crowded molecular environments in cryo-em images. *eLife* **6**, e25648 (2017).

[10] Volkmann, N. Methods for segmentation and interpretation of electron tomographic reconstructions. *Methods Enzymol* **483**, 31–46 (2010).

[11] Fernandez, J.-J. Computational methods for electron tomography. *Micron* **43**, 1010–1030 (2012).

[12] Lucic, V., Fernández-Busnadiego, R., Laugks, U. & Baumeister, W. Hierarchical detection and analysis of macromolecular complexes in cryo-electron tomograms using pyto software. *Journal of structural biology* **196**, 503–514 (2016).

[13] Fernández-Busnadiego, R. *et al.* Quantitative analysis of the native presynaptic cytomatrix by cryoelectron tomography. *J Cell Biol* **188**, 145–56 (2010).

[14] Vargas, K. J. *et al.* Synucleins have multiple effects on presynaptic architecture. *Cell reports* **18**, 161–173 (2017).

[15] Förster, F., Medalia, O., Zauberman, N., Baumeister, W. & Fass, D. Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography. *Proc Natl Acad Sci U S A* **102**, 4729–4734 (2005).

[16] Schur, F. K. *et al.* An atomic model of hiv-1 capsid-sp1 reveals structures regulating assembly and maturation. *Science* **353**, 506–508 (2016).

[17] Wan, W. & Briggs, J. Cryo-electron tomography and subtomogram averaging. In *Methods in enzymology*, vol. 579, 329–367 (Elsevier, 2016).

[18] Bharat, T. A. & Scheres, S. H. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in relion. *Nature protocols* **11**, 2054 (2016).

[19] Sousbie, T. The persistent cosmic web and its filamentary structure–i. theory and implementation. *Monthly Notices of the Royal Astronomical Society* **414**, 350–383 (2011).

[20] Milnor, J. Morse theory, volume 51 of annals of mathematics studies. *Princeton, NJ, USA* (1963).

[21] Forman, R. A user's guide to discrete morse theory. *Seminaire Lotharingien de Combinatoire* **48**, 35pp (2002).

[22] Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).

[23] Pfeffer, S. *et al.* Structure of the native sec61 protein-conducting channel. *Nature communications* **6**, 8403 (2015).

[24] Pfeffer, S. *et al.* Structure of the mammalian oligosaccharyl-transferase complex in the native er protein translocon. *Nature communications* **5**, 3072 (2014).

[25] Pfeffer, S., Dudek, J., Zimmermann, R. & Förster, F. Organization of the native ribosome–translocon complex at the mammalian endoplasmic reticulum membrane. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1860**, 2122–2129 (2016).

[26] Stoyan, D. Fundamentals of point process statistics. In Baddeley, A., Gregori, P., Mateu, J., Stoica, R. & Stoyan, D. (eds.) *Case studies in spatial point process modeling* (Springer, 2006).

[27] Ripley, B. D. *Spatial statistics* (Willey-Interscience, 1981).

[28] Wiegand, T. & Moloney, K. A. Rings, circles, and null-models for point pattern analysis in ecology. *Oikos* **104**, 209–229 (2004).

[29] Xu, M., Beck, M. & Alber, F. Template-free detection of macromolecular complexes in cryo electron tomograms. *Bioinformatics* **27**, i69–i76 (2011).

[30] Xu, M., Tocheva, E. I., Chang, Y.-W., J., J. G. & Alber, F. De novo visual proteomics in single cells through pattern mining. *arXiv:1512.09347* .

[31] Pei, L., Xu, M., Frazier, Z. & Alber, F. Simulating cryo electron tomograms of crowded cell cytoplasm for assessment of automated particle picking. *BMC Bioinformatics* **17**, 405 (2016).

[32] Shrimal, S., Cherepanova, N. A. & Gilmore, R. DC2 and KCP2 mediate the interaction between the oligosaccharyltransferase and the ER translocon. *J Cell Biol* **216**, 3625–3638 (2017).

[33] Braunger, K. *et al.* Structural basis for coupling protein transport and N-glycosylation at the mammalian endoplasmic reticulum. *Science* **360**, 215–219 (2018).

[34] Dunkley, P. R. *et al.* A rapid percoll gradient procedure for isolation of synaptosomes directly from an s1 fraction: homogeneity and morphology of subcellular fractions. *Brain Res* **441**, 59–71 (1988).

[35] Godino, M. d. C., Torres, M. & Sánchez-Prieto, J. Cb1 receptors diminish both ca(2+) influx and glutamate release through two different mechanisms active in distinct populations of cerebrocortical nerve terminals. *J Neurochem* **101**, 1471–1482 (2007).

[36] Fernández-Busnadiego, R. *et al.* Cryo-electron tomography reveals a critical role of rim1α in synaptic vesicle tethering. *J Cell Biol* **201**, 725–740 (2013).

[37] Koster, A. J. *et al.* Perspectives of molecular and cellular electron tomography. *J Struct Biol* **120**, 276–308 (1997).

[38] Danev, R., Buijsse, B., Khoshouei, M., Plitzko, J. M. & Baumeister, W. Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proc Natl Acad Sci USA* **111**, 15635–15640 (2014).

[39] Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using imod. *J Struct Biol* **116**, 71–76 (1996).

[40] Peixoto, T. P. The graph-tool python library. *figshare* (2014).

[41] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[42] Schroeder, W. J., Lorensen, B. & Martin, K. *The visualization toolkit: an object-oriented approach to 3D graphics* (Kitware, 2004).

[43] Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

[44] Ayachit, U. The paraview guide: a parallel visualization application (2015).

[45] Martinez-Sanchez, A., Garcia, I., Asano, S., Lucic, V. & Fernandez, J.-J. Robust membrane detection based on tensor voting for electron tomography. *J Struct Biol* **186**, 49–61 (2014).