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

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

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 provides an accurate representation of molecular architecture of cells. However, detection and precise localization of macromolecular complexes within cellular environments is aggravated by the presence of many molecular species and molecular crowding. We developed a template-free image processing procedure for accurate tracing of complex networks of densities in cryo-electron tomograms, a comprehensive and automated detection of heterogeneous membrane-bound complexes and an unsupervised classification (PySeg). Applications to intact cells and isolated endoplasmic reticulum (ER) allowed us to detect and classify small protein complexes. This classification provided sufficiently homogeneous particle sets and initial references to allow subsequent *de novo* subtomogram averaging. Spatial distribution analysis showed that ER complexes have different localization patterns forming nanodomains. Therefore this procedure allows a comprehensive detection and structural analysis of complexes *in situ*.

Introduction

The cellular environment is characterized by the presence of many different molecular species. Complexes, stable or transient underlie critical cellular functions. Of particular interest are membrane-bound complexes because they are essential for many types of cellular processes, like cell signaling, immune response and synaptic transmission, and are targeted by more than two thirds of all drugs [1].

In cryo-electron tomography (cryo-ET), biological samples are faithfully preserved by rapid freezing, which prevents water crystallization and rearrangements of the biological material and causes sample vitrification. Importantly, samples are imaged in transmission electron microscopy in the same vitrified, fully hydrated state [2, 3]. Therefore, cryo-ET is uniquely suited for high resolution, direct three-dimensional (3D) imaging of unperturbed cellular environments [4, 5].

The potential of cryo-ET to yield cellular maps of molecular complexes is hampered by low signal-to-noise ratio (SNR) in tomograms and the molecular heterogeneity in cells. Because visual detection is limited to large complexes of characteristic shapes [6], image processing methods have been developed to interpret tomograms. In template matching, currently the leading detection method in cryo-ET, a 3D structure of a protein or complex of interest is used to computationally search for instances of that complex [7, 8, 9, 10]. This approach is particularly suited for complexes that do not form part of larger assemblies and critically depends on the already existing 3D structures of complexes of interest. Automated methods were developed for segmentation of cellular components of particular shape, such as lipid membranes and filaments [11, 12, 13, 14], and also

1

pleomorphic, membrane-associated complexes [15, 16, 17]. Subtomogram averaging can yield high resolution 3D structures, but requires that a large number of complex of interest is present [18, 19, 20].

To allow comprehensive, high resolution processing of cellular cryo-tomograms, we developed a software procedure for template-free detection and unsupervised classification of heterogeneous membrane-bound molecular complexes (PySeg). We adapted and further developed methods from other fields such as the discrete Morse theory segmentation, affinity propagation (AP) clustering and spatial point processes and contributed new software. Classes obtained were sufficiently homogeneous to allow standard subtomogram averaging [21, 20]. Validations and comparisons with other methods were performed on phantom and real datasets.

Results

Procedure overview

Our procedure consists of three major parts (the workflow scheme is shown in the Methods). First, densities are traced and complexes 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

For an automated tracing of density in cryo-tomograms, we developed a procedure based on the discrete Morse theory, which is used to calculate topological indices of manifolds and simplicial complexes [22, 23]. We used DisPerSE package, originally developed to identify astrophysical structures in 3D images of the large-scale matter distribution in the Universe [24], to generate Morse complexes comprising the following manifolds: (i) Greyscale minimum points (0-critical points), (ii) Saddle points having minima in two and a maximum in one direction (1-critical points), (iii) Arcs connecting minima and saddle points defined as maximum gradient curves between these points, (iv) 3-manifolds associated with minima (ascending 3-manifolds) (Figure 1 A, C). The minima and arcs visually corresponded well to the distribution of the density.

Complex network of densities and a high level of noise present in cryo-tomograms results in many closely spaced local minima, causing an overly complex structure of the calculated Morse complexes. To solve this problem, we implemented a modified version of the simplification by topological persistence [24], whereby a saddle point and its adjacent minimum that have similar grayscale values are removed (low persistence pairs), and the affected Morse complex elements are reassigned (the algorithm is presented in the Methods; Supplementary Figure S1). This resulted in a greatly simplified Morse complex, and a faithful tracing of density (Figure 1D).

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 densities in cellular cryo-tomograms.

Graph embedding and detection of complexes

To allow further processing, we implemented a procedure that assigns minima to graph vertices and arcs to graph edges (Figure 1B). Vertices and edges keep the information about the spatial location, greyscale density, geometry and the connectivity of the underlying minima and arcs, as well as external information, such as the identity of lipid membranes. These spatially embedded graphs combine precise geometrical and external information to represent the distribution of the biological material (proteins and lipids).

Importantly, they allow computationally efficient queries to extract geometrical and greyscale information and to detect individual complexes. For example, subgraphs can be selected starting from vertices belonging to, or at a specified distance from a previously defined membrane or another cellular structure, and continuing along arcs to trace a complex (Figure 1D). Not all selected subgraphs can be directly matched to complexes, even though subgraphs represent densities. The repositioning or elimination of wrongly selected subgraphs is delegated to the subsequent processing steps. Hence, the Morse-based detection of complexes described so far corresponds to particle picking in the single particle analysis.
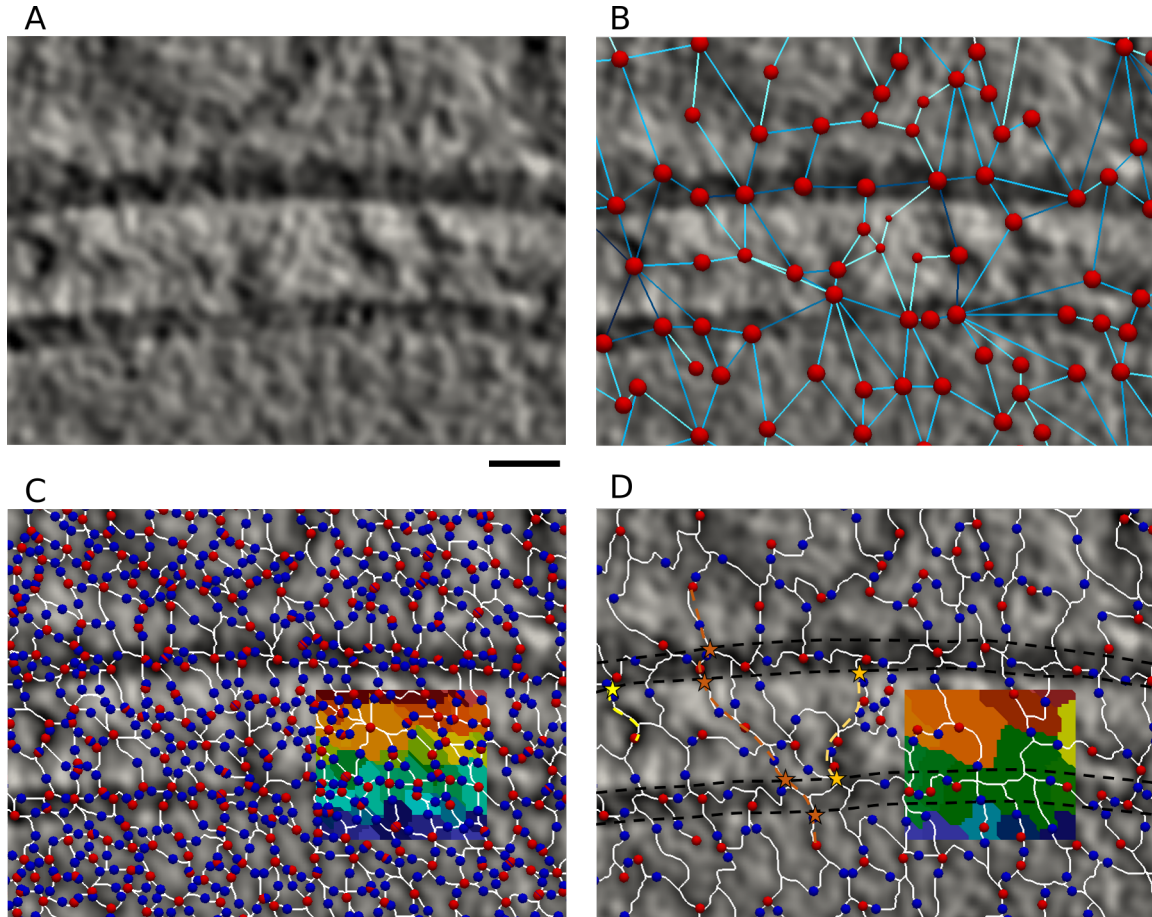
Figure 1: Tracing density at a synapse from rodent cerebrocortical synaptosomal preparation. A Tomographic slice of 1.37 nm thickness. B Graph representation of the Morse complex shown in D. Note that although graph edges are shown by straight lines, they keep the full information about the arcs they represent (shown on 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. Yellow, orange and brown paths represent possible extracellular presynaptic, trans-cleft and extended trans-cleft complexes. Asterisks are intersection points between the selected paths and membrane faces. Black dashed lines outline synaptic membranes. 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 and graph vertices, the blue ones the saddle points, white lines arcs and blue lines graph edges. In B, larger size vertices denote minima of higher density and darker shade edges denote saddle points of higher density. Scale bar 10 nm. Density was traced on N=67 tomograms.

Density tracing validation on phantom data

We applied Morse density tracing to a phantom dataset comprising a rectangular grid with variable amounts of Gaussian noise (Figure S2A). Grid intersections and grid bars were taken as the ground truth features for the detection of minima and arcs, respectively (Figure S2B). Multiple minima occurring at the same grid intersection were avoided by imposing an exclusion distance between particles. For SNR above 0.05, the false positives (FPs) and false negatives (FNs) were below 10% and true positive (TP) minima was above 90%, however for SNR between 0.05 and 0.1 TP arcs was between 80% and 90% (Figure S2C). 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 S2C).

General classification

The application of the Morse detection procedure to complex cellular systems is expected to yield a set of membrane-bound complexes possessing high compositional and conformational heterogeneity. Therefore, a general classification procedure capable of separating highly heterogeneous complexes is required for further processing.

The direction of the vector perpendicular to the membrane is determined for all particles. To optimize these normal vectors and particle positions, we employ constrained particle refinement, whereby the two Euler angles corresponding to the normals are allowed only small changes around the initial values during the alignment step, while the third angle (around the normal) is not constrained. However, imposing a high symmetry diminishes its importance. The initial reference for refinement is obtained from the data, by randomizing the third angle and averaging all particles without alignment. The symmetrization around the normals and the choice of the initial reference reduce the influence of the missing wedge.

Particles are classified by the affinity propagation (AP) clustering, whereby nodes (particles) exchange information between each other to reach the optimal partitioning [25]. The advantages of AP compared to the standard clustering methods used in single particle processing are that this algorithm is unsupervised, the number of classes is not specified in advance but data driven, and it can handle cases where classes have a very different number of particles.

The success of a clustering procedure critically depends on the definition of the clustering distance (similarity). Here, we compute particle rotational averages around their normal vectors (Supplementary Figures S3, S5B, S7B) and define the distance between two particles as the dot product of their normalized rotational averages. In this way, the 2D averages used for clustering are pre-aligned with each other, leaving no further degrees of freedom that could affect the clustering. On the contrary, clustering 3D particles would require the determination of the angle around the normal vector, thus hampering the procedure.

Validation of general classification

To evaluate the AP classification independent of particle picking and to compare it with two other methods used in the field (K-means and hierarchical clusterings), we generated a test dataset from eight available reference structures of membrane-bound complexes. To make realistic particles we added different amounts of noise, rotated them, imposed the missing wedge and translated particles along the membrane (see the Methods; Figure S3A).

We classified the test dataset and compared results against the ground truth using three measures. Fowlkes-Mallows [26] and the Variation of information [27] compare two classifications directly and are currently the state-of-art methods, but are rarely used in cryo-ET [17]. We also made a correspondence between the classes obtained and the ground truth classes based on the number of reference structure elements within each class, and determined F_1 measure (see the Methods).

AP classification was weakly affected by changes in SNR, displacement range and the number of reference structures (Figure S4A-C). The input preference parameter required for AP classification only weakly influenced
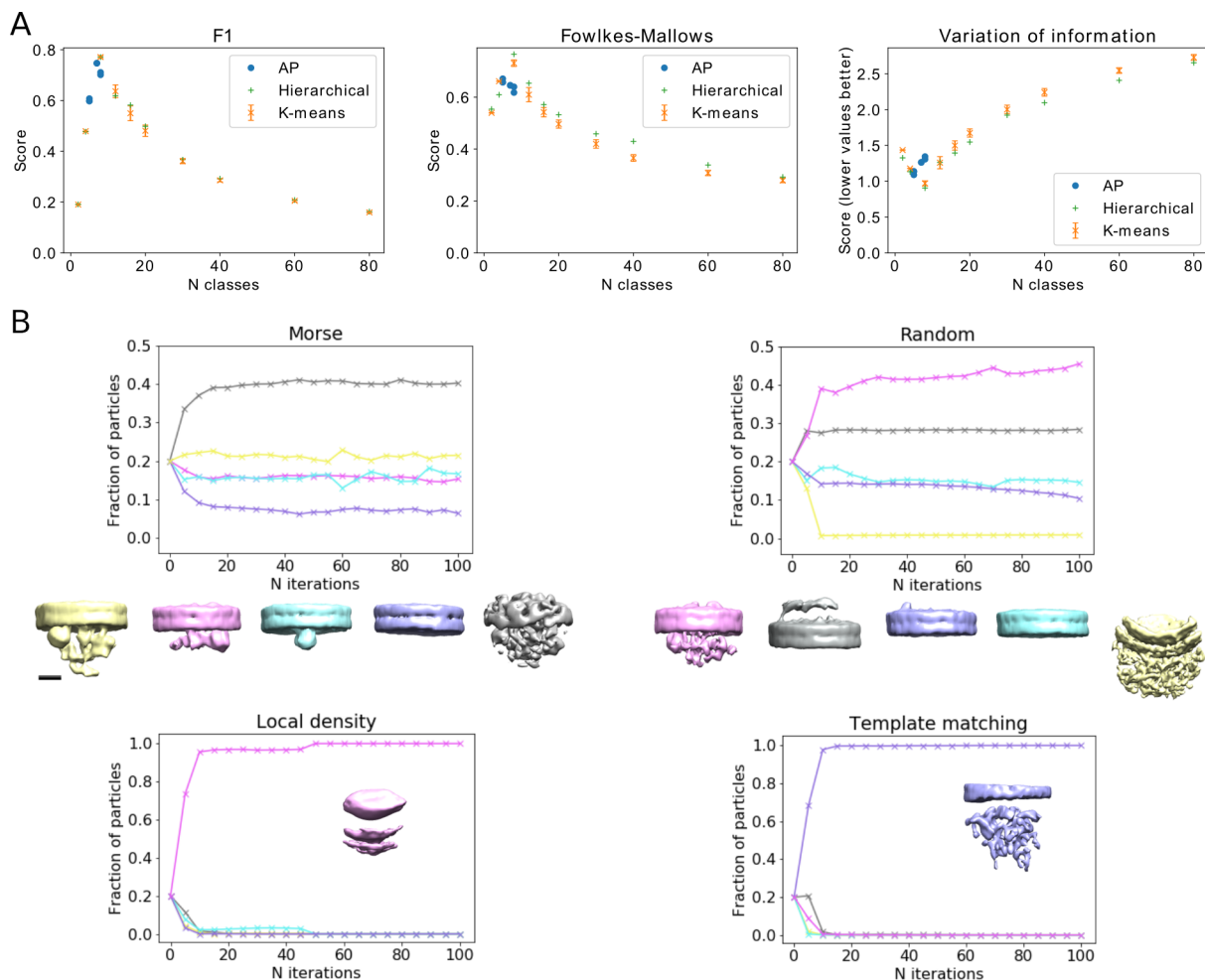
Figure 2: Method validations. A Comparison of the AP with K-means and hierarchical clusterings on the test dataset comprising eight available structures of membrane-bound complexes. F_1 (left), Fowlkes-Mallows (middle) and Variation of information (right). Note that higher values of the Fowlkes-Mallows and F_1 measures, but lower values of the Variation of information signify better agreement. The AP clustering results are shown for the preference parameter taking all integer values between -9 and -3 (several data points overlap). For K-means clustering the data points denote means and the error bars the standard deviation obtained from 200 simulations. B Validation of particle picking on the lumenal particles from the microsomal dataset. Graphs showing the number of particles per class are shown together with the class averages. Morse picking (up, left), random (up, right), local density (bottom left) and template matching (bottom right). Scale bar 5 nm.

the scores. The best results were obtained for the input preference parameter values between -6 and -3 and good results between -10 and -2 defining the values suitable for cryo-tomograms.

AP classification of datasets that include the lipid membrane showed similar results (F_1= 0.89-0.98 for 800-4000 particles). 3D class averages matched the reference structures (Figure S5). The high similarity between the AP class representatives and the ground truth class averages confirmed the strength of the AP classification (Figure S3B).

K-means and hierarchical clustering methods require the number of classes as an input parameter. Our data showed that when the correct number of clusters was specified, the performance of K-means and hierarchical clusterings was similar to the AP using the default input preference values. However, it was drastically reduced for realistic cases when non-optimal number of classes was chosen (Figure 2A). These conclusions remained the same when r-weighting was used (see the Methods), only the scores improved (Figure S4D). Therefore, the distinctive advantage of the AP classification is that it yields optimal results without requiring external information.

Detection and classification on microsomal membranes

To test the particle localization and general classification methods introduced above, we used a subset (26%) of previously analyzed cryo-ET data depicting canine pancreatic microsomes [28]. 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) [29].

We applied Morse detection to a subset (26%) of previously analyzed cryo-ET dataset depicting canine pancreatic microsomes [28]. (Figure 3A, Video 1). Particles at the cytoplasmic and lumenal faces of the ER membrane were located independently of each other, using geometrical constraints (see the Methods). Particle positions and membrane normals vectors were optimized, yielding an average densities having a well-positioned density and a resolved lipid bilayer (Figure 3B).

Classification of cytosolic particles by AP yielded more than 100 classes (Figure S6A). Constrained *de novo* refinement of these classes showed different species of ribosome-translocon complexes (Figure 3C, D; S7). All cytosolic particles were subjected to three rounds of 3D classification, resulting in structures comparable to those previously reported (Figures 3E; S8) [30, 28]. These included fully assembled ribosomes bound to the fully assembled and partial translocon complexes, as well as ribosomal large subunits (resolved to 18 Å, 22 Å and 21 Å respectively) (Figure S6C). This shows that our procedure generated sufficiently homogeneous particle sets and provides internal reference structures to allow further *de novo* classification and averaging.

Lumenal particles were classified by AP and the class showing the best resolved lumenal density by visual inspection was refined to yield a *de novo* 3D average of the translocon (Figures 3F top left; S6B). 3D classification, using the translocon average as the initial reference, yielded ribosome-translocon complex (15%), fully assembled (TRAP, OST and Sec61) ribosome-free translocon (68%) and a likely individual ribosome-free OST complexe (17%) (resolved to 22 Å, 14 Å and 16 Å, respectively) (Figures 3F; S6D). Importantly, without the initial reference generated from AP classification, the same 3D classification procedure failed.

Therefore, our Morse theory-based detection is capable of picking small complexes, like the translocon or even smaller individual OST complexes ($\approx$260 kDa and $\approx$200 kDa lumenal mass, respectively). The unsupervised classification by AP was instrumental to carry the processing to a level where the standard 3D classification and refinement procedures could be used.

Validation of density tracing and particle picking on microsomal membranes

Next, we compared the performance of the Morse detection method with the standard methods used in the field: template matching, local density detection (after low-pass filtering) and random picking. In all cases we analyzed a subset of tomograms used in the previous section and picked particles on the lumenal side of the microsomes, up to 20 nm to the membrane. For template-matching, we used our ribosome-free translocon average density (Figure 3F).
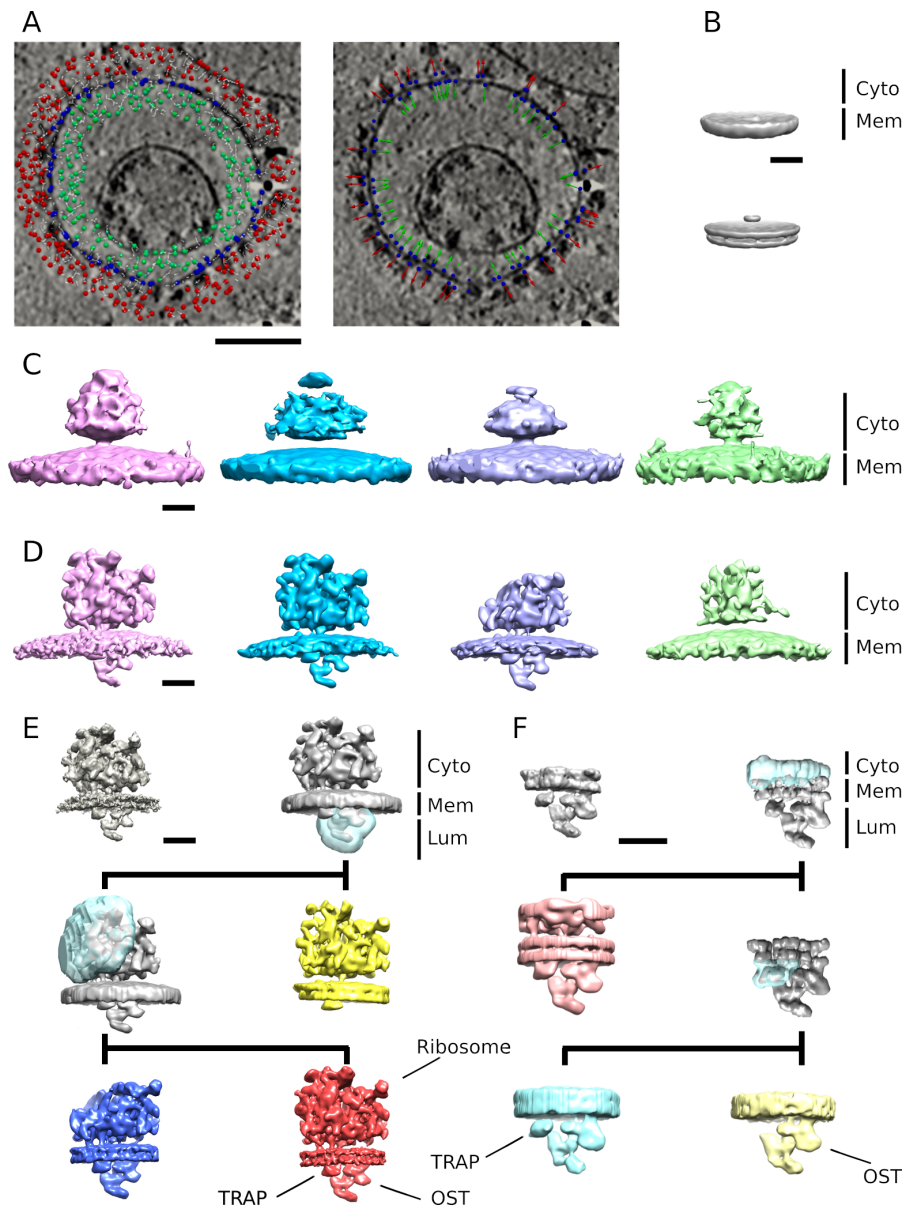
Figure 3: Processing of ER membrane-associated complexes from the microsomal dataset. A Tomographic slices (EMD-10449) showing density tracing. Density minima are shown as small spheres (left) and particles as spheres with arrows representing the associated membrane normal vectors (right). Position of the spheres is color-coded: Red cytosol; blue membrane; green lumen. (N=55 tomograms) 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 AP 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 references. E 3D classification and refinement of cytosolic particles (ribosomes). Densities of ribosomes bound to the fully assembled (red, EMD-0074) and partial translocon complex (yellow, EMD-0084), as well as the large ribosomal subunit bound to the fully assembled translocon (blue, EMD-0075) are shown. F 3D classification and refinement of lumenal particles (translocon). Densities of the ribosome-translocon complex (light red, EMD-0085), the ribosome-free fully assembled translocon (light blue, EMD-0086) and the non-translocon associated individual OST complex (light yellow, EMD-0087) are shown. In both (E) and (F) initial references obtained from the AP 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.

To compare the methods directly, we skipped the classification and subjected the four picked particle sets to the same 3D constrained refinement with a high symmetry, followed by 3D classification that used our ribosome-free translocon density as the initial reference. In all cases, the classification converged to a stable number of particles after 15-20 iterations (Figure 2B). All five classes from Morse-detected particles contained a significant number of particles. One class represented the translocon complex, while another two classes likely represented two other small membrane-bound complexes. Among the last two classes, one probably contained false positive picks because the average showed just a well-resolved membrane bilayer, while the other class likely contained a mixture of other complexes. This shows that small complexes can be reliably detected by the Morse method.

Four out of five classes of random-picked particles contained a significant number of particles, but none of them yielded recognizable densities. The local density detection and template matching yielded particles that all converged to one class and produced poor averages. Therefore, among the particle picking methods tested, the Morse method was the only one that successfully detected small complexes.

Detection and classification in situ

We applied Morse density tracing to tomograms of intact P19 embryonic carcinoma cells and picked particles in the vicinity of endoplasmic reticulum (ER) membranes (Figure 4A, B). Because the orientation of the membranes could not be always determined and the cellular interior is densely populated, we included three rounds of AP classification and obtained the following datasets: large densities located on the same side of the membrane where particles were picked ("large same side"), large densities on the side opposite from the particles ("large opposite side") and the small densities (Figure S9).

Ribosomes were prominent in the AP classes derived from the large same side dataset. Some other classes showed densities similar to the known structures, allowing us to tentatively assign them to the inositol trisphosphate (IP3) receptor and two different assembly states of the MHC-I peptide loading complex (PLC) (Figures 4, S10A) [31, 32].

The ribosome-translocon complex was detected in the large opposite side dataset (Figures 4, S10B). Morse procedure detected the translocon complex directly, while the large densities (ribosomes) seen in the AP class averages arose during the classification.

One AP class average obtained from the small densities dataset was visually consistent with the ribosome-free full translocon and OST complexes from the microsomal data, as well as with the ribosome-bound translocon from the *in situ* data (Figures 4, S10C, 3F). Several other AP averages showed small densities, the identification of which would require a more in-depth analysis. Inspection of these classes showed homogeneous sets of particles, confirming the validity of the classification.

Refinements of randomly chosen small density particles after the first and the second AP classification resulted in averages showing only the membrane, confirming the importance of the AP classification. Therefore, our procedure is applicable to cryo-tomograms of intact cellular interior and allows the detection of even small ER membrane-bound complexes and their separation into classes that are sufficiently homogeneous for standard processing.

Spatial distribution analysis

Methods available for the analysis of spatial point processes can provide further information about the distribution and clustering of complexes and assist with classification. Monovariate first order distribution functions are based on the distance to the closest point, either from other points (nearest neighbor distribution), from arbitrary locations (spherical contact function), or both (J-function) [33]. 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 [34, 35]. 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
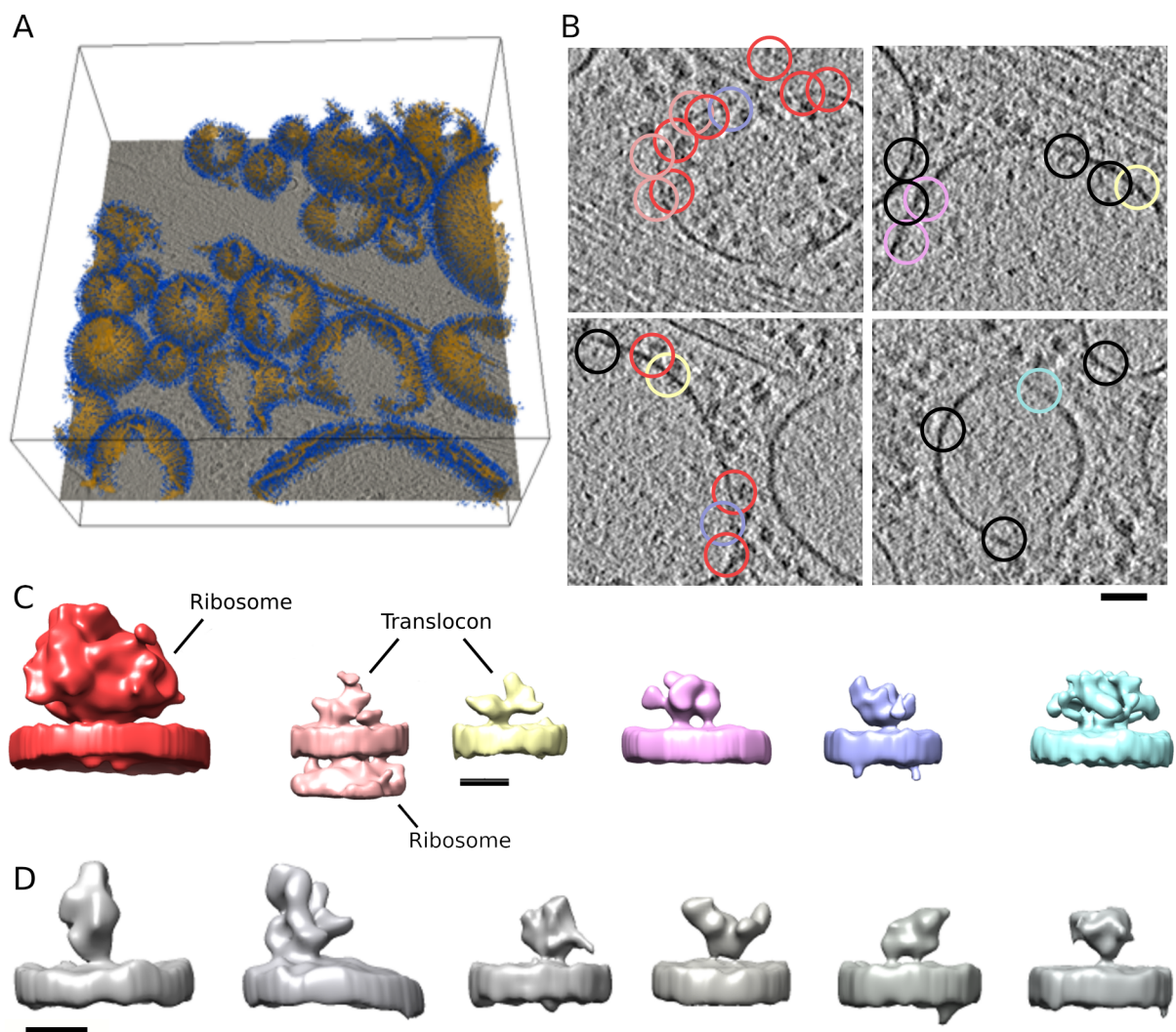
Figure 4: Complexes resolved *in situ* from the intact P19 cells dataset. A Tomographic slice. Superposed are segmented internal membranes (ochre-yellow) and picked particles including the normals (blue). B Tomographic slices (EMD-10439) showing x and y positions of some particles classified as complexes shown in C (colors match). The exact particle positions can be on tomographic slices up to 20 nm away from the shown slices. Scale bar 50 nm. C Average densities, from left to right: membrane-associated ribosome (red, EMD-10432), ribosome-associated translocon (light red, EMD-10433), ribosome-free translocon (yellow, EMD-10434), two putative PLC complexes (magenta EMD-10435 and purple EMD-13436), putative IP3 receptor (blue, EMD-10437). Scale bar 10 nm. D Selection of the other class averages. Scale bar 10 nm. N=12 tomograms.
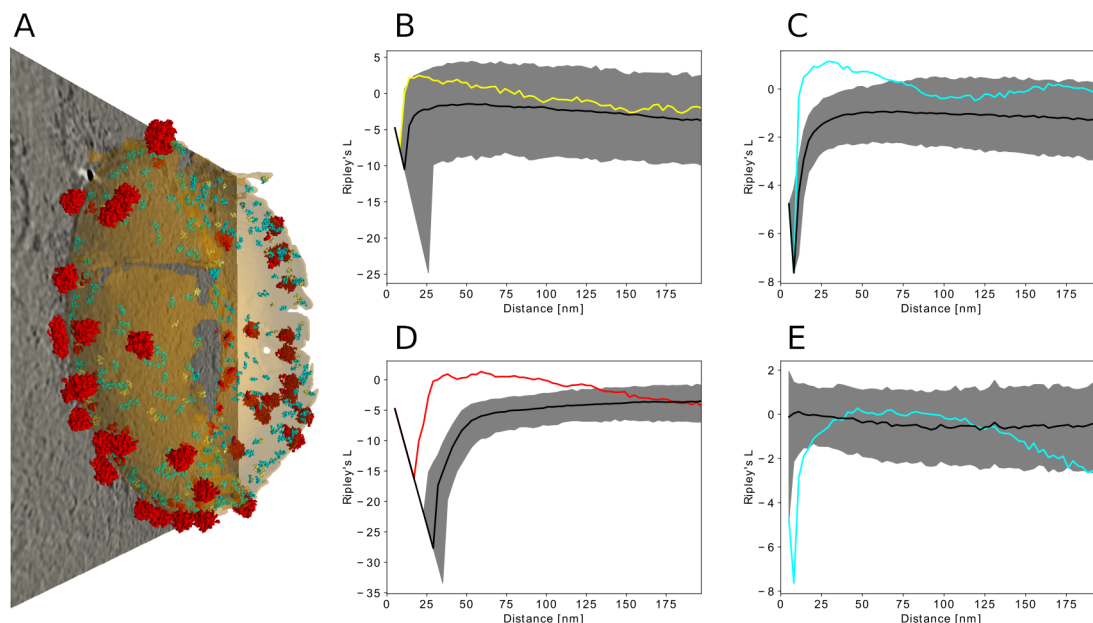
Figure 5: Spatial distribution of microsomal particles. A Distribution of the particles is shown on one microsome (EMD-10450) (all three cytosolic ribosome classes red, ribosome-free full translocon: cyan, individual OST: yellow). N=55 tomograms. B-D Univariate Ripley's L function of the three classes, 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 (N=1220 random distributions) and the gray areas represent regions of p>0.05 confidence.

within the particle region. To this end, we implemented a Monte-Carlo method that generates multiple random distributions of a specified number of particles within an arbitrary space (see Methods).

A clustering method based on Voronoi tessellation was recently developed for the detection of large point clusters on a dense background present in super-resolution fluorescence data [36]. This is substantially different from our system, where points (particles) are less numerous but more significant because they were previously selected by stringent classification steps.

Spatial organization of microsomal complexes

Visually, microsomal particle classes followed distinct distributions (Figure 5A, Video 2). Univariate Ripley's function showed a significant clustering of the ribosome-free translocon complex at length scales of 8- 50 nm, while the individual OST was borderline significant (at p=0.05 level) at 10-20 nm (Figure 5B,C). The nearest neighborhood, spherical contact distribution and J-functions confirmed these results (Figure S11).

As expected, the distribution of ribosomes, comprising all three final classes of the cytosolic particles (Figure 3E), also showed significant clustering (Figure 5D), likely induced by polyribosome formation. Using the bivariate Ripley's L function, we did not detect a significant colocalization between ribosomes and ribosome-free translocon complexes (Figure 5E). These examples show that spatial point process methods allow a quantitative characterization of the organization of molecular complexes.

Discussion

The mining of the rich information content of cellular cryo-electron tomograms is hindered by the crowded nature of cells populated by many different molecular species. To overcome this problem, we created a template-

free procedure for detection and classification of pleomorphic membrane-bound molecular complexes visualized in cellular cryo-tomograms.

In order to accurately trace density in cryo-electron tomograms and to automate detection and localization of complexes, we adapted and extended an existing discrete Morse theory-based segmentation method [24]. Direct comparison on real cryo-tomograms showed that Morse detection performs much better than other commonly used methods and that it was the only one that could detect small complexes.

Because the Morse detection is rather comprehensive, it yields highly heterogeneous particles as well as false positives. We found that the combination of unsupervised classification by AP clustering [25] and the use of 2D particle rotation averages around the membrane normals allows an efficient and successful classification.

K-means and hierarchical clusterings require setting the number of classes in advance, or in case of Ward's agglomerative hierarchical clustering imposing a heuristically determined threshold for internal variance in order to limit merging of classes. A distinctive advantage of AP is that it determined the optimal number of classes from the data, requiring one parameter that can take only few discrete values, which we here determined and showed to be applicable to both real and synthetic data.

The results obtained from *in situ* and isolated ER membranes, expected to contain several hundred different proteins [37], provide structural proof for the presence of ribosome-free translocon complexes and OST, small ER membrane-associated complexes of ≈260 and 200 kDa lumenal mass, respectively. They are likely capable of being engaged in post-translational processes because the majority of ribosome-free translocon complexes is known to already comprise all constituents present in the ribosome-associated translocon. The individual OST complexes we localized on the ER membrane likely correspond to STT3B type OST, known to be excluded from the translocon in metazoa [38].

In both applications, this procedure allowed us to determine *de novo* average 3D densities. The generally lower resolution of the *in situ* data can be attributed to a more dense and heterogeneous cellular environment and to technical differences between these datasets.

The procedure presented here can be applied to small complexes in their native state, that were beyond the reach of template-matching. It is not limited by the availability of high resolution structures and does not introduce an initial bias that might affect subtomogram classification and averaging. Furthermore, being template-free, our procedure can localize heterogeneous complexes, or complexes that adopt different compositional and conformational states.

Because Morse detection is independent of membranes and the ability to determine normal vectors is the only membrane-related requirement for classification, our procedure can be applied to complexes attached to any cellular membrane and also larger structures such as the cytoskeleton. Previous template-free approaches, based on automated pattern mining, deep learning and the difference of Gaussians picking methods, were successful only on large complexes, and did not reach resolution comparable to ours, neither on simulated nor on real datasets [39, 40, 41, 14]. The reason for the success of our procedure may be due to the rotational averaging of membrane-bound complexes, which eliminates the problem of incorrect alignment that can hinder classification and diminishes the effect of the missing wedge, while keeping sufficient structural information for classification.

The failure of methods based on 3D rotation-invariant properties for classification, a promising method for further developments, to detect membrane-attached complexes [41] argues that processing membrane-bound and membrane-independent complexes are two distinct problems that may require different approaches. Because of the major importance of membrane-bound complexes for biochemical signaling pathways, such as those involved in the immune response, development or synaptic transmission, as well as their dominant role in drug development [1], the procedure described here may be applied to a wide range of fundamental biological processes.

The implementation of the spatial distribution functions to characterize point-particle patterns in regions of arbitrary shape, we observed a significant clustering of ribosome-free translocon complexes on microsomal membranes. 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, our results demonstrate that the template-free procedure presented here is uniquely suited to accurately trace density in cryo-electron tomograms, localize and classify heterogeneous membrane-bound molecular complexes, and provide initial references for 3D refinement. The pipeline consisting of our procedure,

the existing automated membrane segmentation and subtomogram averaging allowed us to obtain 3D structures and analyze localization of complexes directly from cellular tomograms. Therefore, it extends the applicability of cryo-ET to small and heterogeneous membrane-bound molecular complexes in their native state and makes large-scale, non-invasive detection, localization and *de novo* structure determination of molecular complexes *in situ* possible.

Acknowledgements

Author Contributions

AM-S and VL conceived and designed the research. AM-S designed and implemented the software; ZK, UL and SC acquired original tomograms; SP provided expertise 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.

References

[1] Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nature reviews. Drug discovery* **16**, 19–34 (2017).

[2] Taylor, K. A. & Glaeser, R. M. Electron diffraction of frozen, hydrated protein crystals. *Science* **186**, 1036–7 (1974).

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

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

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

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

[7] 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).

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

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

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

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

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

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

[14] Chen, M. *et al.* Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. *Nature methods* **14**, 983–985 (2017).

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

[16] 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).

[17] 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).

[18] 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).

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

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

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

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

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

[24] 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).

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

[26] Fowlkes, E. B. & Mallows, C. L. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* **78**, 553–569 (1983).

[27] Meila, M. Comparing clusterings - an information based distance. *Journal of Multivariate Analysis* **98**, 873–895 (2007).

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

[29] 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).

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

[31] Fan, G. *et al.* Gating machinery of insp3r channels revealed by electron cryomicroscopy. *Nature* **527**, 336–341 (2015).

[32] Blees, A. *et al.* Structure of the human mhc-i peptide-loading complex. *Nature* **551**, 525–528 (2017).

[33] 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).

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

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

[36] Andronov, L. *et al.* 3dclustervisu: 3d clustering analysis of super-resolution microscopy data by 3d voronoi tessellations. *Bioinformatics (Oxford, England)* **34**, 3004–3012 (2018).

[37] Han, D. K., Eng, J., Zhou, H. & Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature biotechnology* **19**, 946–951 (2001).

[38] 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).

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

[40] 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).

[41] Xu, M. *et al.* De novo structural pattern mining in cellular electron cryotomograms. *Structure (London, England : 1993)* (2019).

[42] 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).

[43] 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).

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

[45] Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J Struct Biol* **152**, 36–51 (2005).

[46] 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).

[47] Zheng, S. Q. *et al.* Motioncor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature methods* **14**, 331–332 (2017).

[48] Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-em using a 2.6 Å reconstruction of rotavirus vp6. *eLife* **4**, e06980 (2015).

[49] 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).

[50] Nash, C. & Sen, S. *Topology and geometry for physicists* (Academic Press, Harcourt Brace Jovanovic, London, 1990).

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

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

[53] Meyerson, J. R. *et al.* Structural mechanism of glutamate receptor activation and desensitization. *Nature* **514**, 328–334 (2014).

[54] Karakas, E. & Furukawa, H. Crystal structure of a heterotetrameric nmda receptor ion channel. *Science (New York, N.Y.)* **344**, 992–997 (2014).

[55] Herguedas, B. *et al.* Structure and organization of heteromeric ampa-type glutamate receptors. *Science (New York, N.Y.)* **352**, aad3873 (2016).

[56] Wu, J. *et al.* Structure of the voltage-gated calcium channel ca(v)1.1 at 3.6?Å resolution. *Nature* **537**, 191–196 (2016).

[57] Morales-Perez, C. L., Noviello, C. M. & Hibbs, R. E. X-ray structure of the human ?4?2 nicotinic receptor. *Nature* **538**, 411–415 (2016).

[58] Tao, X., Hite, R. K. & MacKinnon, R. Cryo-em structure of the open high-conductance ca, javax.xml.bind.jaxbelement@741fc2e3, -activated k, javax.xml.bind.jaxbelement@4349c43b, channel. *Nature* **541**, 46–51 (2017).

[59] Park, E., Campbell, E. B. & MacKinnon, R. Structure of a clc chloride ion channel by cryo-electron microscopy. *Nature* **541**, 500–505 (2017).

[60] Zhang, Y. *et al.* Cryo-em structure of the activated glp-1 receptor in complex with a g protein. *Nature* **546**, 248–253 (2017).

[61] 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).

[62] Hrabe, T. *et al.* Pytom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *Journal of structural biology* **178**, 177–188 (2012).

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

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

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

[66] Pettersen, E. F. *et al.* Ucsf chimera, a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–12 (2004).

[67] Martinez-Sanchez, A. & Vladan, L. Pyseg: Template-free detection and classification for cryo-et. *Code ocean* https://doi.org/10.24433/CO.0526052.v1 (2019).

Online methods

Synaptosomal preparation

Cerebrocortical synaptosomes were extracted from 6–8 week old male Wistar rats as described previously [42, 43, 16] in accordance with the procedures accepted by the Animal facility of the Max Planck Institute for Biochemistry, in accordance with the German legislation. 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.

Preparation of P19 cells

Murine P19 cells were cultured in alpha-MEM containing nucleosides supplemented with 10% (v/v) FBS, 100 mg/mL each of penicillin and streptomycin at 37˚C with 5% CO_2. Cells were cultivated on Gold Quantifoil grids (R2/1, Au 200 mesh grid, Quantifoil Micro Tools, Jena, Germany). Additional carbon (20–25 nm) was deposited on the film side of the grids in a carbon evaporator (MED 020, BAL-TEC) and plasma cleaned for 45 s prior to their use. Next, grids were sterilized under UV irradiation for 30 min and immersed in culture medium in a CO_2 incubator for 30 min. Cells were detached from cell culture flasks using 0.05% trypsin-EDTA, seeded on 4-6 pre-treated Quantifoil grids in 35-mm dishes and kept in an incubator overnight to allow adhesion.

Cryo-ET of synaptosomes and P19 cells

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, or the concentrated BSA-gold (Aurion) was mixed with synaptosomes at 1:10 volume ratio. A 3-μl drop of synaptosomes was placed onto the grid, blotted with filter paper (GE Healthcare), and plunged into liquid ethane.

Grids with P19 cells were blotted from the reverse side by placing a Teflon sheet on the front and vitrified by plunging into a liquid ethane/propane mixture at liquid nitrogen temperature using a Vitrobot Mark 4 (FEI Company, Eindhoven, Netherlands). The Vitrobot was set to 37˚C, 90 % humidity, blot force 10, blot time of 10 s and 2 s drain time.

Tilt series were collected under a low dose acquisition scheme using SerialEM [44, 45] 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 size was 0.34 nm at the specimen level. Volta phase-plate with nominal defocus between -0.5 and -1 μm [46] was used. The total dose was kept at 60-100 e⁻/Å². Individual frames were aligned using Motioncor2 [47] and dose filtered [48] (P19 cells only). Tilt series were aligned using gold beads as fiducial markers (synaptosomes) or performed by patch-tracking (P19 cells) and 3D reconstructions were obtained by weighted back projection (WBP) using Imod [49]. During reconstruction, the projections were binned once (final voxel size of 0.68 nm) and low pass filtered at the post-binning Nyquist frequency (synaptosomes only).

Procedure overview

The complete procedure implemented in PySeg package contains the following steps:

1. Discrete Morse theory-based density tracing and particle picking

   - Tracing of density by the Discrete Morse theory based algorithm (DisPerSe)
   - Simplification by topological persistence
   - Spatially embedded graph representations of the EM density
   - Selection of complexes - particle picking

2. General classification based on Affinity propagation

   - Determination of membrane normal vectors
   - Constrained refinement (Relion)
   - Unsupervised classification of rotationally averaged complexes by Affinity propagation clustering

3. Spatial analysis and averaging

   - Standard 3D classification and constrained refinement (Relion)
   - Spatial distribution analysis within or between classes

Density tracing and particle picking

Electron densities were traced using DisPerSE software package which is based on discrete Morse theory (see [24] for outline of DisPerSE, [23] for a more rigorous presentation of discrete Morse theory and [50] more generally for mathematical topology). Briefly, in Morse theory, a Morse function is defined on a *n*-dimensional manifold. A critical point of the Morse function (points where gradient is 0) that has order *k* has minima in *n-k* and maxima in *k* directions. Gradient paths starting and ending a critical point of order *k* define ascending and descending manifolds of dimensions *n-k* and *k*, respectively. These include local minima, maximum gradient paths connecting critical points and 3D catchment basins. Intersections of ascending and descending manifolds determine Morse-Small cells of dimension *0* to *n*, which partition the space according to the gradient paths and provide connectivity information between critical points. Importantly, 1-critical points are always connected by arcs to two minima. In our case, 3D tomogram greyscale values were used to define a Morse function. Of particular interest are 0-critical points (local minima, Morse-Small *0*-cells), the Morse-Small *1*-cells, that is maximum gradient arcs that connect *0* and *1*-critical points (minima and saddle points that have minima in two and a maximum in one direction). In other words, these trace local density maxima and the "most dense" paths between the densities, thus providing a network that represents EM density. The discrete Morse theory is defined in a similar manner, except that a Morse function is defined on a simplical complex (instead on a manifold), in our case the 3D voxel-based Cartesian grid. The Morse function is then used to determine critical *k*-dimensional simplices, ascending and descending manifolds, and discrete Morse-Small cells.

We modified the topological persistence simplification method and implemented it in PySeg package because the complexity of cryo-tomograms precluded the use of the Morse complex simplification procedure implemented in DisPerSE. 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. Formally, this method can be presented as the following algorithm:

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

This is equivalent to introducing small perturbations in the greyscale values that remove some minima, effectively reducing the contribution of noise. To determine the persistence thresholds, we run multiple simplifications to find the value that gives a specified surface density of minima on membranes. In this way, the density detection is normalized across tomograms. In the same way, (volume) density of minima could be used, so this method is independent of membranes. Furthermore, because the minima are counted, this normalization depends on the total density rather than on the number of complexes of interest. The number of arcs is kept to be twice the number of minima.

Density tracing information (specified in the form of Morse complexes) is converted to spatially embedded graphs in the following way. Each arc is assigned to a graph edge and its two adjacent minima are assigned to graph vertices associated with the edge. Vertices and edges keep the greyscale density and a precise location of the underlying minima and arcs. Furthermore, geometrical information, such as the Euclidean and the geodesic lengths of arcs are associated with edges. Additionally, these graphs may also contain external information provided by segmentation of large cellular structures, such as lipid membranes, organelles or cytoskeleton.

Creation of spatially embedded graphs from Morse complexes and their manipulation was implemented in PySeg. For some of the standard graph tasks, the graph-tool library was used [51]. Methods to query properties associated with graph vertices, and edges and methods to extract particles were also implemented in PySeg. All together, representing density by spatially embedded graphs significantly increased the computational efficiency.

Density tracing in phantom data

The phantom dataset contained a 6x6x3 grid with a 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 using Morse density tracing. 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. 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). Numbers of TPs, FPs and TNs were normalized to the total number of the corresponding 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.

3D classification and refinement

3D classification and refinement steps were performed in Relion [21]. 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. Except for the validation of particle picking, 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) to remove the missing wedge, that is no external reference was used.

The constrained refinement and 3D classification were 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 the 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.

General classification

The general classification was done using the AP clustering algorithm [25]. In short, the algorithm separates elements (here particles) in clusters and determines the representative of each class ("exemplar"). It is based on the iterative calculation of two properties for each pair of elements: the responsibility quantifies how well suited is the first element to be the cluster representative of the second element, while the availability shows how appropriate it is for the first element to have the second element as the representative of its class. In general, initial preferences of each element to be a cluster representative are specified as input parameters. In all our applications, these preferences are taken to be the same for all elements, thus leaving only one parameter. This parameter influences the number of clusters found (the larger the preference the more clusters), but it is not possible to tune the value of the preference so that the classification yields a specified number of classes.

For classification by AP, particles were rotationally averaged around membrane normals by computing mean greyscale values of 2 pixel-wide rings around the normal vectors and the resulting rotational averages were normalized to a density mean of 0 and standard deviation of 1. As an alternative method, weighting the rotational averages by the distance to the axis (r-weighting) emphasizes off-axis elements. For the classification, the similarity between two particles was defined as the dot product between the rotational averages. Unless noted otherwise, the masks used for the AP classification cylindrical.

Rotational averaging of subtomograms around the membrane normal vectors and an interface for the scikit-learn [52] implementation of the AP algorithm are provided in PySeg.

Validation of general classification

To generate a realistic set of particles, eight available high resolution structures of membrane-bound complexes of different size and shape (4UQJ, 4PE5, 5IDE, 5GJV, 5KXI, 5TJ6, 5TQQ and 5VAI [53, 54, 55, 56, 57, 58, 59, 60]) were selected as the reference structures. First, they were low-pass filtered at 0.524 nm, random noise was added to SNRs of 0.05, 0.01 and 0.005. Particles were randomly tilted by 0 - 20° with respect to the membrane normal and in the other two directions (full range) to impose the missing wedge of $\pm 30°$, tilted up to 5° (datasets that include membranes), rotated around membrane normals, and displaced along the membrane 0-9 pixels. To test specific parameters, the following test datasets were created from these particles: (i) one for each SNR (Figure S4A), (ii) one for each displacement range (0-3, 0-6 and 0-9 pixels; Figure S4B) at 0.01 SNR, (iii) derived from 4-8 structures at 0.01 SNR (Figure S4C). (iv) For the comparison between the clustering methods, all parameters were randomized, including the number of particles obtained from each reference structure (800 particles in total), and SNR was between 0.01 and 0.005 (Figure 2A). The last set was also used for the clustering methods comparison with r-weighted rotational averages (Figure S4D). In addition, we generated datasets similar to the dataset iv that included membrane and contained 800 - 4000 particles (used for Figure S5). The class averages shown were obtained by one round of the constrained refinement using Relion. Except when noted, 8 structures were used and 100 particles were generated for each structure. For each dataset, the corresponding ground truth classes contained particles obtained from one reference structure.

Particles were rotationally averaged around the normals (Figure S3B) and the cross-correlation between all particle pairs was calculated, to be used as the clustering distance. K-means classification was preceded by principal component analysis where the eigenvectors corresponding to the eigenvalues that contributed to 95% of the variability were kept. Each K-means clustering was repeated 200 times and the mean and the standard deviations were calculated. We optimized the agglomerative (ascendant) hierarchical clustering on our test datasets to produce the most similar results to the known classes, according to the Fowlkes-Mallows, Variation of information and F_1 measures. Specifically, we found that it is necessary to reduce the dimensionality of the data by principal component analysis, which was done in the same way as for K-means clustering. Furthermore, Ward linkage, which minimizes the variance increase at each merging step, turned to be superior to other linkages. In all cases, it was necessary to specify an additional parameter, the number of classes or the maximal distance for merging clusters, as required by the implementation in scikit-learn package [52].

Classifications were evaluated by comparing them with the ground truth using the following three measures. Fowlkes-Mallows [26] and the Variation of information [27] were implemented in Pyto package as previously reported [17]. To calculate F_1, for each evaluated class we determined the number of particles belonging to

the reference classes. The reference class that contributed the highest number of particles was then associated to the evaluated class, thus defining a mathematical map from the evaluated to the reference classes. We note that this mapping does not have to be one-to-one or onto. Nevertheless, it allows us to calculate the following properties for each reference class i :

- True positives (TP_i): Number of elements of all evaluated classes associated with reference class i that belong to the class i

- False positives (FP_i): Number of elements of all evaluated classes associated with reference class i that do not belong to the class i

- False negatives (FN_i): Number of elements of all evaluated classes not associated with reference class i, that belong to the reference class i

Now we can define precision (P), recall (R) and the F_1 measure:

$$P = \frac{1}{N_R} \sum_{i=1}^{N_R} \frac{TP_i}{TP_i + FP_i}$$

$$R = \frac{1}{N_E} \sum_{i=1}^{N_R} \frac{TP_i}{TP_i + FN_i}$$

$$F_1 = \frac{2\,P\,R}{P + R}$$

where N_R and N_E and the number of reference and evaluated classes, respectively. In this way, persistence and recall show the expected behavior, that is precision increases and recall decreases with the number of classes (Figure S4A and B).

Processing of microsomal complexes

The processing work-flow is schematically shown in Figure S7.

55 out of 210 the tomograms previously obtained from canine pancreatic microsomes [28] were used in this study. From these tomograms (1.048 nm pixel size), 166 microsomal membranes were segmented using the automated procedure TomoSegMemTV [61]. We note that other available software, or manual membrane segmentation could be also used for this step. 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 topological simplification were performed in 3D by our procedure as described above. To account for different quality of tomograms, the persistence threshold was set so that the density of vertices (minima) on all microsomal membrane was 0.0025 nm^{-3}.

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 connected membrane vertex was detected, and the intersection of this path with the membrane interface was chosen to represent a particle, 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 [49] at a pixel size 0.524 / 0.262 nm and with a box size 120 / 160 pixels (for the cytoplasmic / lumenal sides).

Further processing was done essentially in the same way for the cytoplasmic and lumenal particles, as follows. Particle position and normal vectors were optimized using Relion by the constrained refinement with C10 symmetry. They were rotationally averaged, and the general classification was performed using the AP clustering, yielding about 100 classes. The initial preference was set to -10, to prevent getting a large number

of classes. The resulting classes were visually inspected to select the cytosolic class showing the best resolved cytoplasmic and lumenal densities, and the lumenal class showing the best lumenal density located at the distance to the membrane. These classes were averaged by constrained refinement to yield densities to be used as initial references for the further processing.

Unless noted otherwise, the masks used for the AP and for the following 3D 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.

We then performed three rounds of 3D classification. 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 first variant ("bulk cleaning"), 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) (Figures 3F, S8, S7 dotted line). In the second variant ("AP cleaning", used for cytosolic particles only), each AP class was subjected to 3D classification and the best classes were selected (7100 cytosolic particles from 15 AP classes) (Figures 3E, S7 dashed line). 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 2nd and 3rd classification steps were performed without alignment, using the alignment parameters from the first 3D classification round (the masks used are shown in Figure 3 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 3 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 3F). 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 (Figure S6).

Validation of density tracing and particle picking on microsomal membranes

For this validation, we analyzed 6 out of 55 microsomal tomograms that we used before. The Morse tracing and picking was done in the same way as before. For local density detection, tomograms were low-pass Gaussian filtered at 5 nm and density maxima were chosen. Template matching was done using Pytom software [62], our ribosome-free translocon complex average (Figure 3F) low-pass filtered at 2.5 nm served as the template and the angular increment was 12.85°. In all cases, particles were picked on the lumenal side, up to 20 nm from the membrane, at 1.05 nm pixel size, yielding 13 000 - 17 000 particles. Further processing was done with Relion at 0.26 nm pixel size. All classes were first subjected to the constrained refinement with C10 symmetry and then 3D classified into five classes with constrained alignment, using the same ribosome-free translocon average, filtered at 3.0 nm, as the initial reference.

Detection and classification in situ

Twelve tomograms of intact P19 cells that contained ER were used here. Membranes were segmented with TomoSegMemTV [61] without any manual intervention. Consequently, the lumenal and cytosolic sides were not distinguished in the subsequent analysis. Tomograms were low-pass Gaussian filtered at $\sigma = 0.8$ pixels, the density was traced using our procedure, vertex density on membranes was 0.0035 nm^{-3} (all at 1.368 nm pixel size). Particles were picked in the region 3-10 nm away from membranes, having geodesic distance to the membrane of 3-30 nm. The exclusion distance was 5 nm to allow a rather comprehensive picking of small complexes. All together, 172 000 particle subvolumes were reconstructed at 0.342 nm pixel size, with box size of 160 pixels.

Particle position and normal vectors were optimized using Relion by the constrained refinement with C10 symmetry. They were rotationally averaged and the dot product between them was used as the similarity measure for the subsequent three rounds of AP classification, the initial preference parameter was -6.

The first AP classification (at 0.684 nm pixel size) served to discard classes that showed two, or did not show any well resolved membrane (34 000 particles), and to group other classes into datasets defined by the following features: large densities on the particle side of the membrane (28 000 particles), large densities on the side opposite from the particles (25 000 particles) and the small densities (85 000 particles) (Figures S9, S10). In the second round (at 0.342 nm pixel size), these three datasets were AP-classified separately. Classes were were visually evaluated and grouped for further processing based on desired features (ribosome-containing and other classes from the large densities on the particle side dataset, and ribosome containing classes from the large densities on the opposite side dataset), and some classes were discarded. The selected particles were subjected to the constrained refinement with C10 symmetry by Relion, to optimize their positions and normals. The mask used for the first two AP rounds contained membranes and the region on both sides of the membrane. The third round of AP-classification (at 0.342 nm pixel size) was performed using a mask that did not contain the membrane, to improve the position and orientation of particles. Selected classes were subjected to constrained refinement, and in some cases also to 3D classification, by Relion. The third AP classification was not needed for classes containing ribosomes, from the large density on the particle side dataset.

Averages of selected classes were obtained by constrained refinement using Relion. In some cases the refinement was preceded by 3D classification. No symmetry was used except that C4 was used for the class tentatively identified as the IP3 receptor. No external structures were used, all initial references were provided from the AP classes. Final numbers of averaged particles are: 875 ribosomes, 180 ribosome-associated translocons, 63 ribosome-free translocons, 41 IP3 receptors, and 108 and 177 PLC complexes. Overlapping particles were removed. All masks used were cylindrical. We note that because several hundred different proteins are expected to reside on ER, many of the picked particles that were not included in the final averages likely represent different complexes.

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:

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

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 [34, 35]. 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 that contained at least 10 ribosomes (122 out of 166) 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 1220 in 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).

General software methods

The software developed for the work presented here 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).

PySeg uses Pyto package [17], Numpy package, surface meshes were stored using VTK [63] and graphs are plotted using *matplotlib* library [64]. We parallelized (shared memory model) some of 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 [65] and the UCSF Chimera package from the Computer Graphics Laboratory, University of California, San Francisco [66] software packages were used.

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

Statistics

K-means clustering for the validation of general classification were repeated 200 times for each data point (number of classes) because of the random nature of the initialization (Figures 2A; S4D).

For the determination of the statistical significance of the spatial distribution functions results, 1220 randomly distributed particle sets were generated (Figures 5B-D; S11).

Data availability statement

The following EM densities have been deposited in the EMDataBank. Subtomogram averages obtained from the microsomal dataset, 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 microsomal dataset 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). Microsomal tomograms EMD-10449, EMD-10450 and EMD-10451. Subtomogram averages obtained from the in situ P19 cells dataset: membrane-associated ribosome (EMD-10432), ribosome-associated translocon (EMD-10433), ribosome-free translocon (EMD-10434), two putative PLC complexes (EMD-10435, EMD-13436), putative IP3 receptor (EMD-10437). In situ P19 cells tomogram shown on Figure 4 EMD-10439).

Source data for Figure 2 is available with the paper. Data for Supplementary Figure S4 are available from the corresponding authors upon request.

Code availability statement

The complete software, together with all dependencies, is installed as PySeg capsule on Code Ocean [67]. The software is also available upon demand.

Supplemental Data

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

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

Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

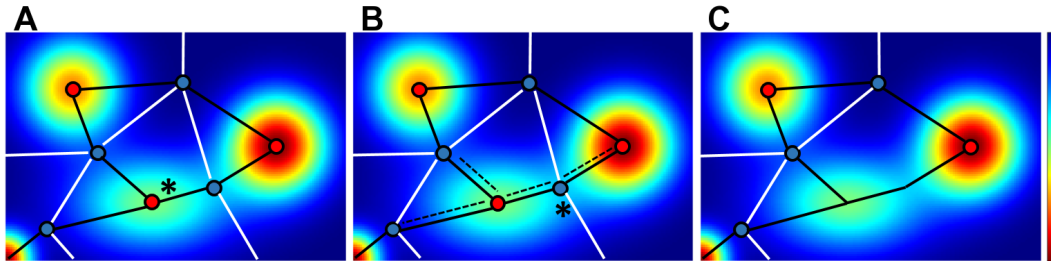[*]Corresponding authors; AM-S martinez@biochem.mpg.de and VL vladan@biochem.mpg.de

Supplementary figure 1

Figure S1: 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 approximate borders between ascending 2-manifolds.

Supplementary figure 2
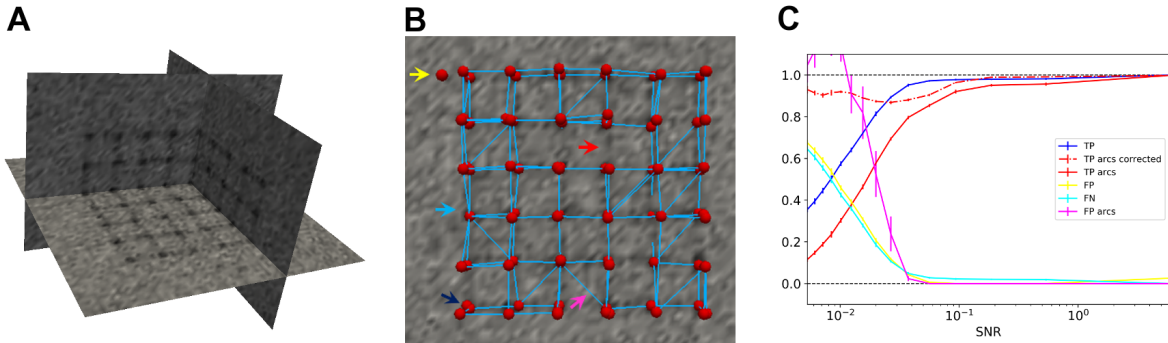
Figure S2: Validation of the density tracing and the simplification on phantom data. **A** Tomographic slices of the phantom data set. **B** Detected minima and arcs at SNR = 0.025 in the graph representation superposed on a slice of the simulated tomogram. Red spheres denote detected minima (vertices) 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).
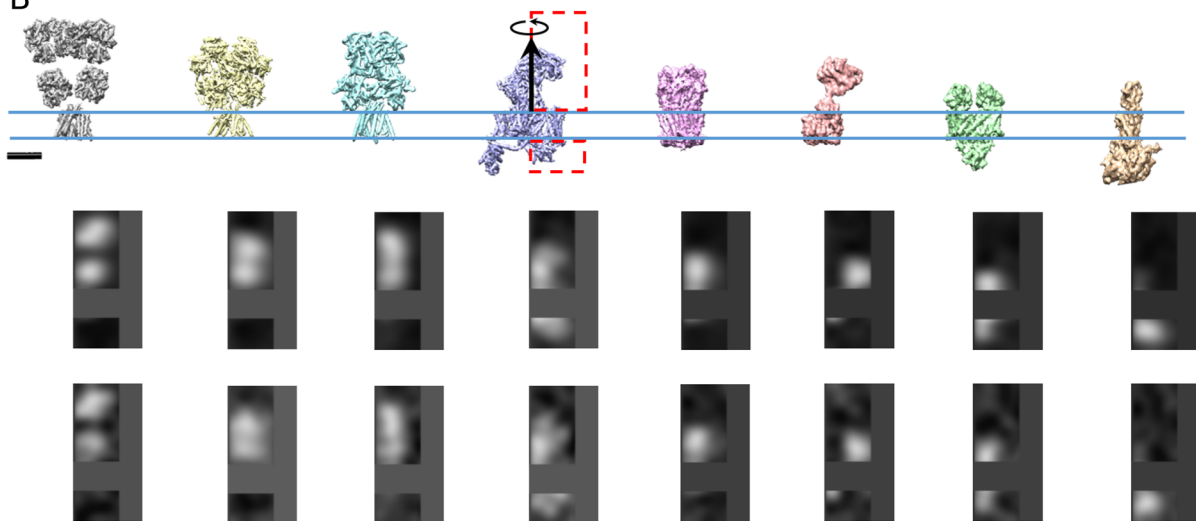
Supplementary figure 3

A
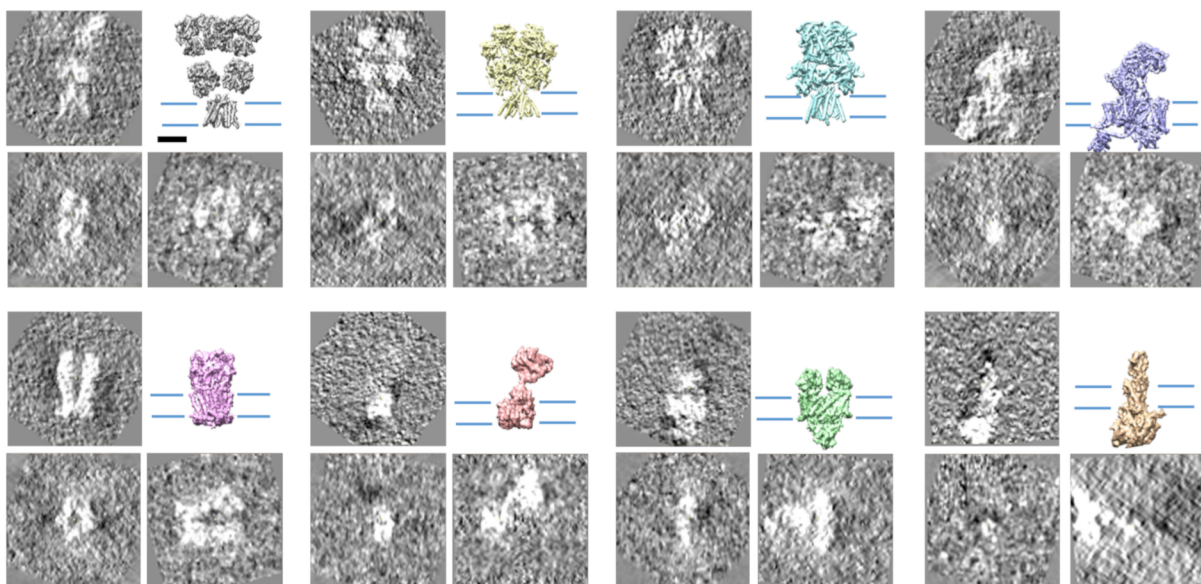
B

Figure S3: AP classification of test datasets that does not include membranes. A Examples of test particles (x-, y-, and z-slices) together with the corresponding high-resolution reference structures at SNR=0.05. B High resolution structures used to make test data sets (top row), the corresponding AP class averages (middle row) and the AP class representatives (bottom row). The red rectangle show the position of the mask. Scale bars 5 nm. N=800 particles
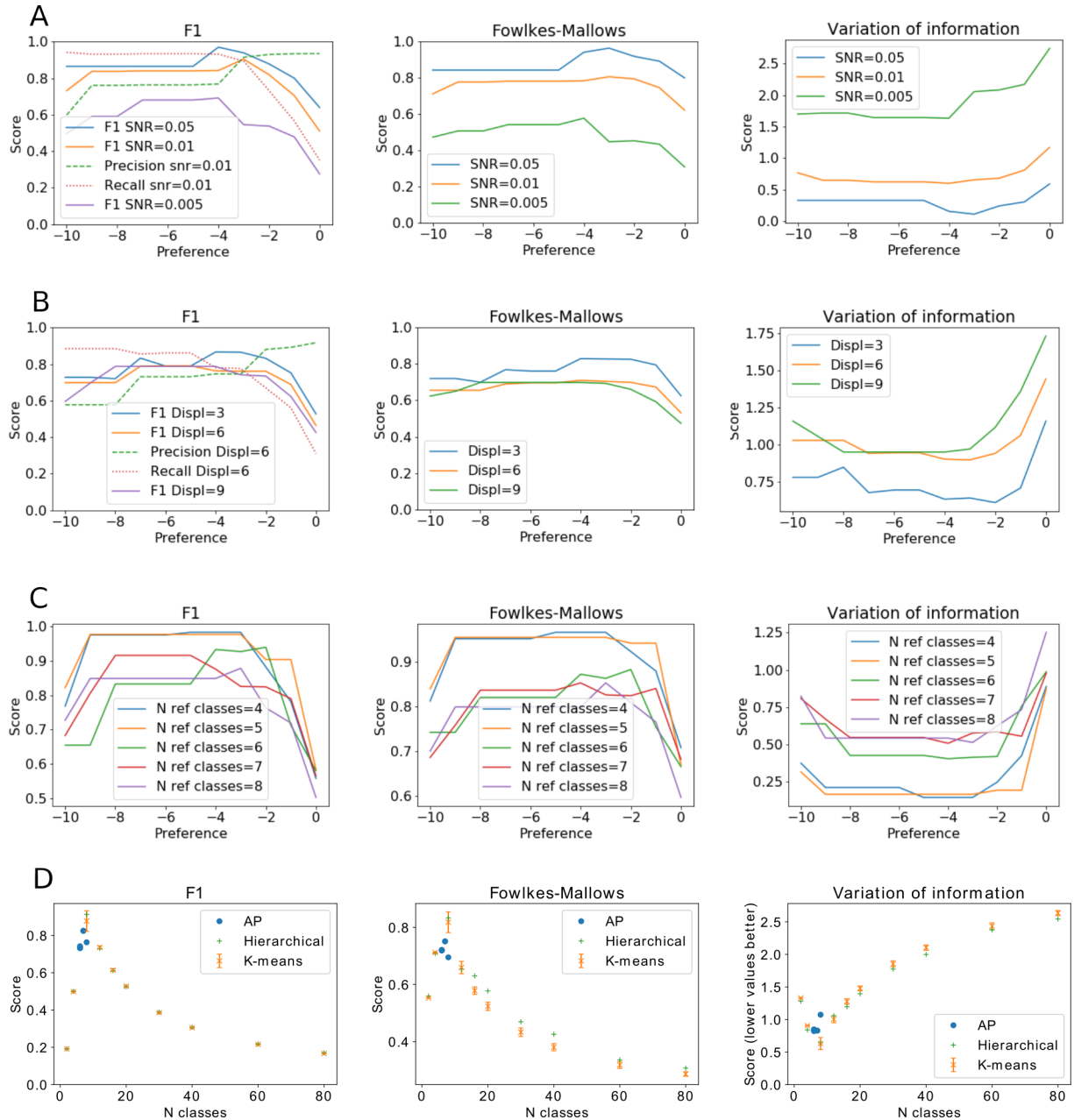
Figure S4: Evaluation of AP classification against the ground truth using F_1 (left column), Fowlkes-Mallows (middle column) and Variation of information (right column) measures. A Evaluation of AP classification at different values of the preference parameter, on datasets having different SNR's. In addition to F_1, the graph on the left shows precision and recall at SNR=0.01. B The same as (A) except that the three datasets used had displacement range along the membrane of 0-3, 0-6 and 0-9 pixels (labeled as 3, 6 and 9, respectively). C The same as (A) except that the datasets contained different number of reference classes (4-8). D Evaluation of AP, K-means and hierarchical clustering. The same as Figure 2A of the main text except that here particles were rotationally averaged with r-weighting. The AP clustering was performed for the preference parameter taking all integer values between -9 and -3 and the resulting number of classes is shown on the x-axis (note that several data points overlap). The number of classes was set as a parameter for K-means and hierarchical clustering, as shown on the x-axis. For K-means clustering, the data points denote means and the error bars the standard deviation obtained from 200 simulations. Higher values of the Fowlkes-Mallows and F_1 measures, but lower values of the Variation of information signify better agreement. The number of classes increases with the preference parameter.
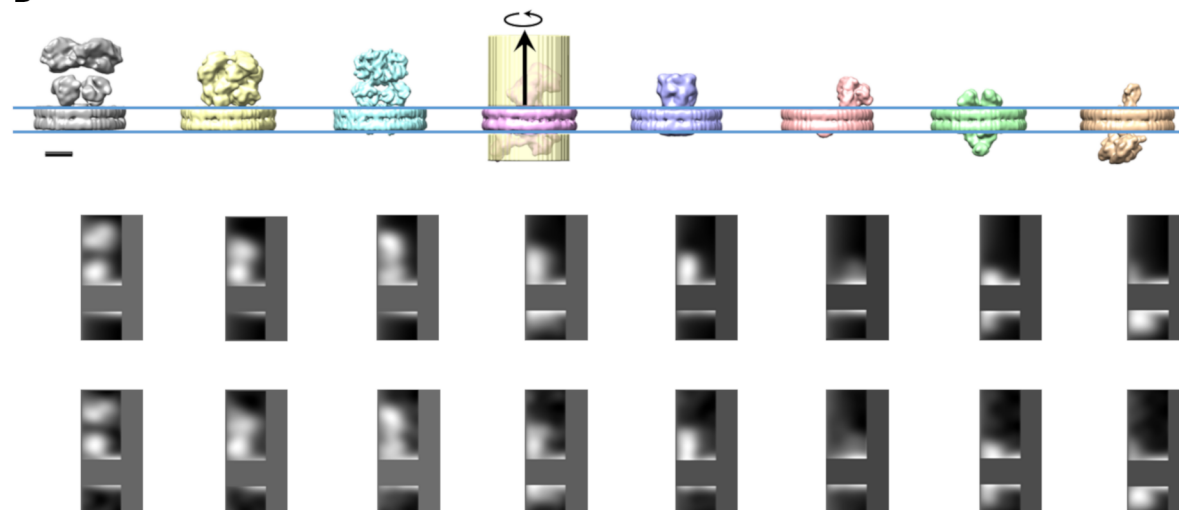
A
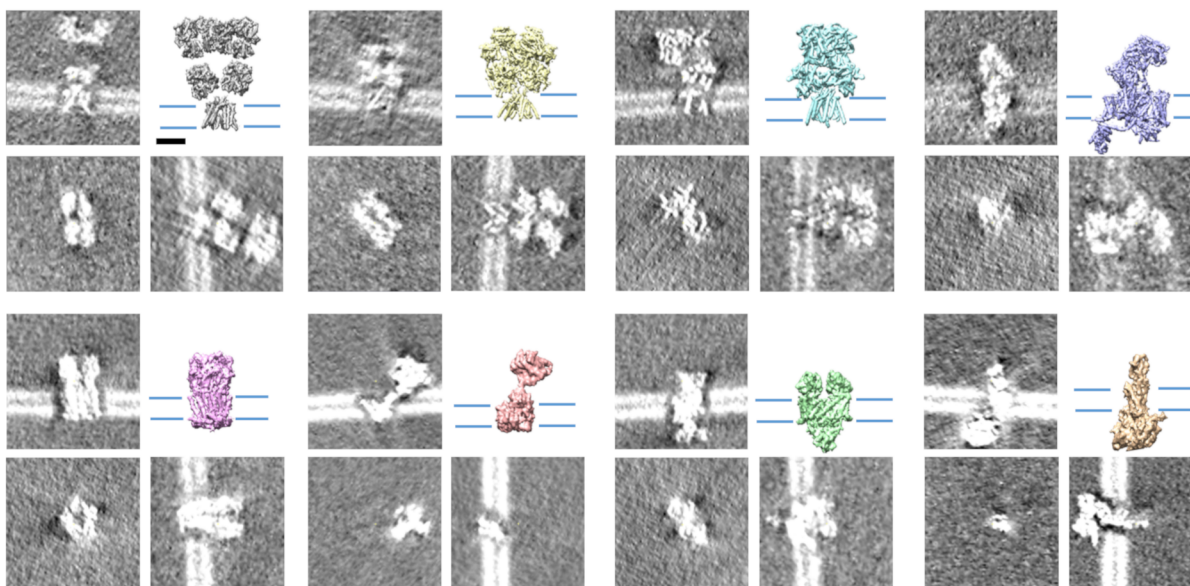
B

Figure S5: AP classification of test datasets that include membranes. A Examples of test particles (x-, y-, and z-slices) together with the corresponding reference structures at SNR=0.05. B 3D structures obtained by constrained refinement of AP classes (top row), the corresponding AP class averages (middle row) and the AP representatives (bottom row). Scale bars 5 nm. The yellow cylinder shows the mask. Scale bars 5 nm. N=4000 particles.
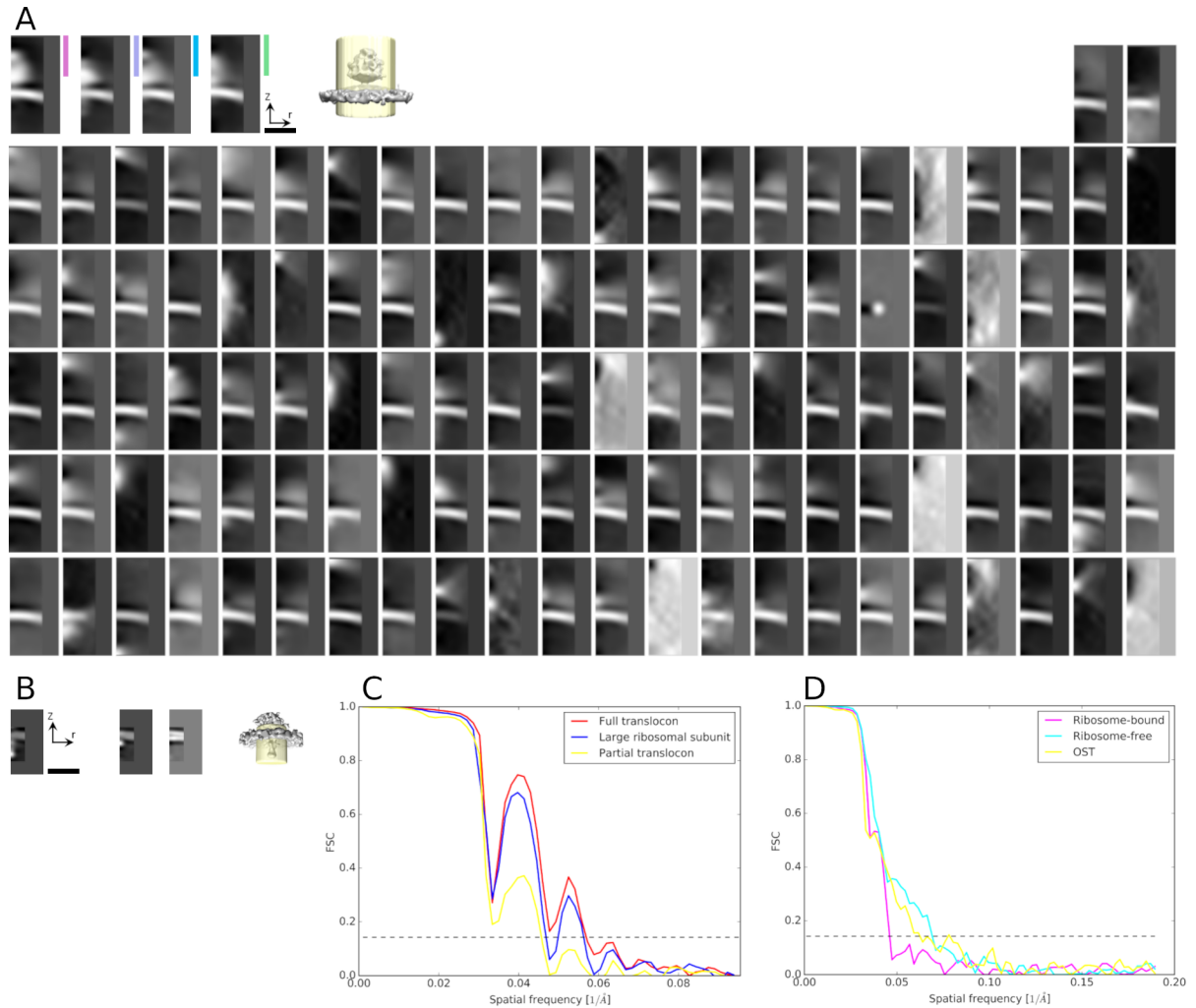
Supplementary figure 6

Figure S6: Processing of microsome-attached complexes. A AP classification of cytosolic particles (ribosomes). All 2D class averages are shown. The four color bar labeled images are from the same classes as those shown in the same colors on Figure 3C, D. The class shown in the upper left corner was used to generate the initial reference for the subsequent 3D classification. 3D class average obtained without alignment is shown on the right, the mask is shown in transparent yellow. B AP classes of lumenal particles (translocons). 2D class averages are shown on the left. The class shown on the left was used as the initial reference for the subsequent 3D classification. The two 2D classes on the right are examples of bad classes because they do not show lumenal density at a distance from the membrane. 3D class average obtained without alignment is shown on the right, the mask is shown on the right in transparent yellow. C FSC plots of the 3D refined ribosomal class averages obtained using the AP cleaning variant (see Figure S7), shown on Figure 3E (colors match). D FSC plots of the 3D refined translocon class averages shown on Figure 3F (colors match). FSC plots in C and D were obtained by the AP cleaning variant. Scale bars 20 nm.
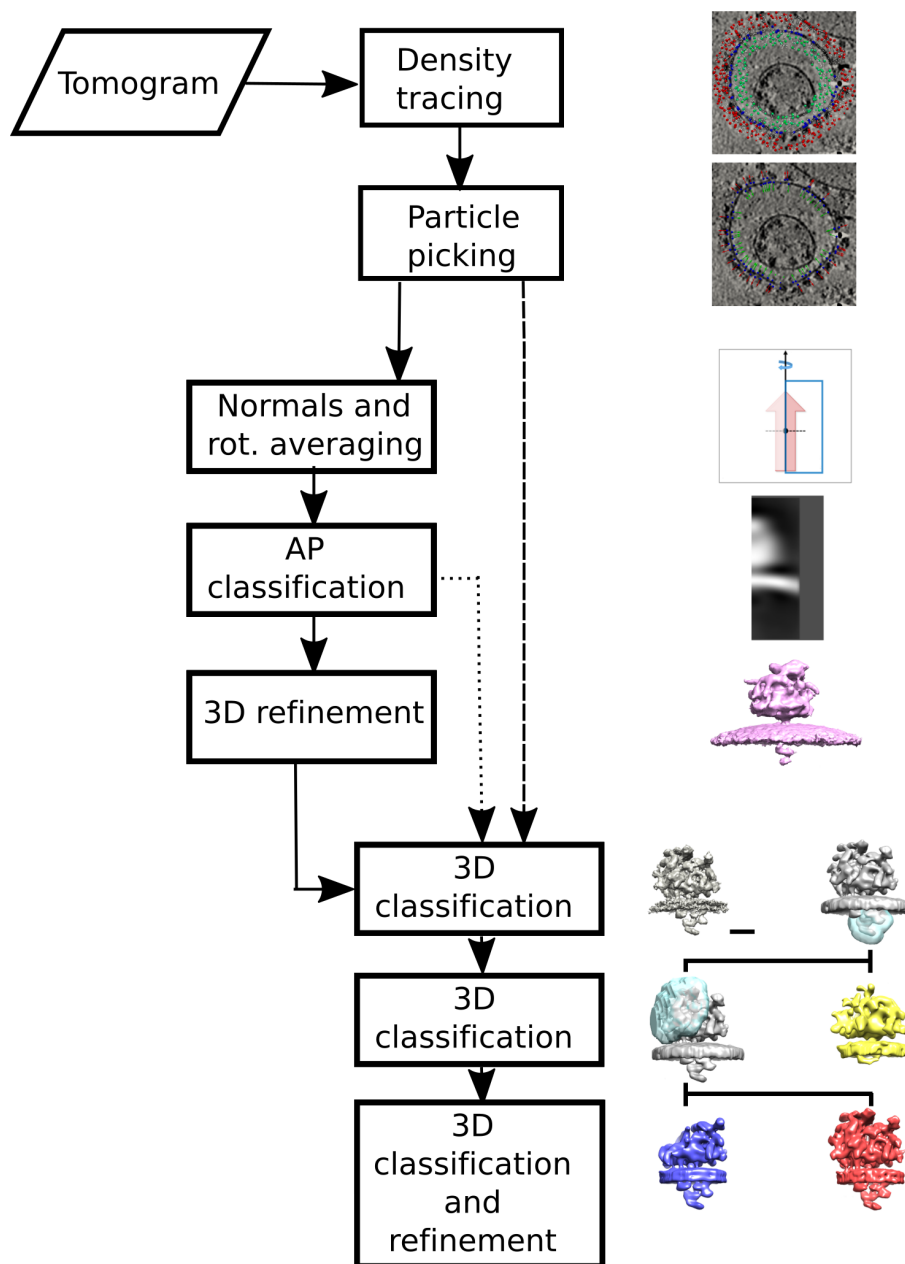
Figure S7: Flowchart of the processing steps used for microsome-attached complexes. Rectangles show the computational steps. Solid lines and the dotted line together show the processing flow that includes the AP cleaning variant of the first 3D classification step. The dashed line together with the solid lines show the bulk cleaning variant. Images on right side graphically represent results of the corresponding processing steps.
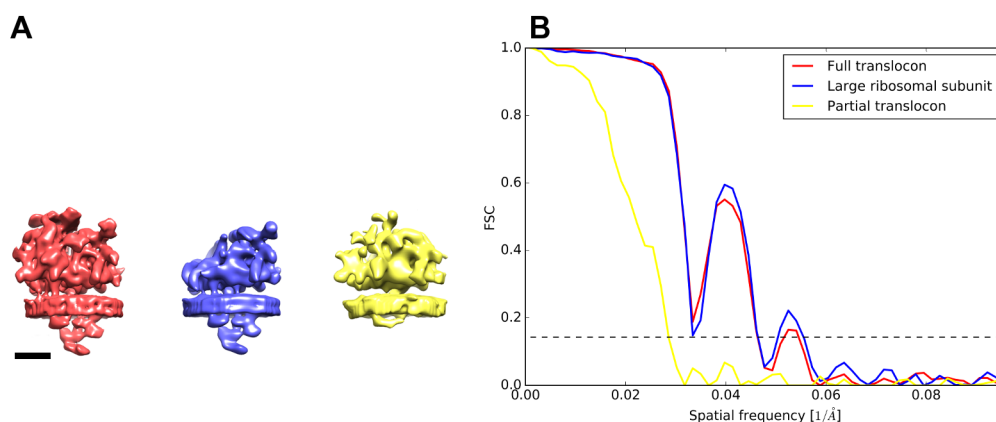
Supplementary figure 8

Figure S8: Processing lumenal particles using the bulk cleaning variant (see Figure S7). A 3D densities of fully assembled (left) and partial translocon complex (right), as well as the large ribosomal subunit bound to the fully assembled translocon (middle). B FSC plots of the classes shown in A. Colors correspond to the densities shown in Figure 3E. Scale bar 10 nm.
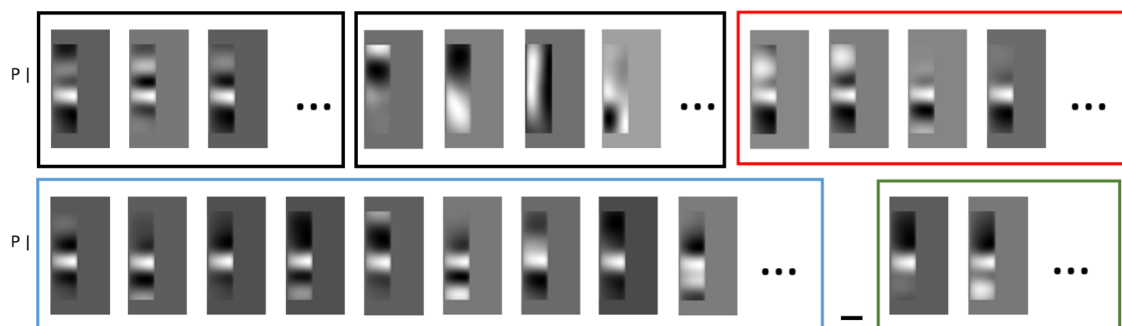
Supplementary figure 9

Figure S9: Examples of class averages obtained after the first AP classification round of in situ complexes. Vertical lines on the left labeled with "P" show the level where particles were picked. The two groups boxed in black show classes containing double membranes (left) and classes not containing a clear membrane (middle). The group boxed in red is shown in Figure S10A, blue in Figure S10B and green in Figure S10C. Scale bar 10 nm.
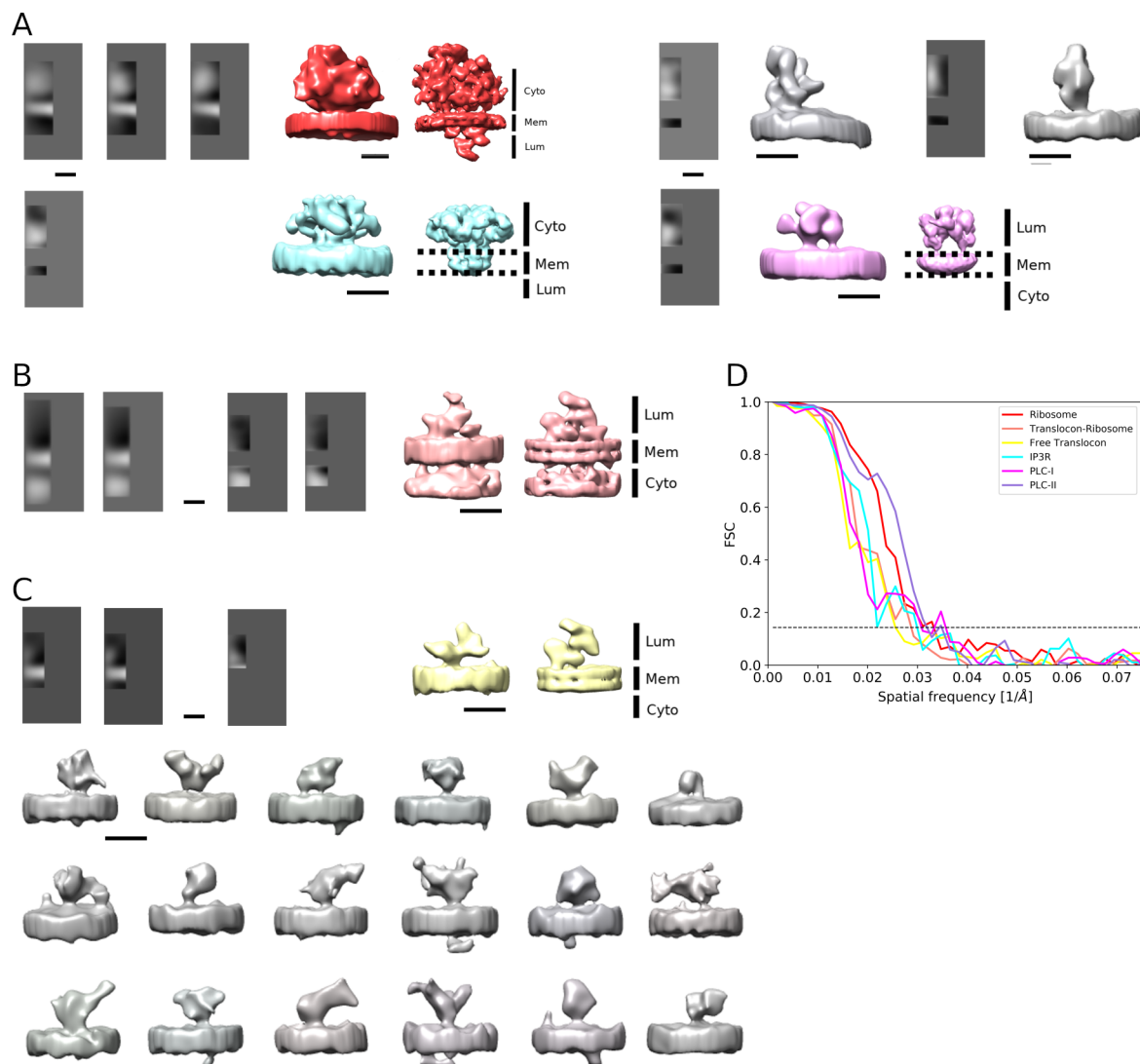
7

Supplementary figure 10

Figure S10: Classification and refinement of *in situ* complexes. A The third AP classification and refinement of the dataset featuring large densities on the particle side of the membrane. Top left: ribosomes (AP averages, *in situ* 3D average EMD-10432 and the 3D average obtained from microsomes EMD-0074). Top right: two unidentified structures (AP average, *in situ* 3D average). Bottom left: Putative IP3 receptor (AP average, *in situ* 3D average EMD-10437 and EMD-6369). Bottom right: putative PLC (AP average, *in situ* 3D average EMD-10435 and EMD-3905). B Dataset featuring large densities on the side opposite of the particle. Ribosome-bound translocon: the second and the third AP classifications (two classes each), 3D *in situ* average EMD-10433 and the 3D average obtained from microsomes EMD-0085. C Dataset showing small densities. Top row, ribosome-free translocon: the second (two classes) and the third AP classifications, 3D *in situ* average EMD-10434 and the 3D average obtained from microsomes EMD-0087. Below: a gallery showing other average densities from the same dataset. *In situ* 3D averages are also shown on Figure 4, 3D average obtained from microsomes are also shown on Figure 3. D FSC plots of the average densities shown in color. All scale bars 10 nm.
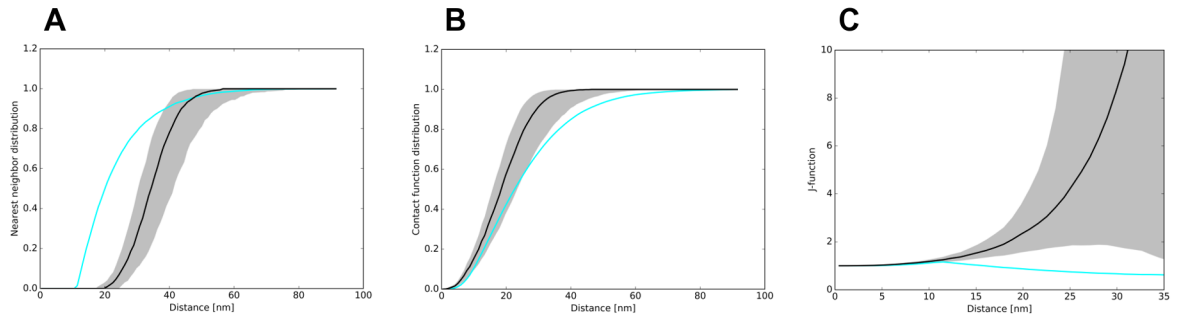
Supplementary figure 11

Figure S11: Spatial distribution analysis of microsomal ribosome-free full translocon particles by the first order distribution functions. A Nearest neighborhood distribution. B Contact distribution. C J-function. Blue lines show the data from experimental particles. Black lines show the median of the respective functions for a set of random particle distributions (N=1220 random distributions) and the gray areas represent regions of p>0.05 confidence.

Video 1: Detection of microsome-attached complexes and localization of lumenal particles (tomogram EMD-10450). Density minima are shown as small spheres (red: cytosolic; blue membrane; green lumenal), arcs as grey lines and green arrows denote membrane normal vectors. N=55 tomograms.

Video 2: Localization of microsome-attached complexes (tomogram EMD-10450). Ribosomes derived from the cytosolic particles are shown in red, ribosome-free fully assembled translocon complexes in blue and the non-translocon associated OST complexes in yellow. N=55 tomograms.