Deep Learning Improves Macromolecules Localization and Identification in 3D Cellular Cryo-Electron Tomograms

Emmanuel Moebel[1], Antonio Martinez-Sanchez[2], Damien Larivière[3], Eric Fourmentin[3], Julio Ortiz[4], Wolfgang Baumeister[2], Charles Kervrann[1]

[1] Serpico Project-Team, CNRS-UMR 144, Inria, CNRS, Institut Curie, PSL Research University, Campus Universitaire de Beaulieu, 35 042 Rennes Cedex, France
[2] Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
[3] Fourmentin-Guilbert Scientific Foundation, 2 Avenue du Pavé Neuf, 93 160 Noisy-le-Grand, France
[4] Ernst Ruska-Centre, Wilhelm-Johnen-Straße, 52425 Jülich, Germany

abstract

Cryo-electron tomography (cryo-ET) allows one to visualize and study the 3D spatial distribution of macromolecules, in their native states and at nanometer resolution in single cells. While this label-free cryogenic imaging technology produces data containing rich structural information, automatic localization and identification of macromolecules are prone to noise and reconstruction artifacts, and to the presence of many molecular species in small areas. Hence, we present a computational procedure that uses artificial neural networks to accurately localize several macromolecular species in cellular cryo-electron tomograms. The DeepFinder algorithm leverages deep learning and outperforms the commonly-used template matching method on synthetic datasets. Meanwhile, DeepFinder is very fast when compared to template matching, and is better capable of localizing and identifying small macromolecules than other competitive deep learning methods. On experimental cryo-ET data depicting ribosomes, the localization and structure resolution (determined through subtomogram averaging) results obtained with DeepFinder are consistent with those obtained by experts. The DeepFinder algorithm is able to imitate the analysis performed by experts, and is therefore a very promising algorithm to investigate efficiently the contents of cellular tomograms. Furthermore, we show that DeepFinder can be combined with a template matching procedure to localize the missing macromolecules not found by one or the other method. Application of this collaborative strategy allowed us to find additional 20.5% membrane-bound ribosomes that had been missed or discarded during manual template matching-assisted annotation.

Introduction

The last decades of research in cell biology have revealed that cellular processes are performed by groups of interacting macromolecules in a crowded environment. This is in contradiction with the previous cell models where macromolecules were considered as isolated objects floating randomly in the cytoplasm. Deciphering the underlying interaction mechanisms is thus of paramount importance to gain a deeper understanding of the cell. To address this issue, cryo-electron tomography (cryo-ET) is generally considered because it is capable of producing 3D views of large portions of a cell while having enough resolution to localize and identify macromolecules. The localization and identification results can be exploited to build a complete molecular atlas of the cell. One advantage of cryo-ET is to avoid the use of markers, such as fluorescent probes in light microscopy, which generate damages in the cell. The conventional protocol for image acquisition in cryo-ET can be summarized as follows. In a preliminary step, the samples are vitrified in order to preserve both the spatial distribution and the structure of macromolecules in the cell during the image acquisition process. A set of tilted 2D images of the same specimen rotated around an axis perpendicular to the electron beam serves as inputs to 3D reconstruction algorithms[1;2;3]. Unfortunately, due to technical limitations, the specimen cannot be rotated up to ± 90 degrees, and acquisition is typically restricted in the range ± 70

degrees at best. The resulting 3D volumes are then severely corrupted by noise, and reconstruction artifacts due to the missing wedge of data in Fourier space. Consequently, one typically observes undesirable elongated structures along the Z-direction, and also a loss of information in the X–Y plane[4;5], along the direction perpendicular to the tilt axis. For all these reasons, the identification of macromolecules in experimental tomograms becomes very challenging and requires powerful and sophisticated algorithms.

A well-established method in cryo-ET for localizing macromolecules is based on template matching[6] (TM). A fine-detailed template depicting the macromolecule of interest is typically used to scan a 3D cryo-electron tomogram. While template matching is efficient for localizing large molecules such as ribosomes, it is necessary to apply a series algorithms to localize and identify smaller molecules as sketched in Supplementary Fig. 6. In practice, several regions of interest in the tomogram are inspected, and the localizations of macromolecules are estimated from the locations of higher scores template matching (e.g. correlation score) values. Important difficulties arise when template matching is used to localize several macromolecular species that are structurally similar, or to differentiate specific states of the same macromolecular specie (e.g. membrane-bound ribosome vs cytosolic ribosome). In real experiments, the method is applied separately to localize all macromolecules of all species (mono-class procedure). Nevertheless, dedicated classification procedures[7;8] are required to differentiate the molecules (also named particles) potentially belonging to sub-classes or corresponding to false positives. The classification remains a challenging problem in cryo-ET if the number of considered classes or subclasses is high. Currently, such complex and time-consuming processing chains are routinely applied to localize and identify the molecules. Typically, a given template matching/subtomogram classification round can take 10 to 30 hours of computation on specialized CPU clusters. The whole procedure is generally applied to analyze two or three classes/subclasses on the same volume. However, cryo-electron tomograms contain many more macromolecular species, albeit hardly discernable from noise and reconstruction artifacts. To address this challenging issue, one needs to apply powerful pattern recognition techniques.

In this paper, we describe a deep learning-based framework to fastly localize and identify macromolecules in cryo-electron tomograms. It is now well established that deep learning[9] produces state-of-the-art results in various fields (e.g. computer vision[10], language processing[11], image classification[12] and segmentation[13], and object recognition[14]). In bioimage analysis also, convolutional neural networks (CNN) produced spectacular results[15;16;17;18;19], including in super-resolution microscopy[20] and in microscopy image denoising[21]. A conventional neural network is defined as an architecture generally composed of successive connected neuron layers. The neurons are small processing units capable of performing linear and non-linear operations. Each neuron is controlled by parameters which are optimized during the learning process. In the case of CNNs, the neurons are applied in a convolutive manner, which allows dealing with the information redundancy of neighboring pixels. Thus, a neuron can be thought of as a filter, and a neuron layer as a filter bank. Each layer automatically extracts various and specific features from the data. Applying the layers sequentially enables to progressively compute high-level features, which results in a hierarchical or multiscale representation of the data. In human face recognition, the first layers typically encode basic features such as image contours/edges and textures, which allows the next layers to gradually capture more complex shapes (e.g. eyes, ears), and object ensembles (e.g. faces). Those multiscale representations are learned from the data, and turn out to be more efficient than conventional handcrafted representations which require time-consuming manual expertise.

Convolutional neural networks have been recently investigated to learn high-level generic features in electron microscopy. Previous 2D processing approaches[22;23] are intended for supervised segmentation, but are either not designed for a complex environment (e.g. crowded cell), or require an additional classification step to achieve satisfactory results. A promising 3D processing approach for the supervised classification of subtomograms has been proposed in Che *et al.*[24]. Nevertheless, this approach, evaluated on our dataset, supposes that the molecular complexes of interest have already been found by another algorithm. To overcome the limitations of the aforementioned methods[22;23;24], we propose a 3D deep learning method to both localize and identify macromolecules in a native crowded cell environment as illustrated in Fig. 1B. Our DeepFinder algorithm is capable to handle multiple macromolecular species or macromolecular sub-categories in one pass. We demonstrate that manipulating several macromolecular species at the same time, is actually the key approach to improve performance of CNNs in 3D cryo-ET. Moreover, complex and time-consuming post-classification steps are no longer required to produce reliable results contrary to the aforementioned approaches[23]. We also show that DeepFinder is very flexible, and can be efficiently combined with template matching to improve localization sensitivity on experimental cryo-ET tomograms.

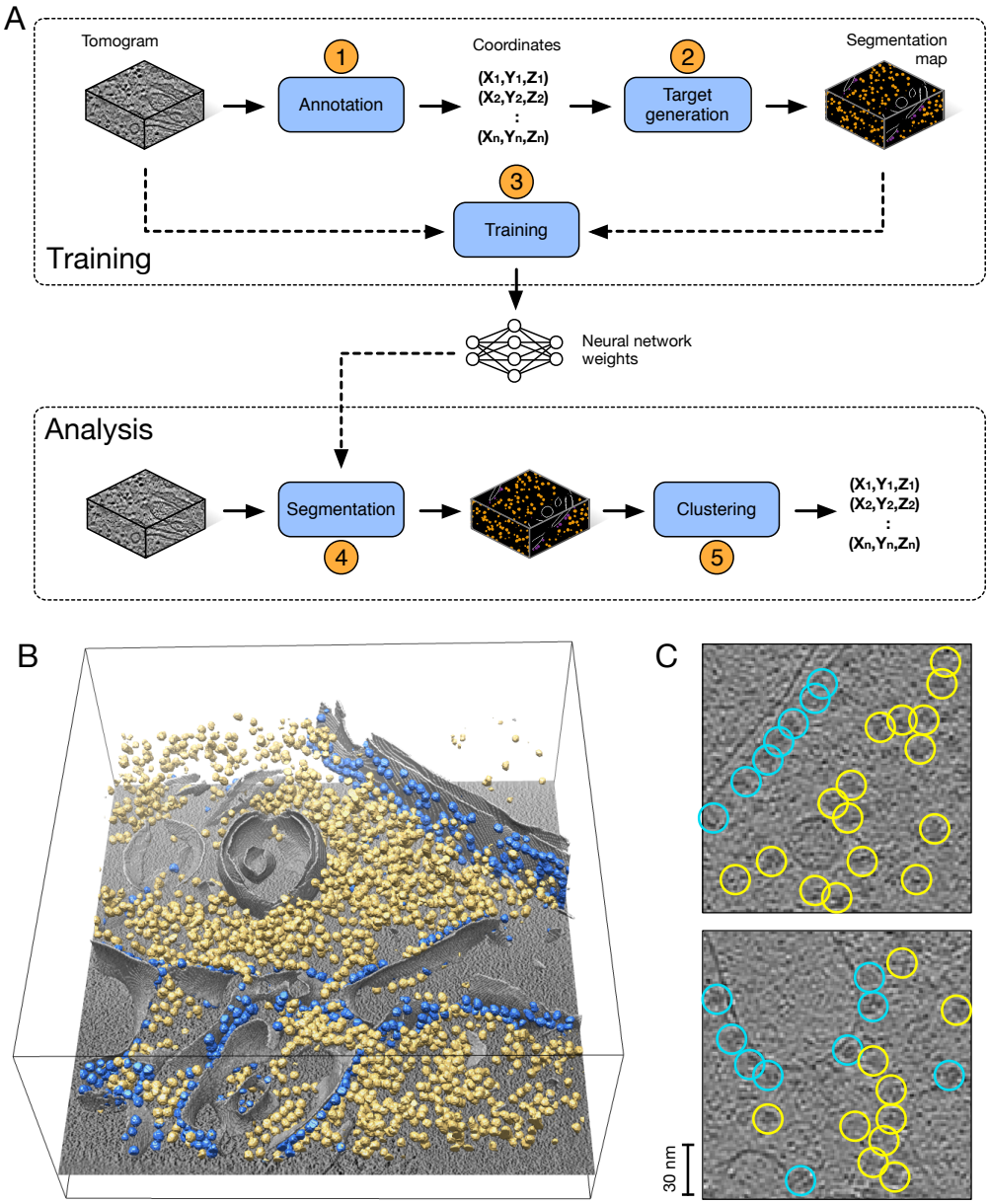

Figure 1: **Overview of DeepFinder.** (A) DeepFinder consists of 5 steps represented by blue boxes. These steps constitute a workflow that allows one to locate macromolecular complexes in crowded cells. The DeepFinder package provides for each step a graphical interface to allow non-computer scientists to use the algorithm. Meanwhile, each step can be executed with scripts using the API. (B) Ribosomes localization with DeepFinder in a cryo-electron tomogram of a *Chlamydomonas Reinhardtii* cell. Tomographic slice with superimposed segmented cell membrane (gray) and ribosomes classified with respect to their binding states: membrane-bound (blue) and cytosolic (yellow). (C) Tomographic slices showing coordinates of detected ribosomes (colors match with (B)). The positions are obtained by analyzing the segmentation map shown in (B) (outputs of the clustering step).

Results

Overview of our 3D deep-learning-based approach

We propose a flexible and powerful strategy based on 3D convolutional neural networks, and hereafter called DeepFinder, to automatically localize in one pass several macromolecules of different species, with various sizes and shapes, in 3D cellular tomograms. DeepFinder, which showed high performance on the synthetic SHREC19 challenge dataset[25] (`https://www2.projects.science.uu.nl/shrec/cryo-et/2019/`), is a two-step algorithm (see details in Methods). In the first step of the algorithm, a supervised 3D CNN-based method is used to classify the 3D tomogram voxels into different categories or sub-categories of macromolecules. In the second step, a clustering algorithm is applied to aggregate voxels into groups and determine the location of particles (gravity center of clusters) in the volume (see Fig. 1B). The particles are further picked and exploited for structure resolution determination through subtomogram averaging. DeepFinder is a user-friendly algorithm implemented in python (see Fig. 1A) with a graphical user interface (see Supplementary Fig. 7 for illustration) and a documentation available at `gitlab.inria.fr/serpico/deep-finder`. It is currently embedded into the new release of Scipion (`http://scipion.i2pc.es/`), an open-source image processing framework for cryo-electron microscopy.

DeepFinder finds multiple macromolecular species in one pass

In the SHREC'19 challenge (`https://www2.projects.science.uu.nl/shrec/cryo-et/2019/`), seven different methods (including DeepFinder) were compared to evaluate the performance of deep learning techniques and template matching on challenging synthetic 3D cryo-electron tomograms. The SHREC'19 dataset (Dataset #1) is composed of 10 synthetic tomograms, each containing 12 different classes of macromolecules, with different sizes and structures[25]. These macromolecules have been categorized by size along 4 groups (large, medium, small, tiny). The number of macromolecules per class and per tomogram varies slightly. In average, a tomogram contains 208 macromolecules per class. Participants had access to 9 out of 10 ground truths (i.e. nine training tomograms and one testing tomogram). DeepFinder achieved the best global F_1-score among all competing methods[25]. Below, we discuss the most competitive DL method and the template matching method, evaluated on the SHREC'19 dataset.

A conventional template matching approach has been implemented by the organizers[25] and served as a reference method. The most serious competitor to DeepFinder is the DoG-CB3D method[24] which separates the localization and classification tasks. For all tested methods, as shown in Fig. 2A, it turns out that the smaller the macromolecule size is, the poorer the performance of methods is. This is especially true for template matching, where the smallest macromolecules have got F_1-scores close to 0. Unlike DoG-CB3D, DeepFinder outperformed template matching for all macromolecular species considered in the benchmark dataset. Even so, template matching is still a competitive method for detecting the largest macromolecules.

On real cryo-ET cellular images, we explored the ability of our method to localize and identify two ribosome subfamilies ('membrane-bound ribosomes' and 'cytosolic ribosomes' denoted *mb-ribo* and *ct-ribo* respectively throughout the paper), and compared the results to those obtained with template matching. We used a cryo-ET dataset (Dataset #2, see Methods) composed of 57 tomograms of *Chlamydomonas Reinhardtii* cells and annotations of membrane-bound 80S ribosomes performed by experts[26]. In our experiments, we therefore considered four classes: membrane-bound ribosome (*mb-ribo*), cytosolic ribosome (*ct-ribo*), membrane, and background. Expert annotations contained positions of 8795 *mb-ribos*. The training datasets corresponding to the *ct-ribo* and membrane classes were obtained by using automatized computational tools, without careful expert supervision (see Methods). In this situation, we got highly imbalanced classes since the background class represents 99.5% of training voxels. As this scenario is common in cryo-ET, the training procedure (especially the generation of batches) needs to be adapted by integrating techniques such as re-sampling (see Methods).

Figure 1B-C illustrates an example of our 4-class segmentation, and gives an insight into the spatial distribution of two binding states of the ribosome. When computing the overlap with the expert annotations, DeepFinder achieved higher scores than template matching. For *mb-ribos*, we obtained a F_1-score of 0.86 for DeepFinder and 0.50 for template matching (Dataset #2", see Methods). Also, we noticed that a number of detected *mb-ribos* are not part of the annotations performed by experts. Part of these additional macromolecules are false positives, but, after a detailed inspection, the other part is made of macromolecules that have been missed or discarded during the annotation process.

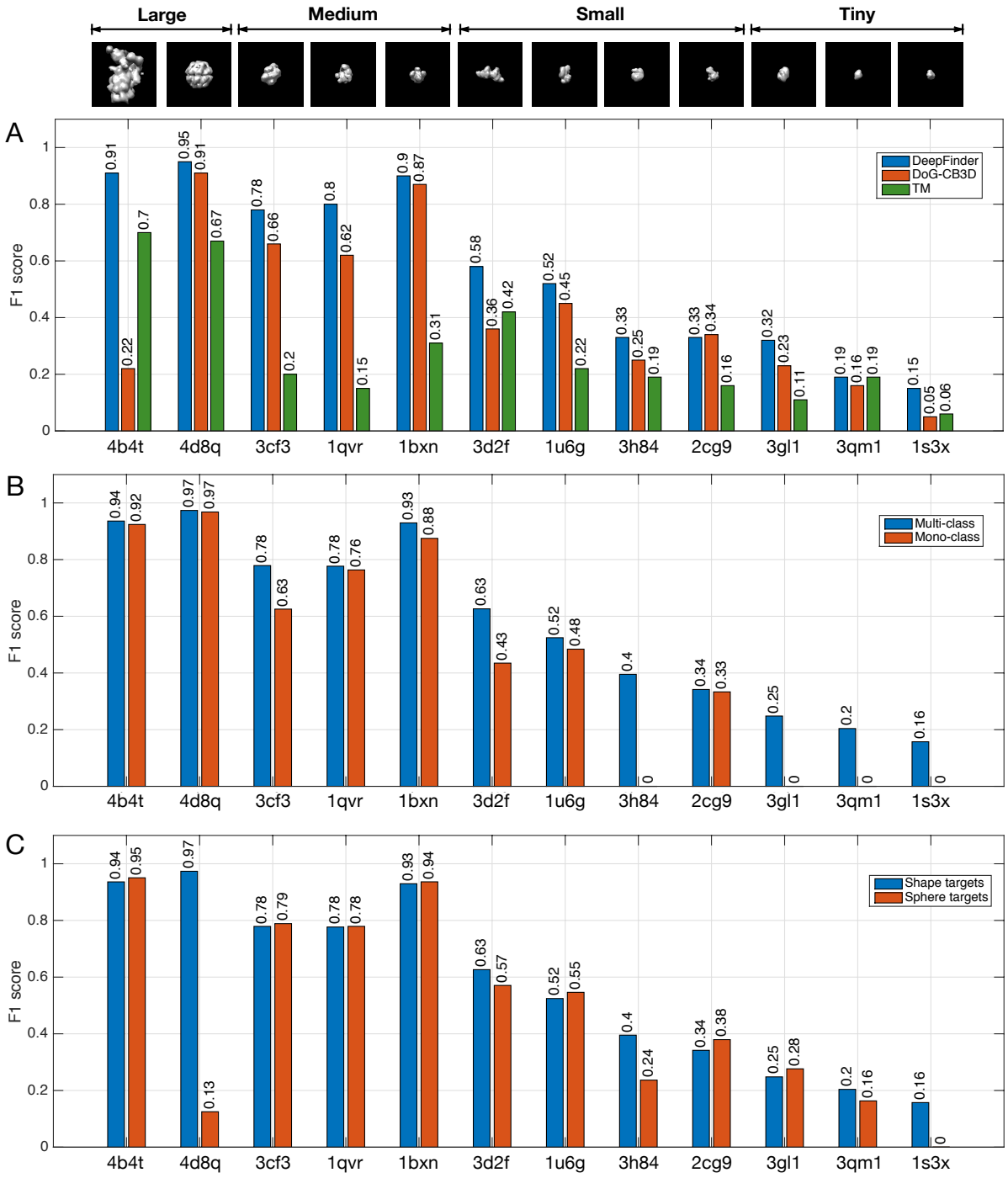

Figure 2: **Analysis of performance (F_1-score) of algorithms on the synthetic dataset (SHREC'19 challenge).** (A) Performance (F_1-score) of DeepFinder, DoG-CB and template matching algorithms and ability to discriminate between 12 classes/subclasses of macromolecules. The tiny macrolecules are 1s3x (human Hsp70 ATPase domain), 3qm1 (Lactobacillus Johnsoni Cinnsmoyl esterase LJ0536 S106A mutant in complex Ethylferulate Form II), and 3gl1 (ATPase domain of Ssb1 chaperon, Saccharomyces cerevisiae (see `https://www.rcsb.org/structure/` for description). The scores of template matching were provided by courtesy of SHREC'19 challenge organizers (Utrecht University, Department of Information and Computing Sciences and Department of Chemistry). (B) Performance of DeepFinder implemented as a multi-class network architecture and as an architecture made of 12 binary networks. These two architectures differ only by the number of output neurons. (C) Influence of the training target generation method ("shapes" versus "spheres"). In the case of "shapes", the exact shapes of the macromolecules have been used to annotate the tomograms. In the case of spheres", the shape and the orientation of macromolecules are not needed to generate the training targets.

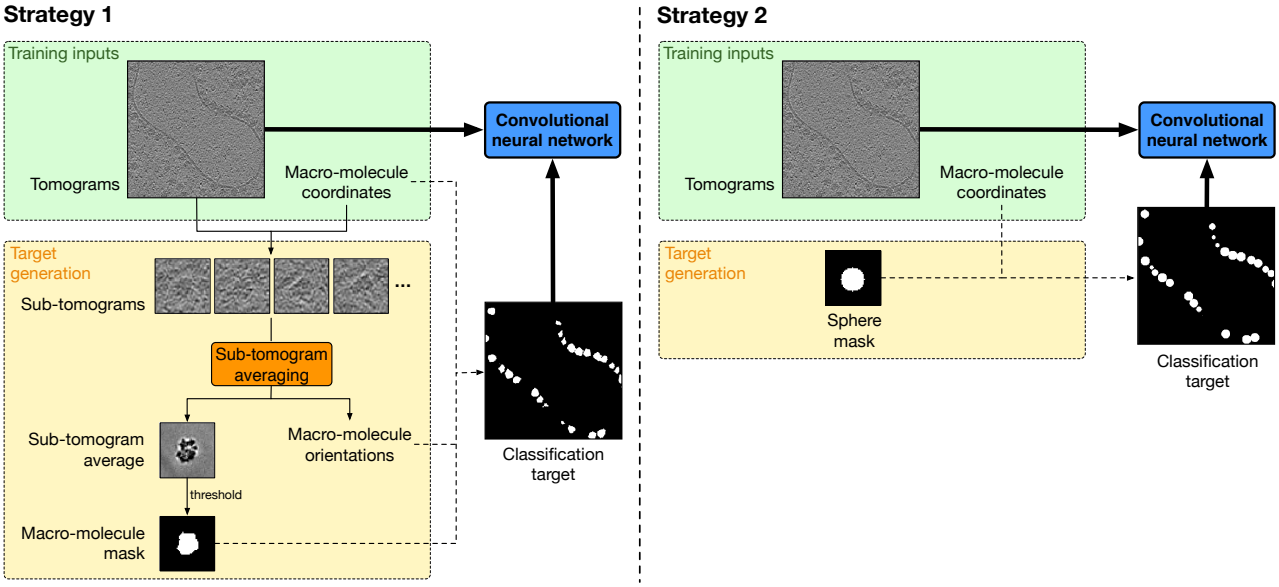

Figure 3: **Target generation strategies for training.** The DeepFinder software includes a graphical unit interface (see Supplementary Fig. 7) to to guide the experts during the whole learning procedure described below, from the selection of 3D coordinates of macromolecules to voxel-wise annotation. **Strategy 1 (left):** voxel-wise annotations are automatically obtained from position-wise annotations using a subtomogram averaging procedure. Subtomogram averaging is a registration algorithm designed to obtain higher resolution structures by averaging thousands of aligned subvolumes containing the molecular specie. The macromolecule coordinates serve as inputs to a subtomogram averaging procedure. The subtomograms centered at the annotated positions are extracted, aligned and finally averaged. The alignment procedure outputs the object orientations, whereas the averaging process provides a clean and missing wedge free density map of the macromolecule. From this density map, it is possible to create a binary mask of the macromolecule by thresholding the averaged subtomogram. Furthermore, the resulting 3D mask is pasted into an empty volume at each annotated position with the estimated 3D orientation. The resulting volume with well delineated macromolecules is then used as a target to train the 3D CNN parameters. **Strategy 2 (right):** the macromolecule masks are approximated and replaced by sphere masks. Hence the subtomogram averaging procedure is by-passed and the target generation process is much faster. The training targets contain however more *"label noise"*.

Handling several molecular species simultaneously is better than performing localization of several macromolecular species separately

In Fig. 2B, we compared the performance of our multi-class network versus twelve binary (one-class) networks as illustrated in Supplementary Fig. 8. Applying multiple binary networks is typically the strategy used in Chen *et al.*[23]. In such a scenario, the network does not need to be as deep as multi-class networks, and accordingly can run faster while requiring less computing power. Also, each binary network may work on a different resolution (i.e. voxel size), which may be useful for adapting the tomogram resolution to the size of the target macromolecule. As shown in Fig. 2B, the binary networks have a similar performance when compared to multi-class networks for larger macromolecules (e.g. 4b4t, 4d8q), but perform worse (or fail) for the smaller ones (e.g. 1s3x, 3gl1, 3h84 and 3qm1). In the end, the multi-class training provides overall better performance. Hence, we can expect to improve the localization of *mb-ribo* particles in Dataset #2 by adding the *ct-ribo* and *membrane* classes in the training set. Also, this suggests that smaller and under-represented macromolecule classes can benefit from features learned from larger and better represented macromolecule classes.

DeepFinder only needs 3D macromolecule positions for training

In the learning step, as the manual annotation of voxels can be very cumbersome and time-consuming in cryo-ET, DeepFinder converts the input 3D coordinates of macromolecules supplied by the experts into voxel-wise annotations (see Methods). It is worth noting that the DeepFinder software includes a graphical unit interface to help the experts to explore the training tomogram and select the coordinates of macromolecules used to constitute the training

dataset (see Supplementary Fig. 7). Next, converting the positions of macromolecules into voxel-wise annotations is not a trivial task in cryo-ET. We propose two approaches which have a high and a low time-computing complexity, respectively. The first approach consists in exploiting the segmentation maps built from well-delineated shapes of macromolecules estimated by a subtomogram averaging procedure[27]. In the second approach, the shapes of macromolecules are approximated by 3D spheres (see Fig. 3). Actually, the sphere-based representation is appealing since neither information about the shape nor the orientation of the macromolecules are needed. Unlike the first approach, applying time-consuming subtomogram averaging is not necessary. However, using 3D spheres instead of 3D shapes induces more "label noise", which is not desirable in the training procedure. In our experiments, it turned out that the sphere-based representation was able to provide good scores for large and medium sized macromolecules, as shown in Fig. 2C. For small and tiny macromolecules, the shape-based representation provides better results. For larger macromolecules, the "label noise" induced by the sphere-based representation is actually compensated by the high amount of labels in the segmentation map (see Rolnick *et al.*[28] for discussion). Unfortunately, the number of labels associated to small macromolecules is very small, and the "label noise" finally hinders more seriously the learning procedure. In conclusion, the "cheap" sphere-based representation is recommended, provided that the target macromolecule is not too small.

More time-consuming data annotation is required to localize small macromolecules

We analyzed the influence of the size of the training set on the performance of our method (see Supplementary Figs. 9-10). It is desired to get a good performance even if the size of training sets is small as manual annotation of data is time consuming and requires considerable effort in 3D imaging. As shown in Supplementary Fig. 10 (synthetic SHREC'19 dataset), the performance of DeepFinder progressively decreases with the size of macromolecules. For large and medium macromolecules, the performance is remarkably stable, even when using only one training tomogram (i.e. 206 macromolecules per class). For small and tiny macromolecules, the drop in performance is more significant and larger training sets are required to get high score values. It is not a surprise since the number of labeled voxels associated to large macromolecules is high in the segmentation maps. Accordingly, less annotated macromolecules are needed. In Supplementary Fig. 9, note that we do not observe a monotonous increasing of the F_1-score with respect to the number of training tomograms, mainly because of the stochastic nature of neural network training and the heterogeneous numbers of macromolecules in the training tomograms. On real cryo-electron tomograms, the F_1-score is not significantly worsen if we consider a *mb-ribos* training dataset made of one-fifth of the complete annotation dataset (8795 *mb-ribos*) (see Fig. 10). All these results suggest that DeepFinder does not require large amounts of annotations for localizing large macromolecules, but more annotations are necessary to localize small macromolecular species.

DeepFinder imitates expert annotations on experimental cryo-ET data

We have carefully examined the performance of template matching and DeepFinder methods on experimental cryo-ET data. In Fig. 4, it is clear that the score map of template matching is much noisier than the score maps generated by DeepFinder. Actually, the responses of template matching based on the constrained cross correlation are high at ribosome locations, but also at undesirable locations corresponding to highly contrasted structures with similar sizes (e.g. cell membrane in Fig. 4). Consequently, template matching is mainly used to crudely discard the voxels with no significant structural information. In the next step, the experts apply post-processing techniques to separate the desired macromolecules from the false positives. Unlike template matching, DeepFinder produces less noisy score maps, with meaningful score values in well-localized blobs. If we examine the histograms of local maxima in the score maps in Fig. 4, a significant mode is clearly visible in (C) (emphasized in red with a Gaussian distribution). The width of this mode is a relevant feature for evaluating the discrimination power of macromolecule localization algorithms. In the histogram of local maxima of scores of template matching (B), this secondary mode weakly emerges, suggesting a lower discrimination power than DeepFinder. In (D), the secondary mode is hardly visible (*ct-ribos*). In our experiments, the *mb-ribos* have have been carefully annotated by experts unlike the *ct-ribos* which where automatically annotated with no biologist-assisted supervision. The "label noise" is then expected to be higher for the *ct-ribos*, and consequently DeepFinder discriminates better the *mb-ribo* particles than the *ct-ribo* particles.

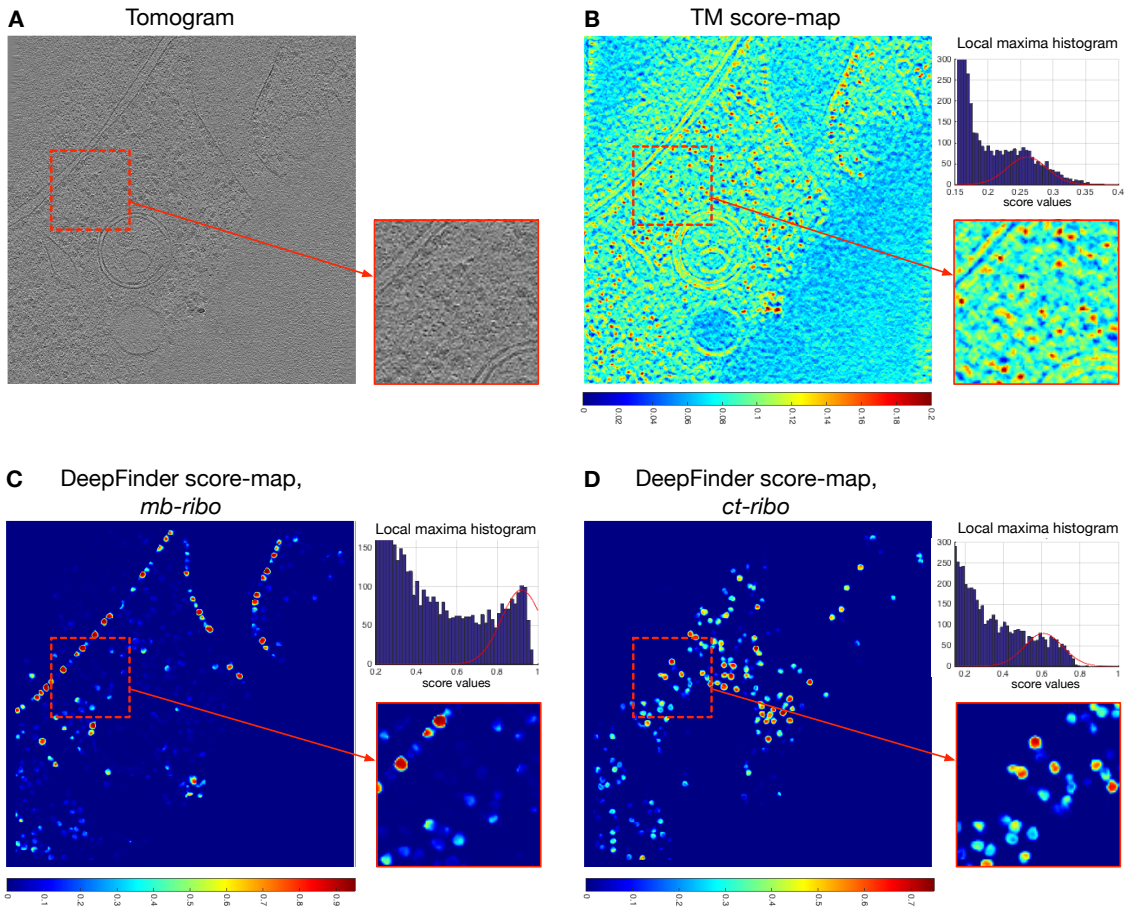

Figure 4: **Comparison of score maps obtained with template matching and DeepFinder.** (A) Experimental cryo-electron tomogram depicting a *Chlamydomonas Reinhardtii* cell (Dataset #2, see Methods). (B) Localizations of ribosomes with template matching, and histogram of local maxima of score maps of template matching. (C,D) Localizations of *mb-ribos* and *ct-ribo* particles with DeepFinder, and histograms of local maxima of score maps.

Furthermore, DeepFinder allows accurate localization of two ribosome binding states in experimental cryo-ET data as illustrated in Supplementary Fig. 6A. The *mb-ribos* (in blue) are primarily located close to cell membranes, whereas the *ct-ribos* (in yellow) are mainly located in the cytosol. This is confirmed by the analysis of probability density of Euclidean distances of *mb-ribos* and *cyt-ribo* particles to the nearest cell membranes (see Supplementary Fig. 11B). We observe a significant mode located at 136.8Å, which approximately corresponds to the radius of ribosomes; this suggests that the majority of the voxels classified as *mb-ribos* are very close or linked to the membrane. As to the *ct-ribos* class, the probability distribution is very broad with no significant mode. The distances of *ct-ribos* to the cell membranes are more heterogenous, suggesting random locations of *ct-ribos* in the cytosol.

Analysis of consensus responses. The next step of our evaluation process consists in measuring the overlap between the expert annotations and the *mb-ribos* found by DeepFinder. We compared the results with the template matching outputs, that is, before applying sophisticated post-classification methods. In Supplementary Fig. 12, we plotted the *Recall*, *Precision* and F_1 scores with respect to DeepFinder and template matching algorithm parameters. We focused on the thresholds used to detect objects with template matching (thresholding of score values) and DeepFinder (thresholding of object sizes). We obtained a maximum F_1-score of 0.86 for DeepFinder and a F_1-score of 0.50 for template matching (with no post-classification step, see Fig. 6A) (Dataset #2"). Furthermore, we also examined the complementarity between the two sets of *mb-ribo* macromolecules found by the experts and DeepFinder.

In what follows, we respectively denote S_E and S_{DF} the sets obtained by the experts and DeepFinder. While the overlap $S_E \cap S_{DF}$ between both sets is substantial (1516 particles), there is also a significant amount of particles belonging to $S_E \setminus S_{DF}$ (220 particles), i.e. the particles annotated by the expert but not found by DeepFinder, and to $S_{DF} \setminus S_E$ (356 particles), i.e. particles found by DeepFinder but missed by the expert. We can benefit from the two complementary sets of particle positions to improve the overall validation rates. Actually, the union $S_E \cup S_{DF}$ of the two sets enables to increase the list of potential *mb-ribo* macromolecules, for which a confidence level can be assigned to each set member depending on whether it belongs to $S_E \cap S_{DF}$, $S_{DF} \setminus S_E$ or $S_{DF} \setminus S_E$. The particles belonging to $S_E \cap S_{DF}$, i.e. found by both methods, are very likely to be true positives. Meanwhile the particles belonging to $S_E \setminus S_{DF}$ and $S_{DF} \setminus S_E$ can be labeled as "suspicious" and need more investigation. These two sets are relatively small and the experts may focus on the detected macromolecules that may correspond to rare conformations observed in the cryo-electron tomogram.

Analysis of structure resolution. DeepFinder allows one to achieve a 3D structure resolution comparable to the expert, as determined through subtomogram averaging. We analyzed the particles by computing subtomogram averages for each ribosome binding state (*ct-ribos* and *mb-ribos*) (Fig. 5). This way, we compared the subtomogram averages obtained from particles localized by DeepFinder on one hand, and on the other hand, from expert annotations. We applied the same registration procedure (rotational matching[29]) before computing the averages. The resulting density maps of *mb-ribos* are composed of two parts: the ribosome density and the membrane density. The membrane density varies with the number of *mb-ribo* particles involved in the average. As expected, the membrane density computed from the set S_{DF} of *mb-ribo* particles is high, while the membrane density computed from the set of *ct-ribo* particles detected by DeepFinder, is very weak or null. We noticed that the ribosomes density is similar in both ribosome subfamilies, the difference lying at the interface with the membrane.

Furthermore, we studied the Fourier shell correlation (FSC) from the subtomogram averages for each ribosome subfamily. According to Fig. 5, the best structure resolution has been obtained with the expert annotations (*mb-ribos*): 24Å versus 24.7Å. As to *ct-ribo*, the average resolution is lower (32.7Å), mainly because of the *ct-ribo* annotations were not supervised by experts, as explained earlier. As we would expect better results with DeepFinder, we carefully analyzed the difference between the two *mb-ribo* sets which served to compute the subtomogram averages. We investigated the set S_{DF} of *mb-ribo* particles found by DeepFinder and considered the two following hypotheses (see Appendix A):

H_1 : The set S_{DF} contains too many false positives.

H_2 : The set contains particles with a low quality (due to noise and blur), suggesting that our method found supplementary "true" *mb-ribo* particles, but missed or discarded during the supervised annotation.

So we aligned and computed the average of *mb-ribo* particles taken in the set $S_{DF} \setminus S_E$, i.e. found by DeepFinder but not present in the set of experts annotations. We noticed that the corresponding density map depicts a ribosome structure (see Supplementary Fig. 13). Accordingly, the hypothesis H_1 is not likely. We suggest that DeepFinder found additional membrane-bound ribosomes, missed or discarded by experts (hypothesis H_2). Therefore, it appears that the number of actual *mb-ribos* is higher than expected: in our test set, we have detected $+20.5\%$ of *mb-ribos* when compared to the S_E set. This result highlights the benefit of combining several analysis methods to investigate cryo-electron tomograms. The collaboration between the expert processing chain and DeepFinder allows one to decrease overall false negative rate and to generate a larger set $S_E \cup S_{DF}$ of membrane-bound ribosomes. Such a collaborative approach between complementary computational methods is a promising way to better investigate cryo-ET data.

Discussion

The analysis of cellular cryo-electron tomograms is a computational challenge, given the molecular crowding and high diversity of molecular species. Hence, we developed the DeepFinder solution, based on supervised deep learning for multiple macromolecular species localization. Our two-step procedure first produces a segmentation map where a class label is assigned to each voxel. The classes can represent different molecular species (e.g. ribosomes,

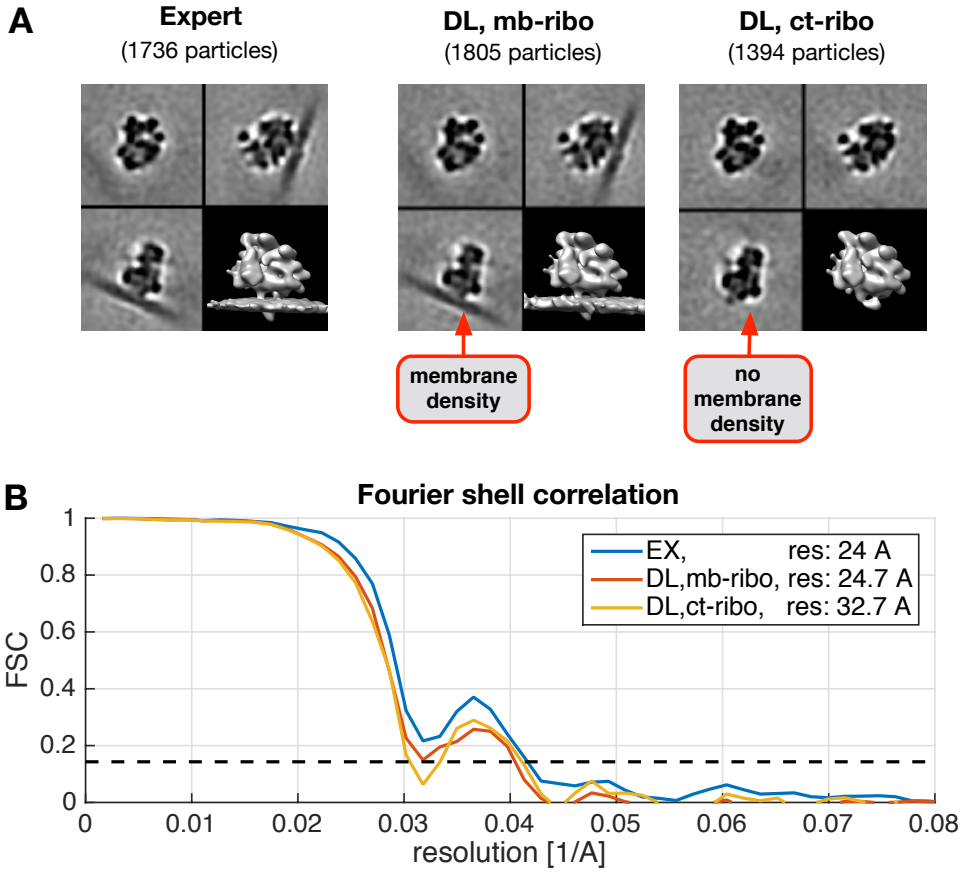

Figure 5: **Analysis of structure resolution through subtomogram averaging (Dataset #2).** (A) subtomograms obtained from expert annotations (left) and from particles localized with DeepFinder (*mb-ribos* (middle) and *ct-ribos* (right)). The density averages have been obtained with the same alignment procedure and parameters setting. For visualization purposes, the averages have been low-pass filtered at 40Å resolution. (B) gold-standard FSC curves in each case with estimated resolutions. The resolution (24Å) corresponding to expert annotations (dedicated to *mb-ribo* particles) is very similar to the resolution (24.7Å) obtained with *mb-ribo* particles localized with DeepFinder.

ATPase), states of a molecular species (e.g. binding states, functional states) or cellular structures (e.g. membranes, microtubules). In the second step, the segmentation map is used to extract the positions of macromolecules.

To perform image segmentation, we use a 3D CNN whose architecture and training procedure have been adapted for large datasets with unbalanced classes. With feasibility in mind, we developed graphically-assisted training strategies to guide the user in producing multiple segmentation maps from the spatial coordinates of macromolecules. In the first step, DeepFinder classifies the voxels into different categories of macromolecules. In the second step, a "mean-shift" algorithm[30] is applied to cluster the voxels with the same label class. Hence, the detected clusters correspond to individual macromolecules and their positions can then be derived. The "mean-shift algorithm has the advantage that it is controlled by only one parameter (depending on the macromolecules diameter) and that the number of clusters does not need to be known in advance (as opposed to k-means algorithm for instance).

While identifying large macromolecule complexes (e.g. ribosomes) is a relatively feasible task, the next computational challenge is to identify smaller complexes, with lower molecular weights. To investigate if DeepFinder has the capability to face this challenge, we evaluated its performance on a recent benchmark dataset (SHREC'19 challenge). This benchmark dataset has been designed to test the limits of macromolecule localization algorithms, including species having molecular weights as small as 70 kDa. Most of competing algorithms were based on Deep Learning. From a total of seven participants, DeepFinder has got the best performance, especially on small molecular complexes.

Even though DeepFinder yields better result for small macromolecules than the other competing methods, the algorithms need to be improved a lot. A roadmap can be derived from our experiments and observations. First, unlike large macromolecules, larger amounts of training examples are required for localizing and identifying small macromolecules. Unfortunately, the size of datasets made of tiny annotated macromolecules, is too small regarding the current needs in supervised Deep Learning. Second, handling several molecular species (large, medium, small) at once helps finding the smallest macromolecules. A relative good annotation of large molecular complexes enables to guide the identification of smaller detected macromolecules. A few sets of features learned with CNNs are actually common for the large and small complexes, and serve to efficiently discriminate the smallest detected particles. As a result, DeepFinder has a low false positive rate, which gives it a clear advantage over template matching. The price to be paid for better results is to resort to supervised machine learning algorithms guided by annotations performed by experts.

In the case of experimental data, annotations are a subset of the unknown ground truth. Therefore, we evaluated the performance of DeepFinder by applying subtomogram averaging to *mb-ribos* and *ct-ribo* particles. From the FSC, we found that the structure resolution computed from the set of *mb-ribo* particles localized with DeepFinder (24.7Å), is very similar to the one computed from the experts annotations set (24 Å). By investigating the difference between the two sets, we brought to the fore that the set of supplementary *mb-ribo* particles found by DeepFinder were missed or discarded during the annotation process. Both sets are therefore complementary, and combining the tow sets allows one to decrease the overall false negative rate while increasing the set of true positive *mb-ribo* particles (20.5% in our experiments). The combination of algorithms with different designs, paves the way towards successful identification of macromolecules in cryo-ET. Having an accurate picture of the actual number of macromolecules in situ has paramount importance to analyze their spatial distribution[8].

Meanwhile, cryo-electron tomograms are rarely processed to detect more than two or three molecular species. Yet they contain rich information on the whole proteome, albeit hardly discernable from noise and artifacts. To move towards visual proteomics[31] and study the interaction between several macromolecular species, more powerful pattern recognition techniques are required. We believe that deep learning-based methods, such as DeepFinder, are on the right way to address this challenge. Unlike template matching, they proved to be invariant to diffeomorphisms[32] (i.e. able to cope with non-rigid, elastic deformations within a class) and are therefore able to localize complexes which exhibit different conformational states.

In conclusion, our results demonstrated that DeepFinder is very appropriate to efficiently localize and identify molecular complexes with variable sizes and structures. For the first time, we showed objectively that we can obtain a structure resolution with a Deep Learning-based method, similar to the one obtained by experts. Moreover, once trained, DeepFinder is relatively fast (once trained) when compared to the common template matching and subtomogram classification pipeline. Larger datasets can be processed in one day with DeepFinder, while several macromolecular species are simultaneously identified. Furthermore, we provide an open-source implementation of DeepFinder, with a graphical unit interface for both training and localizing/identifying macromolecules in cellular tomograms, aimed towards non-computer scientist users.

Future challenges include developing unsupervised deep learning methods[33] and applying transfer learning[34] in order to reduce human annotation effort and annotation errors. Transfer learning would allow one to download pre-trained networks and directly apply them to newly acquired data (see an example in Supplementary Fig. 14). This would very helpful for research groups with limited computing facilities, which want to benefit from the power of deep learning (segmentation is less computation intensive than training).

Methods

DeepFinder algorithm

In this section, we describe the two-step DeepFinder algorithm.

Step #1 – Multi-class voxel-wise classification. The objective is to provide a segmentation map for which each voxel is assigned to a class label (representing a macromolecular specie). The architecture of DeepFinder is based on the U-net, an "encoder-decoder" type network designed for segmenting images in an end-to-end manner. The U-net is an extension of the fully convolutional network[13;17], which achieves multi-resolution feature representation and produces high resolution label maps. DeepFinder architecture consists of a down-sampling path (i.e. encoder) needed to generate global information and a up-sampling path (i.e. decoder) used to generate high-resolution outputs[35;36], i.e. local information (see Supplementary Fig. 8). Down-sampling is performed with max-pooling layers (factor 2) and up-sampling with up-convolutions[13] (sometimes called "backward convolution"), which is basically a trained and non-linear up-sampling operation. Combining global and local information is performed by concatenating features at different spatial resolutions. The features are then processed with the convolutional layers of the up-sampling path. The fully convolutional nature of DeepFinder allows one to use various input-patch shapes, with the constraint that the patch size must be a multiple of four, because of the two down-sampling stages. Then, arbitrary large tomograms can be processed by using an overlap-patch strategy. Unlike Milletari *et al.*[36], DeepFinder is not that "deep", since we found that using more than two down-sampling stages does not increase the classification results. Also, we used only $3 \times 3 \times 3$ filter sizes as in described by Simonyan *et al.*[37]. The rationale behind this choice is that two consecutive $3 \times 3 \times 3$ filters mimic a larger $5 \times 5 \times 5$ filter but with fewer parameters. Training is then faster and easier and requires less memory. An important concept in neural architectures is the receptive field of deepest neurons layers. It determines the size of the spatial context to be used to make decisions. Considering a large spatial context is essential to handle macromolecule class involving interactions with the environment, for instance interactions with the cell membrane. It is established that adding convolutional layers after down-sampling operations is appropriate to enlarge the spatial context[36]. Accordingly, we added two supplementary convolutional layers in the lowest stage of our architecture. In the end, the receptive field size of our network is $48 \times 48 \times 48$ voxels. To complete the description, we use rectified linear units (ReLU)[12] as activation function for every layer except the last one which uses a *soft-max* function. While ReLU is a popular choice to tackle non-linearities in the network, the *soft-max* function is necessary in order to interpret the network outputs as pseudo-probabilities for each class. In summary, step#1 is capable of robustly classifying the cryo-ET tomogram into N classes with a high accuracy.

Step #2 – Clustering for macromolecules localization. Given the multi-class voxel-wise classification map and classification errors, the objective of step #2 is to estimate the spatial coordinates of each particle of a given macromolecular specie. The voxel labels should be ideally spatially well clustered into well distinct 3D connected components, each cluster corresponding to unique particle. Because of noise, non-stationarities in the background, and artifacts in the tomogram, step #1 generates isolated labels or very small groups of voxels, and groups that contain different labels. Post-processing is then necessary to aggregate neighboring voxels into 3D connected components and to assign a unique label to each component (or cluster). The smallest clusters are considered as false positives and are discarded. The centroids of meaningful clusters are computed to determine the location of the particle. As the centroid is computed by uniformly averaging the coordinates of cluster voxels, we are able to numerically produce positions with sub-voxel precision. As several voxel labels can be spatially grouped in a given cluster, the most frequent label is assigned to the detected particle. To address this task, we used the popular "mean-shift" clustering algorithm[30]. The main advantage of "mean-shift" is that it is controlled by only one parameter, commonly called the bandwidth, which is directly related to the average size of macromolecules. The K-means algorithm was not considered further since the number of clusters must be provided as input parameter by the user.

Training

In the training step, DeepFinder parameters are learned from pairs of tomograms and their corresponding voxel-wise annotation. The underlying 3D CNN actually needs that each voxel is annotated as member of a given macromolecular specie, or as background. While voxel-wise classification examples are naturally available for synthetic data, it is often not the case for experimental data. In our case, the experts only supplied the 3D coordinates of macromolecules of interest without labeling voxels, in several training tomograms. In practice, voxel-wise annotation cannot be performed manually for two reasons in cryo-ET: *i)* it is time consuming to individually label each voxel belonging to a 3D macromolecule; *ii)* data is so noisy that the macromolecule borders are barely visible.

To get voxel-wise annotations, we propose two strategies from 3D spatial coordinates of particles:

- *Sphere-wise annotation:* 3D spheres are positioned at each annotated location in the tomogram. The sphere radius should be similar to the size of the target macromolecule. This strategy is easy to implement and is appealing in terms of computing time. The main limitation is that spheres may introduce significant "label noise"[28]. The resulting errors in the annotations hinder the training procedure, and the segmentation network may have sub-optimal performance (the network gets "confused"). Nonetheless, it has been shown that CNNs have a natural robustness to reasonable amount of "label noise"[28], which is also confirmed in our experiments with DeepFinder.

- *Shape-wise annotation:* To reduce the amount of "label noise", we propose a second automatic strategy to convert macromolecule coordinates into voxel-wise annotations. Instead of using spheres, we generate the target by using the structural shape of the macromolecules. These targets are more difficult to obtain, because information on the shape and the orientation of the macromolecules are needed. In order to obtain these informations, we exploit the commonly used subtomogram averaging procedure[27] (see Fig. 3). While this strategy provides more accurate targets, it still induces "label noise", albeit less than when using spheres. Indeed, as we project several times the same average shape (computed with subtomogram averaging) at different positions into the tomogram for training, we neglect the structural details of individual macromolecules.

In our experiments, all results have been obtained by using shape-wise annotations, except for the comparison between sphere-wise and shape-wise annotations (Fig. 2C).

Training datasets and batch generation. Due to memory limitations, it may not be feasible to load the whole tomogram set with the corresponding targets during training. Instead, for each training iteration, we sample a group of smaller 3D patches to constitute a "batch". The patch size should be large enough to capture sufficient context information. The macromolecule radius being in the range of 10 voxels, we choose a patch size of $56 \times 56 \times 56\times$ voxels. Our implementation is such that only the current batch is loaded into memory. This helps to handle large dataset with limited computational resources, allowing DeepFinder to be applied on lower end consumer GPUs with less RAM.

Another problem is the very high imbalance between class labels. Because of the small size of the macromolecules compared to the tomogram size, more than 99% of voxels are annotated as "background". This causes the trained network to be skewed towards the over-represented class. Therefore, we "guide" the sampling procedure as follows: the patches are sampled only at annotated positions such that each patch contains at least one macromolecule. An additional benefit of sampling only at annotated positions is that the amount of false negatives is reduced (here the negative class is the background). It is indeed common that annotations are not exhaustive. Therefore false "background" labels remain at missed macromolecule positions, contributing to "label noise" increase. The proposed patch sampling procedure does not discard all false negatives (e.g. a false positive could be neighboring a true positive), but the number should be relatively small at the end.

An additional problem is the imbalance between competing macromolecule classes: for instance, 20 macromolecules of class #1 (e.g. *mb-ribo* particles) versus 100 macromolecules of class #2 (e.g. *ct-ribo* particles). To reduce this imbalanced class effect, we apply a bootstrapping algorithm (i.e. resampling), so that the distribution of the positive classes in a batch will be uniform. This stochastic resampling procedure is effective for sampling the under-represented classes more frequently than the over-represented classes.

Finally, it is also common in deep learning to use "data augmentation" to artificially increase the size of training sets. The idea is to apply geometric transform to training images. In our approach, we implement "data augmentation" by applying a $180°$ rotation with respect to the microscope tilt-axis, to each training example. Nevertheless, we do not use "mirror" operations or geometric deformations because the macromolecule structure (including its chirality) is the main clue for detection. Also, we do not use random rotations because of the well-determined orientation of missing wedge artifacts, which is preserved when applying $180°$ rotations with respect to tilt-axis. Finally, we apply random shifts to improve invariance to translations.

Optimization. In our experiments, DeepFinder has been computationally trained with the ADAM algorithms, chosen for its good convergence rate[38], by setting the learning rate to 0.0001, the exponential decay rate to 0.9 for the first moment estimate, and to 0.999 for the second moment estimate. The batch size is set to 25 and the patch size to $56 \times 56 \times 56$ voxels. No regularization (e.g. $L2$ regularizer, "drop out") was needed for processing the datasets. We implemented both the categorical cross-entropy (as implemented in Keras) and the Dice[39] as loss functions. On the synthetic SHREC'19 dataset, the Dice loss function was shown to be able to better localize the smallest macromolecules. However, both yield similar results on real cryo-ET data (localization of *mb-ribo* and *ct-ribo* particles).

Description of synthetic data (Dataset #1)

The synthetic dataset has been generated by the Cryo-Electron Microscopy group of Utrecht University in the frame of the challenge "SHREC 2019: Classification in Cryo-Electron Tomograms", organized for the Eurographics Workshop on "3D Object Retrieval"[25]. The dataset has been especially designed to objectively evaluate the performance of localization and classification algorithms in cryo-electron tomography. The dataset has been generated as follows[25] : *i)* tomogram density maps were generated using Protein Data Bank (PDB) structures (identifiers: 1bxn, 1qvr, 1s3x, 1u6g, 2cg9, 3cf3, 3d2f, 3gl1, 3h84, 3qm1, 4b4t and 4d8q); *ii)* a series of projection images was generated with a $\pm 57°$ tilt-range and a $3°$ tilt-step; *iii)* projection images were degraded by noise and contrast transfer function such that signal-to-noise ratio was 0.02; *iv)* the tomogram was reconstructed using weighted back-projection. The resulting tomograms have a size of $512 \times 512 \times 512$ voxels and a voxel size of 1 nm.

In our experiments, the dataset #1 has been split into training, validation and test sets. The training set is composed of 8 tomograms annotated with a total of 19956 macromolecules. The validation set is composed of one tomogram annotated with a total of 2490 macromolecules. The test set is composed of one tomogram annotated with a total of 2540 macromolecules. The training has been performed until convergence, which represents in average 10000 iterations and 33 hours of computation on a Nvidia Tesla K80 GPU, using the Dice loss function.

Description of experimental cryo-ET data (Dataset #2)

The real cryo-ET dataset is composed of 57 tomograms of *Chlamydomonas Reinhardtii* cells, and was used for membrane-bound 80S ribosome (*mb-ribo*) annotation. To produce the annotations, the experts essentially used template matching, and then filtered out the false positives by applying the CPCA subtomogram classification algorithm[7]), and careful visual inspection (see Supplementary Fig. 6). To reduce the computational cost, the tomograms were under-sampled, resulting into a tomogram size of $928 \times 928 \times 464$ voxels and a voxel size of 13.68Å. Tilt range is $\pm 60°$ with an increment of $2°$.

As starting point, Dataset #2 has been annotated by localizing 8795 *mb-ribo* particles under expert supervision. Then, we introduced two additional classes (without expert supervision): the cytosolic ribosomes (*ct-ribos*) class and the *membrane* class. Annotations for these two additional classes were obtained by applying semi-automatized computational tools to tomograms. First we selected members of the *membrane* class by employing an algorithm dedicated to cell membrane segmentation[40]. Meanwhile, we obtained training (*ct-ribo*) particles by applying template matching and selecting the most isolated candidates, located at a distance higher than 273.6Å (i.e. the ribosome diameter) with respect to the cell membrane. The motivation behind adding new classes to the available annotations was twofold: on the one hand, we wanted to demonstrate the ability of DeepFinder to localize and identify multiple molecular species on real data with one pass; on the other hand, we noticed that the multi-class approach tends to

improve the discriminating power of CNN, therefore better finding *mb-ribo* particles in experimental tomograms. When trained by using only the *mb-ribo* examples, DeepFinder actually finds unwanted cytosolic ribosomes. By considering the *ct-ribo* class in the training, we encourage the network to better discriminate both ribosome subclasses (corresponding to binding states). Note that these annotations of membrane components and cytosolic ribosomes have been obtained without the supervision of an expert. Therefore more errors are expected for these two classes when compared to the *mb-ribo* examples, reliably annotated by the experts.

Dataset #2 has been arbitrarily split into training, validation and test sets. The training set is composed of 48 tomograms annotated with 6834 *mb-ribos* and 6687 *ct-ribo* particles. The validation set is composed of one tomogram annotated with 222 *mb-ribos* and 254 *ct-ribo* particles. The test set is composed of 8 tomograms annotated with 1736 *mb-ribos* and 2594 *ct-ribo* particles. In our experiments, we consider "background" and three classes: *mb-ribo*, *ct-ribo* and *"membrane"*. Training is achieved using the categorical cross-entropy as a loss function (we got very similar results with the Dice loss function). The training has been performed for 6000 iterations and took 20 hours of computation on a Nvidia Tesla K80 GPU (see plots of loss evolution during training in Supplementary Fig. 15).

Description of template matching

The template matching results on the synthetic SHREC'19 dataset were provided by courtesy of Utrecht University (Department of Information and Computing Sciences, Department of Chemistry) and were obtained as follows: *i)* template matching has been applied for each class. As the particles have been simulated, we know the true number of candidates for each tomogram: 2500 particles/12 classes = 211. Then, the top 211 candidates for each class were sequentially selected and extracted, from the largest macromolecules to the smallest ones, in a way that all these candidates do not overlap with the already extracted candidates.

We used the template matching algorithm implemented in the PyTom toolbox[41] to analyze the Dataset #2 (*Chlamydomonas Reinhardtii* cells).

Evaluation

The metric used to assess localization performance is the F_1-score, as commonly used in detection problems. The F_1-score (also known as Dice coefficient) can be interpreted as a weighted average of two well-known metrics depending on the number of true positives (TP):

- *Recall* (also known as sensitivity):

$$R = \frac{\text{Number of TP}}{\text{Number of particles in the ground truth}},$$

- *Precision* (also known as positive predictive value):

$$P = \frac{\text{Number of TP}}{\text{Number of detected particles}}.$$

The F_1-score, defined as follows

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

allows one to evaluate performance by considering a single value.

Moreover, for each test tomogram, bounding boxes are placed at each true location. These bounding boxes are of same size as the target macromolecule. Then, a detected particle is considered as a TP if the estimated centroid is located in the bounding box.

Dataset #1 is very helpful to objectively study the performance of algorithms. In the case of real cryo-ET data (Dataset #2), the spatial coordinates of *mb-ribo* particles supplied by the experts, served to place bounding boxes and to compute F_1 sores. Unlike Chen *et al.*[23], we also analyzed the score distributions and subtomogram averages. The resolution of obtained subtomogram averages is estimated with the commonly-used gold standard Fourier shell correlation (FSC) and determined for a threshold equal to 0.143.

Code availability and implementation details

Each step of DeepFinder in Fig. 1A can be executed with scripts using the API (examples are provided) or with a graphical user interface. These steps may also be embedded in other workflows, e.g. if the user needs only the segmentation step. To implement DeepFinder, we used Keras [keras.io], an open-source toolbox written in python and using the Tensorflow framework. Our code can be downloaded for free from our GitLab website: [`https://gitlab.inria.fr/serpico/deep-finder`] with a documentation on line (`https://deepfinder.readthedocs.io/en/latest/`).

The execution times to process a $928 \times 928 \times 464$ tomogram are 6 min for step #1 (multi-class voxel-wise classification) and 15 min for step #2 (clustering), respectively.

All training procedures were achieved using a Nvidia Tesla K80 GPU, by using Cuda 8 and cuDNN 6. Below, we display the memory consumption of DeepFinder, for different training parameters.

	Batch size	
Patch size (voxels)	**15**	**25**
$40 \times 40 \times 40$	2.56 GigaBytes	3.76 GigaBytes
$56 \times 56 \times 56$	6.16 GigaBytes	9.86 GigaBytes

We used the PyTom toolbox[41] (template matching, subtomogram averaging, fast rotational matching[29] for subtomogram alignment). The alignment has been performed with respect to a reference template. We used Chimera[42] for 3D visualization purposes. TomoSegMemTV[40] was used to annotate the "membrane" class.

Data availability

The synthetic dataset is available on the website of the SHREC19 challenge (`http://www2.projects.science.uu.nl/shrec/cryo-et/2019/`). A tomogram from the experimental dataset of *Chlamydomonas Reinhardtii* cells[26;43] can be found in the Electron Microscopy Data Bank (`https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-3967`) under accession number EMD-3967.

Acknowledgements

This work was jointly supported by Fourmentin-Guilbert Scientific Foundation, and Région Bretagne (Brittany Council). Experiments on real data (courtesy of MPI Biochemistry, Martinsried, Germany) were performed on the Inria Rennes computing grid facilities partly funded by France-BioImaging infrastructure (French National Research Agency - ANR-10-INBS-04-07, "Investments for the future").

We would like to thank S. Pfeffer, B. Engel and S. Albert for providing their *Chlamydomonas Reinhardtii* dataset as well as their annotations. We also would like to thank F. Förster M. Killinger for fruitful discussions about cryo-ET data analysis, and deep learning applied to large 3D volumes analysis, respectively.

We thank the organizers of the SHREC'19 challenge for helpful assistance and for providing the template matching results: Ilja Gubins & Remco C. Veltkamp (Utrecht University, Department of Information and Computing Sciences), Gijs van der Schot & Friedrich G. Forster (Utrecht University, Department of Chemistry).

Finally, we thank S. Prima for careful reading of the paper and valuable suggestions and comments.

Author contributions

E. Moebel designed and implemented the presented method, carried out the experiments and wrote the manuscript. C. Kervrann supervised the project, was in charge of overall direction and planning, and co-wrote the manuscript. A. Martinez-Sanchez and J. Ortiz helped conceiving the experiments. A. Martinez-Sanchez also assisted the conceptualization. W. Baumeister provided access to the dataset and substantively revised the work. D. Larivière and E. Fourmentin devised the project and the main conceptual ideas. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Competing interests

The authors declare no competing interests.

Note

Through the LifeExplorer project, the Fourmentin-Guilbert Scientific Foundation has been a pioneer in approaching the structural modeling and visualization of entire cells. It is expected from the creation of interactive 3D avatars of cellular environments, bridging from the level of atoms to the level of cells, that the rules governing the spatiotemporal organization of the cytoplasm could be revealed. Such an approach requires to make an inventory and a cartography of all the components constituting a single cell.

The technique of choice for such a mapping is cryo-electron microscopy applied on frozen but intact cells. For years, the Fourmentin-Guilbert Foundation has supported the Max Planck Institute of Biochemistry, headed by Prof. W. Baumeister whose team has been capable of delivering whole cryo-electron tomograms of E. coli cells at an unprecedented resolution.

The next big challenge, still an ongoing challenge, was to recognize macromolecular components within the tomograms. Most of the effort of the scientific community was put on the identification of the ribosomes thanks to a methodology relying on single-particle analysis and template matching and giving impressive results. However, it is likely that such an approach will be limited to large macromolecules like the ribosomes. Facing this challenge, the Fourmentin-Guilbert Foundation has solicited the research group headed by Dr. C. Kervrann to develop alternative recognition methods having the potential to help the *in-situ* identification of the thousands of proteins left in the dark. The Serpico Project-Team has then developed and compared with well-established methods new approaches based on deep learning and capable of identifying and counting the “gold standard” ribosomes within a tomogram. Theses methods, as an alternative to template matching, should also have the potential to apply on particles smaller and rounder than ribosomes.

References

[1] Frank, J. Approaches to large-scale structures. *Curr. Opin. Struct. Biol.* **5**, 194–201 (1995).

[2] McEwen, B., Renken, C., Marko, M. & Mannella, C. Principles and practice in electron tomography. *Methods Cell Biol.* **89**, 129–168 (2008).

[3] McIntosh, R., Nicastro, D. & Mastronarde, D. ew views of cells in 3d: an introduction to electron tomography. *Trends Cell Biol.* **15**, 43–51 (2005).

[4] Nicastro, D., Frangakis, A., Typke, D. & Baumeister, W. Cryo-electron tomography of neurospora mitochondria. *J. Struct. Biol.* **129**, 48–56 (2000).

[5] Guesdon, A., Blestel, S., Kervrann, C. & Chrétien, D. Single versus dual-axis cryo-electron tomography of microtubules assembled in vitro: limits and perspectives. *J. Struct. Biol.* **181**, 169–78 (2013).

[6] Best, C., Nickell, S. & Baumeister, W. Localization of protein complexes by pattern recognition. *Methods Cell Biol.* **2007**, 615–638 (2007).

[7] Förster, F., Pruggnaller, S., Seybert, A. & Frangakis, A. S. Classification of cryo-electron sub-tomograms using constrained correlation. *J. Struct. Biol.* **161**, 276–286 (2008).

[8] Martinez-Sanchez, A. *et al.* Template-free detection and classification of membrane-bound complexes in cryo-electron tomograms. *Nat. Methods* **17**, 209–216 (2020).

[9] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

[10] LeCun, Y., Kavukcuoglu, K. & Farabet, C. Convolutional networks and applications in vision. In *Proc. IEEE Int. Symp. on Circuits and Systems*, 253–256 (2010).

[11] Hinton, G. *et al.* Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process. Mag.* **29**, 82–97 (2012).

[12] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proc. Neural Inf. Processing Systems (NIPS)*, 1–9 (2012).

[13] Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proc. Conf. Comput. Vis. Pattern Recognition*, 3431–3440 (2014).

[14] Szegedy, C., Toshev, A. & Erhan, D. Deep neural networks for object detection. In *Proc. Conf. Neural Inf. Processing Systems (NIPS)*, 1–9 (2013).

[15] Sullivan, D., Winsnes C., L., and Akesson & Lundberg, E. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnology* **36**, 820–828 (2018).

[16] Moen, E. *et al.* Deep learning for cellular image analysis. *Nat. Methods* **16**, 1223–1246 (2019).

[17] Falk, T. *et al.* U-net – deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).

[18] Gupta, A. *et al.* Deep learning in image cytometry: A review. *Cytometry Part A* **95**, 366–380 (2019).

[19] Belthangady, C. & Royer, L. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat. Methods* **16**, 1215–1225 (2019).

[20] Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. Deep learning massively accelerates super-resolution localization microscopy. *Nat. Biotechnology* **36**, 460–468 (2018).

[21] Weigert, M. *et al.* Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nat. Methods* **12**, 1090–1097 (2018).

[22] Wang, F. *et al.* DeepPicker : A deep learning approach for fully automated particle picking in cryo-EM. *J. Struct. Biol.* **195**, 325–336 (2016).

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

[24] Che, C. *et al.* Improved deep learning based macromolecules structure classification from electron cryo tomograms. *Mach. Vis. Appl.* **29**, 1227–1236 (2018).

[25] Gubins, I. *et al.* SHREC'19 Track : Classification in cryo-clectron tomograms. In *Eurographics Workshop on 3D Object Retrieval: SHREC - 3D Shape Retrieval Contest (2019)*, 1–6 (2019). URL `https://www2.projects.science.uu.nl/shrec/cryo-et/2019/`.

[26] Pfeffer, S. *et al.* Dissecting the molecular organization of the translocon-associated protein complex. *Nat. Communications* **8**, 14516 (2017).

[27] Förster, F. & Hegerl, R. Structure determination *in situ* by averaging of tomograms. *Cell. Electron Microsc.* **79**, 741–767 (2007).

[28] Rolnick, D., Veit, A., Belongie, S. & Shavit, N. Deep learning is robust to massive label noise. *arXiv Prepr.* (2017). `arXiv:1705.10694v2`.

[29] Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J. M. & Förster, F. Fast and accurate reference-free alignment of subtomograms. *J. Struct. Biol.* **182**, 235–245 (2013).

[30] Comaniciu, D., Meer, P. & Member, S. Mean Shift : a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 603–619 (2002).

[31] Förster, F., Han, B. G. & Beck, M. Visual proteomics. *Methods Enzymol.* **483**, 215–243 (2010).

[32] Mallat, S. Understanding deep convolutional networks. *Philos. Trans. R. Soc. London A Math. Phys. Eng. Sci.* **374** (2016).

[33] Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep clustering for unsupervised learning of visual features. In *Proc. Eur. Conf. Computer Vision*, 132–149 (2018).

[34] Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks ? In *Proc. Conf. Neural Inf. Processing Systems (NIPS)*, 3320–3328 (2014).

[35] Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, vol. 9351, 234–241 (2015).

[36] Milletari, F., Navab, N. & Ahmadi, S.-a. V-Net : fully convolutional neural networks for volumetric medical image segmentation. *arXiv Prepr.* 1–11 (2016). `arXiv:1606.04797v1`.

[37] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Representation*, 1–14 (2015).

[38] Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. *arXiv Prepr.* (2014). `arXiv:1412.6980v9`.

[39] Salehi, S. S. M., Erdogmus, D. & Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. *arXiv Prepr.* (2017). `arXiv:1706.05721v1`.

[40] 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).

[41] Hrabe, T. *et al.* PyTom: A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *J. Struct. Biol.* **178**, 177–188 (2012).

[42] Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

[43] Albert, S. *et al.* Proteasomes tether to two distinct sites at the nuclear pore complex. *Proc. Natl. Acad. Sci.* **114**, 201716305 (2017).

[44] Henderson, R. Avoiding the pitfalls of single particle cryo-electron microscopy : Einstein from noise. *Proc. Natl. Acad. Sci.* **110**, 18037–18041 (2013).

Supplementary information

Deep Learning Improves Macromolecules Localization and Identification in 3D Cellular Cryo-Electron Tomograms

Emmanuel Moebel[1], Antonio Martinez-Sanchez[2], Damien Larivière[3], Eric Fourmentin[3], Julio Ortiz[4], Wolfgang Baumeister[2], Charles Kervrann[1]

[1] Serpico Project-Team, CNRS-UMR 144, Inria, CNRS, Institut Curie, PSL Research University, Campus Universitaire de Beaulieu, 35 042 Rennes Cedex, France
[2] Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
[3] Fourmentin-Guilbert Scientific Foundation, 2 Avenue du Pavé Neuf, 93 160 Noisy-le-Grand, France
[4] Ernst Ruska-Centre, Wilhelm-Johnen-Straße, 52425 Jülich, Germany

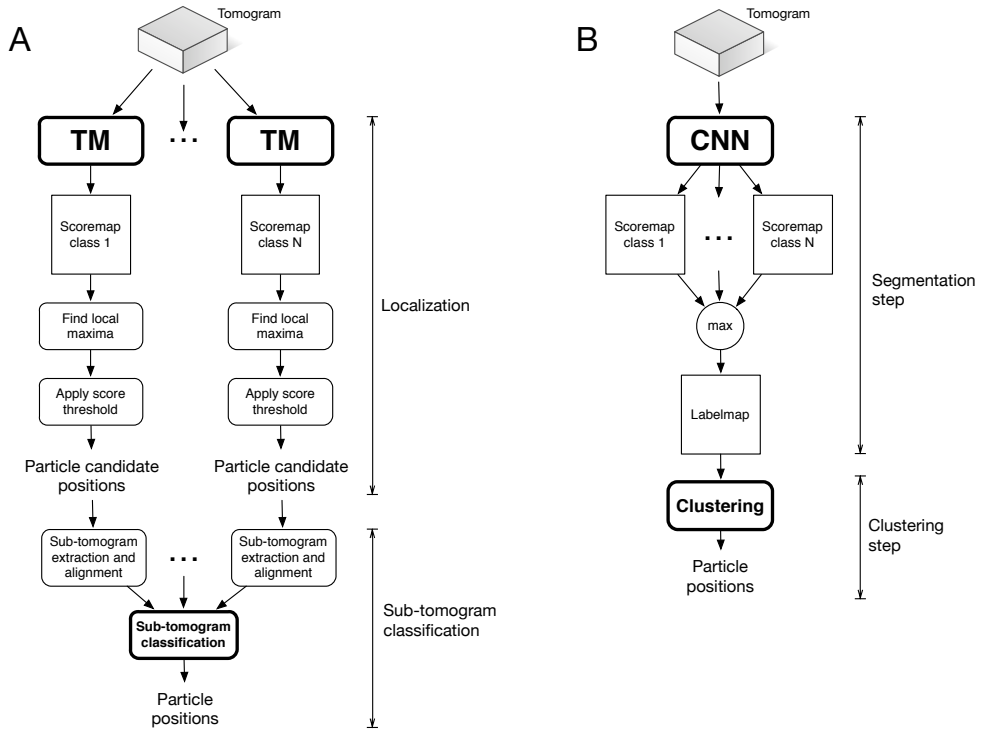

Figure 6: **Two workflows for macromolecules localization in cryo-ET.** (A) Conventional processing pipeline based on template matching. (B) DeepFinder: a multi-class approach able to localize several particles of different macromolecular species in one pass.

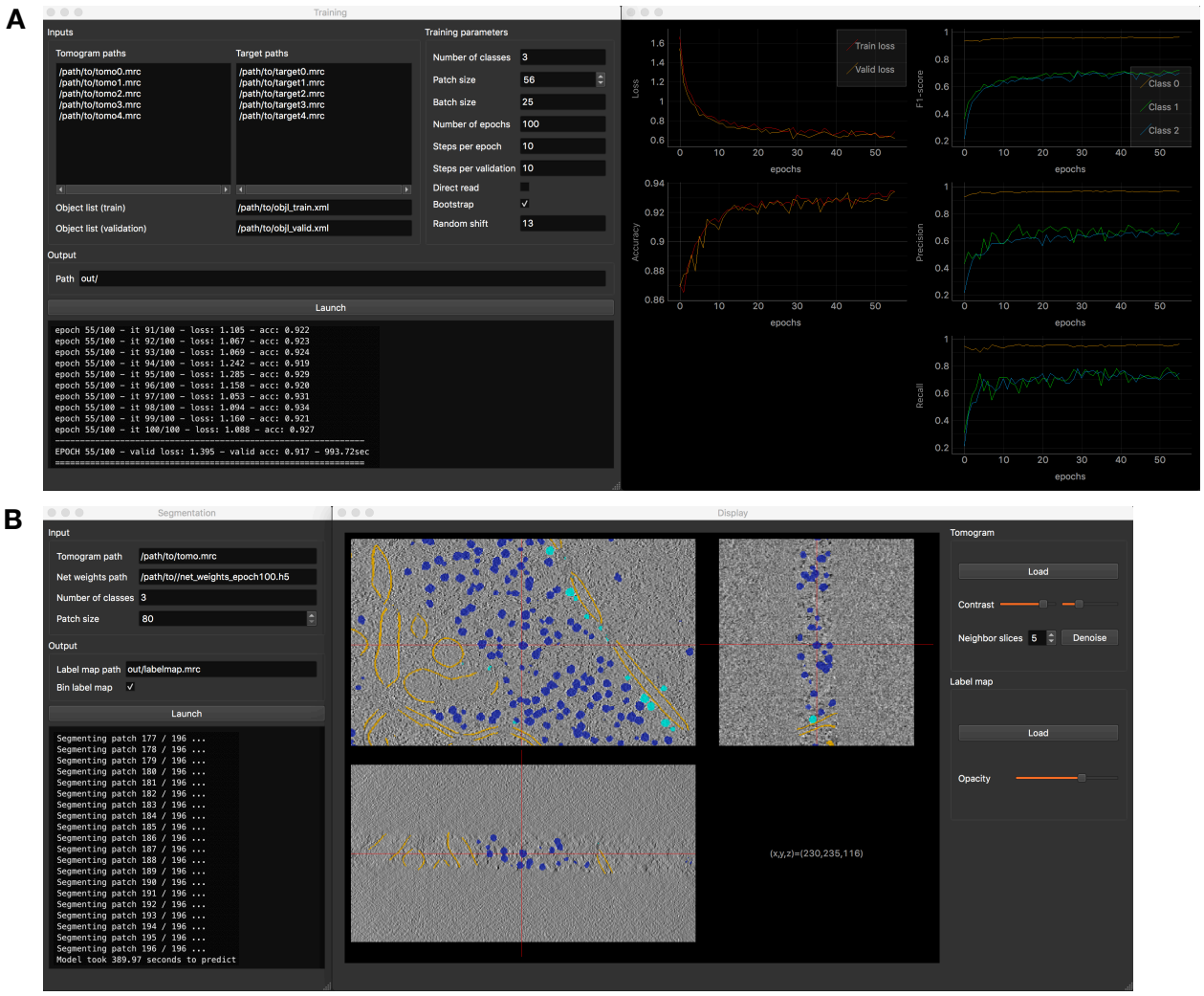

Figure 7: **DeepFinder graphical user interface.** (A) Training interface composed of a first window for parametrising the procedure, and a second window for displaying the training metrics in real-time. (B) Segmentation interface which also opens a data visualisation tool. This tool allows one to explore the tomogram with super-imposed segmentation. In addition, DeepFinder also incorporates interfaces for tomogram annotation, target generation and clustering (see the documentation at `https://gitlab.inria.fr/serpico/deep-finder` for more information).

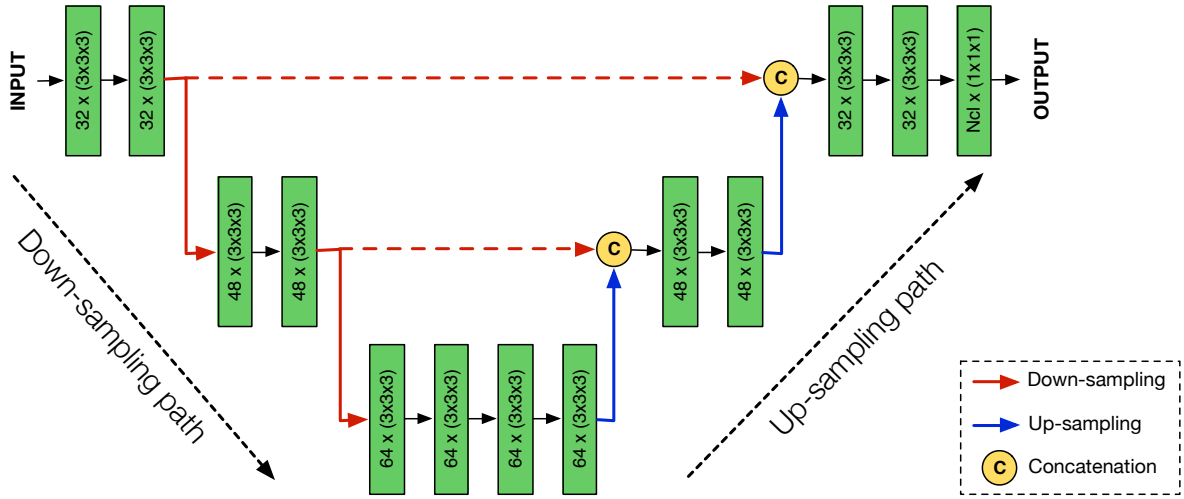

Figure 8: **U-net architecture used in DeepFinder.** Each green box represents a convolutional layer. The number of filters n and the filter size s is labeled as $n \times (s \times s \times s)$.

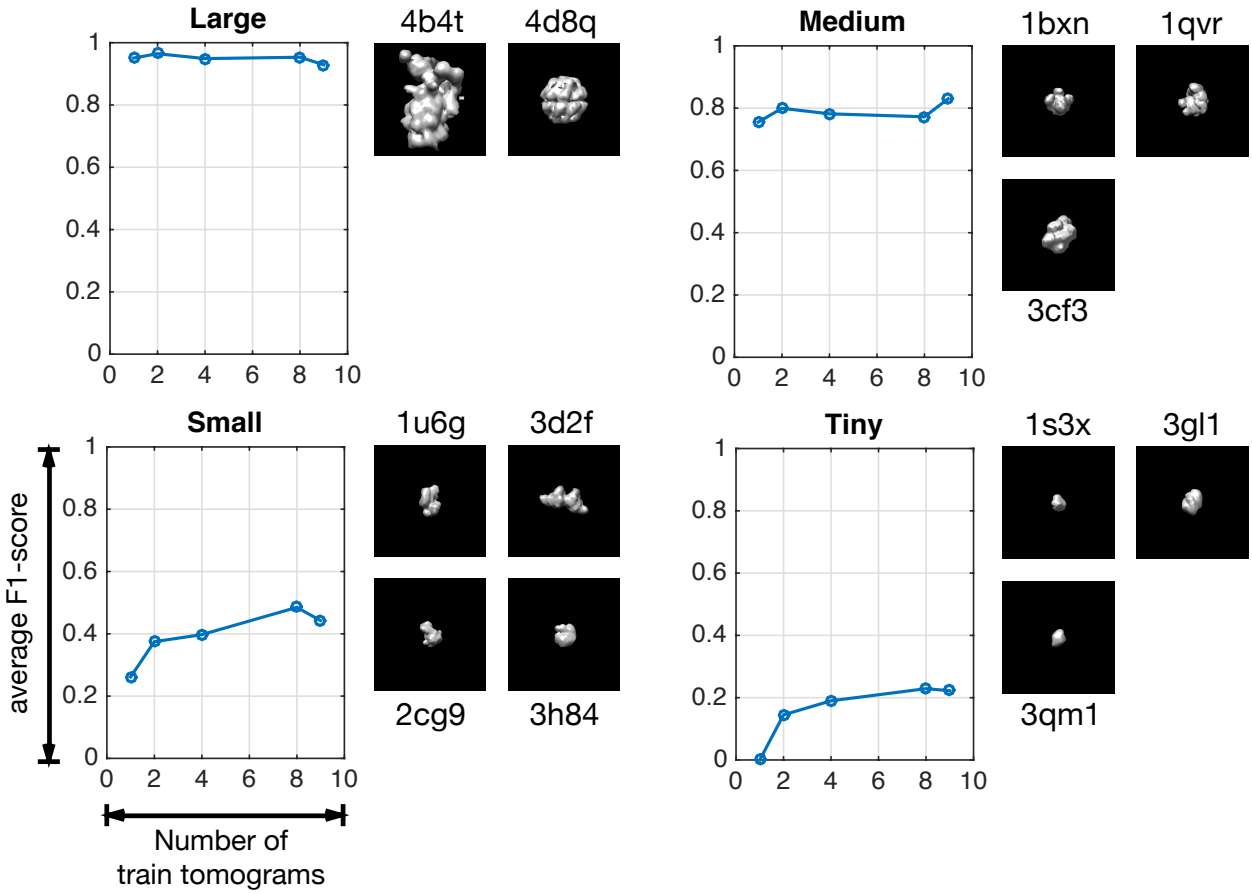

Figure 9: **Evolution of F_1 scores with respect to sizes of the training sets (number of tomograms) on the synthetic SHREC'19 dataset.** Each tomogram contains in average 208 macromolecules per class. The macromolecules have been categorized into 4 groups (large, medium, small and tiny) in order to illustrate the relationships between the size of training sets and the size of macromolecules.

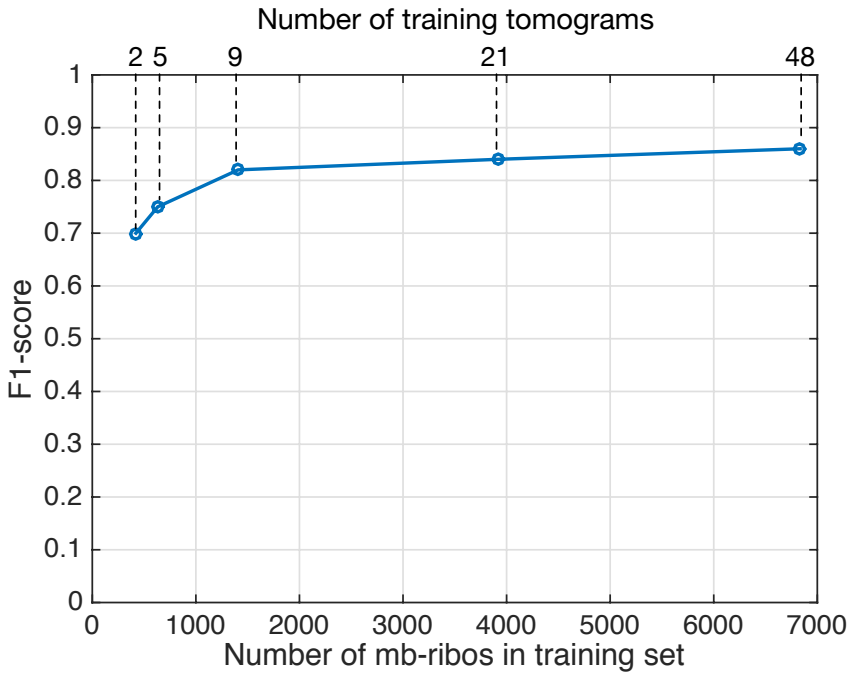

Figure 10: **Evolution of F_1 score with respect to the sizes of the training sets on real cryo-ET dataset (Dataset #2, *Chlamydomonas Reinhardtii*).** The F_1-scores are obtained by comparing the membrane-bound ribosomes (*mb-ribo* particles) found by DeepFinder to expert annotations.

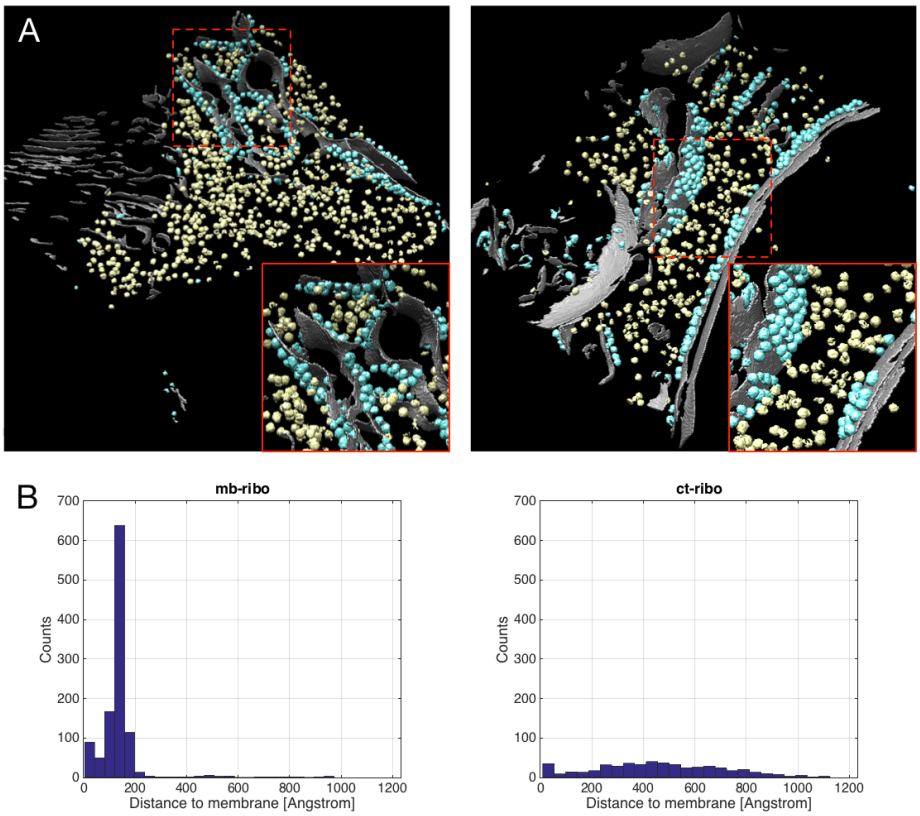

Figure 11: **Analysis of localization of *mb-ribo* and *ct-ribo* particles in the cell (Dataset #2).** (A) 3D voxel-wise classification of experimental tomograms obtained by DeepFinder: most *mb-ribo* particles (in blue) are close to the cell membranes (in gray), while the *ct-ribo* particles are not. (B) Histograms of distances of *mb-ribo* particles (left) and *ct-ribo* particles (right) with respect to the nearest cell membranes computed from the whole test set (8 tomograms).

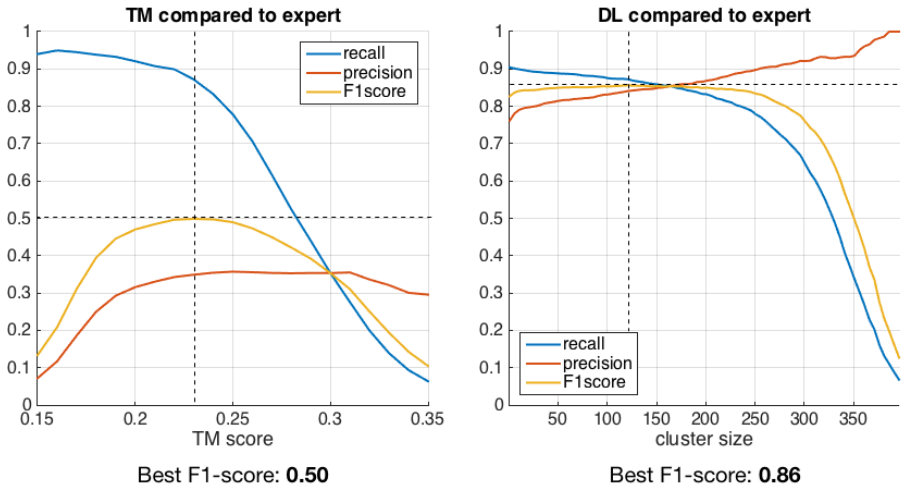

Figure 12: **Quantitative analysis of overlap with expert annotations (Dataset #2).** We varied the thresholds of template matching (left) and DeepFinder (right) to compute the *Recall*, *Precision* and F_1-score curves. template matching and DeepFinder have good *Recall* values but template matching has a lower *Precision* than DeepFinder. This suggests that template matching can be recommended to select many candidates but a time-consuming post-classification is required to improve *Precision*. DeepFinder has much higher *Precision* values, which confirms the results on synthetic dataset (SHREC'19 challenge).

Appendix A

Study of the complementarity of expert annotations and particles found by DeepFinder

In this section, we show how DeepFinder, when combined with template matching, can substantially improve the localizations and the number of particles in experimental cryo-ET data. We investigated the set S_{DF} of *mb-ribo* particles found by DeepFinder and considered the two following hypotheses:

H_1 : The set S_{DF} contains too many false positives.

H_2 : The set contains particles with a low quality (due to noise and blur), suggesting that our method found supplementary "true" *mb-ribo* particles, but missed or discarded during the supervised annotation.

So we aligned and computed the average of *mb-ribo* particles taken in the sets S_{DF}, S_E, $S_{DF} \setminus S_E$.

In Fig. 13, we display the resulting subtomogram average denoted $\mathbf{A}_{DF}$ computed from the set S_{DF} (356 particles) found with DeepFinder. Since it was not guaranteed that all the particles involved in the average were actual *mb-ribo* particles, we have analyzed the difference between $\mathbf{A}_{DF}$ and the average $\mathbf{A}_{DF}^{ref}$ computed from 356 true *mb-ribo* particles (expert annotations) randomly picked from the the set $S_E \cap S_{DF}$. Also, in order to check if $\mathbf{A}_{DF}$ was not biased by the reference template used for subtomogram alignment (as explained in Henderson *et al.*[44]), we considered another average denoted $\mathbf{A}_{DF}^{\circ}$ computed from 356 subtomograms picked from random positions. We compared $\mathbf{A}_{DF}$, $\mathbf{A}_{DF}^{ref}$, and $\mathbf{A}_{DF}^{\circ}$ visually and quantitatively by assessing the underlying structure resolution (Fig. 13). We noticed that $\mathbf{A}_{DF}$ has a lower resolution (36.4Å) than $\mathbf{A}_{DF}^{ref}$ (33.5Å). This suggests that $\mathbf{A}_{DF}$ likely contains false positives and/or very noisy particles. Nonetheless, $\mathbf{A}_{DF}$ has a higher structure resolution than $\mathbf{A}_{DF}^{\circ}$ (48.6Å), suggesting that the reference template bias is not significant. Moreover, a ribosome pattern visually appears in $\mathbf{A}_{DF}$. This means that DeepFinder has actually found additional *mb-ribo* particles that had been missed or discarded during the annotation process.

To perform a fair comparison, we applied a similar procedure to $S_E \setminus S_{DF}$, i.e. to the set of *mb-ribo* particles annotated by the expert but missed by DeepFinder. As before, we computed $\mathbf{A}_E$ from the 220 particles belonging to $S_E \setminus S_{DF}$, $\mathbf{A}_{DF}^{ref}$ by randomly sampling 220 *mb-ribo* particles from $S_E \cap S_{DF}$, and $\mathbf{A}_{DF}^{\circ}$ from 220 random positions. The resulting structure resolutions are very similar to the one achieved with DeepFinder. Here again, $\mathbf{A}_E$ has a lower resolution (37.6Å) than the reference $\mathbf{A}_E^{ref}$ (34.2Å). It is noteworthy that in both cases, the *mb-ribo* particles belonging to $S_E \cap S_{DF}$ lead to a higher resolution than those obtained with the complementary sets $S_E \setminus S_{DF}$ and $S_{DF} \setminus S_E$. This study clearly demonstrates that combining sets obtained from different methods is a good strategy to increase the confidence levels of detections. The lower structure resolutions of $\mathbf{A}_{DF}$ and $\mathbf{A}_E$ suggest that the sets $S_E \setminus S_{DF}$ and $S_{DF} \setminus S_E$ are made of more heterogeneous particles, potentially including particles depicting unusual structural conformations. This study emphasizes that template matching and Deep learning-based methods can be combined to better investigate experimental cryo-electron tomograms. For instance, the set of common particles found by the two methods enables to focus on "suspicious" particles found either by the experts (assisted by template matching) or by DeepFinder. In Fig. 5, we show that, while the Fourier Sell Correlation (FSC) curves obtained with DeepFinder are below the curve obtained with expert annotations, it is hasty to definitely decide that the additional macromolecules are not ribosomes (Fig. 13). The structure resolutions are actually lower than expected, but mainly because the particles are corrupted by noise and other sources of signal degradation. It would be appropriate that experts carefully analyze the supplementary noisy particles since they may be valuable *mb-ribo* candidates. Finally, it appeared that the number of actual *mb-ribos* was higher than expected: in our test set, we have detected $+20.5\%$ of *mb-ribo* particles when compared to the S_E set. In summary, DeepFinder found additional noisy *mb-ribos* that were missed or discarded during the annotation process.

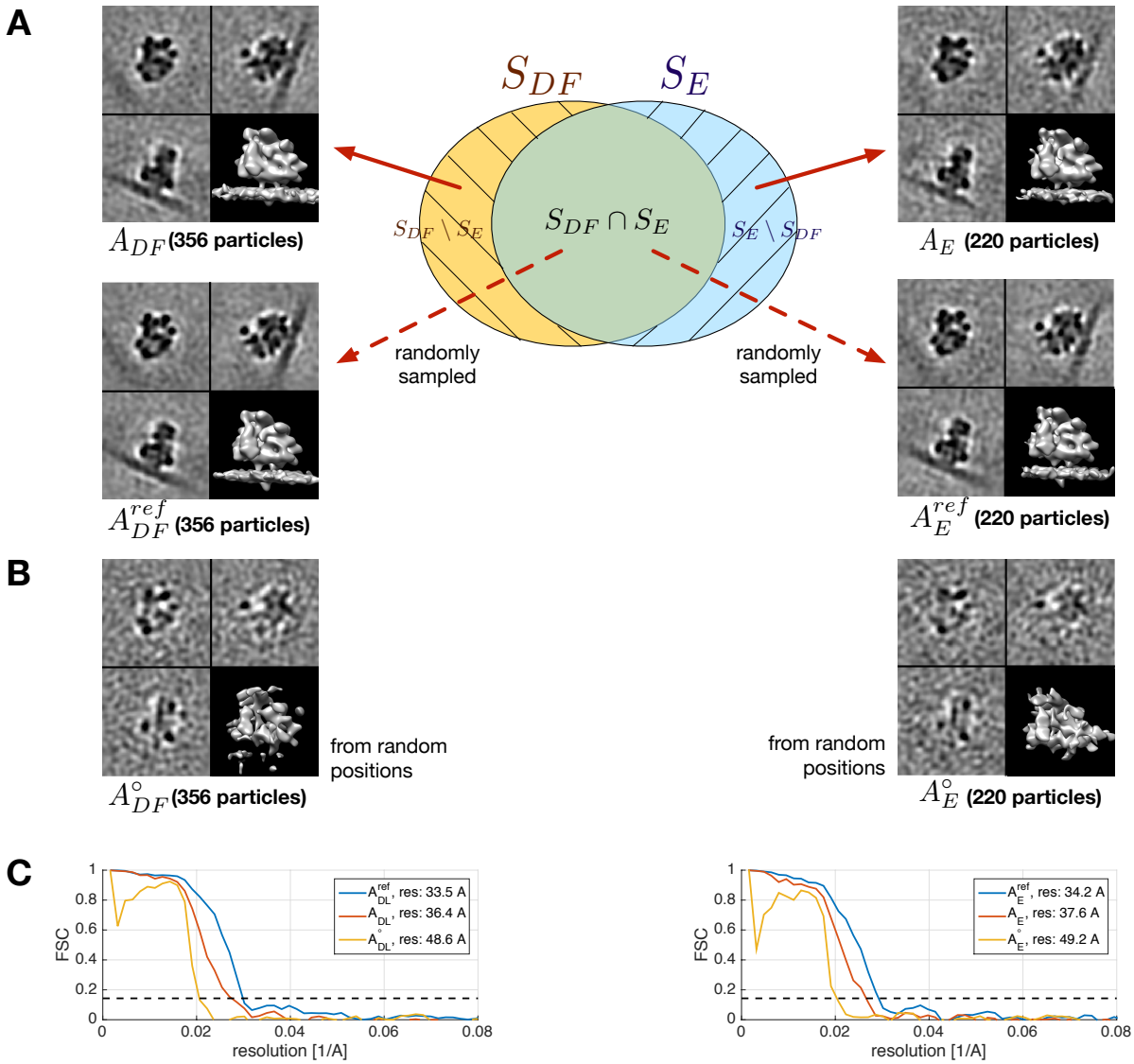

Figure 13: **Analysis of consensus decisions and overlap sets (Dataset #2).** (A) The central diagram represents the overlap between the *mb-ribo* sets S_{DF} (found by DeepFinder) and S_E (annotated by expert). Thus, $S_E \cap S_{DF}$ is the subset of *mb-ribo* particles found by both DeepFinder and the experts, $S_{DF} \setminus S_E$ is the subset of *mb-ribos* found by DeepFinder only (and missed by the experts), and $S_E \setminus S_{DF}$ is the subset of *mb-ribo* particles found by the experts only (and missed by DeepFinder). The origin of red arrows pointing to the subtomogram averages $\mathbf{A}_{DF}$, $\mathbf{A}_{DF}^{ref}$, $\mathbf{A}_E$, $\mathbf{A}_E^{ref}$ indicate the particles subsets which served to compute the averages. A ribosome density is clearly visible in $\mathbf{A}_{DF}$, therefore one can safely assume that the FP rate in $S_{DF} \setminus S_E$ is low. (B) The subtomogram averages $\mathbf{A}_{DF}^{\circ}$ and $\mathbf{A}_E^{\circ}$ have been computed using subtomograms sampled from random positions. These averages serve to estimate a lower bound for the FSC curve. The correlation values equal or below this bound are considered "noise" values, and are caused by alignment bias[44]. (C) The averages $\mathbf{A}_{DF}^{ref}$ and $\mathbf{A}_E^{ref}$ have both led to a higher resolution than $\mathbf{A}_{DF}$ and $\mathbf{A}_E$, implying that the *mb-ribo* particles in the set $S_{DF} \setminus S_E$ and in the set $S_E \setminus S_{DF}$ are more heterogenous than the *mb-ribo* particles in the set $S_{DF} \cap S_E$. Also, $\mathbf{A}_{DF}$ and $\mathbf{A}_E$ have led to a higher resolution than $\mathbf{A}_{DF}^{\circ}$ and $\mathbf{A}_E^{\circ}$, meaning that the impact of alignment bias is not significant.

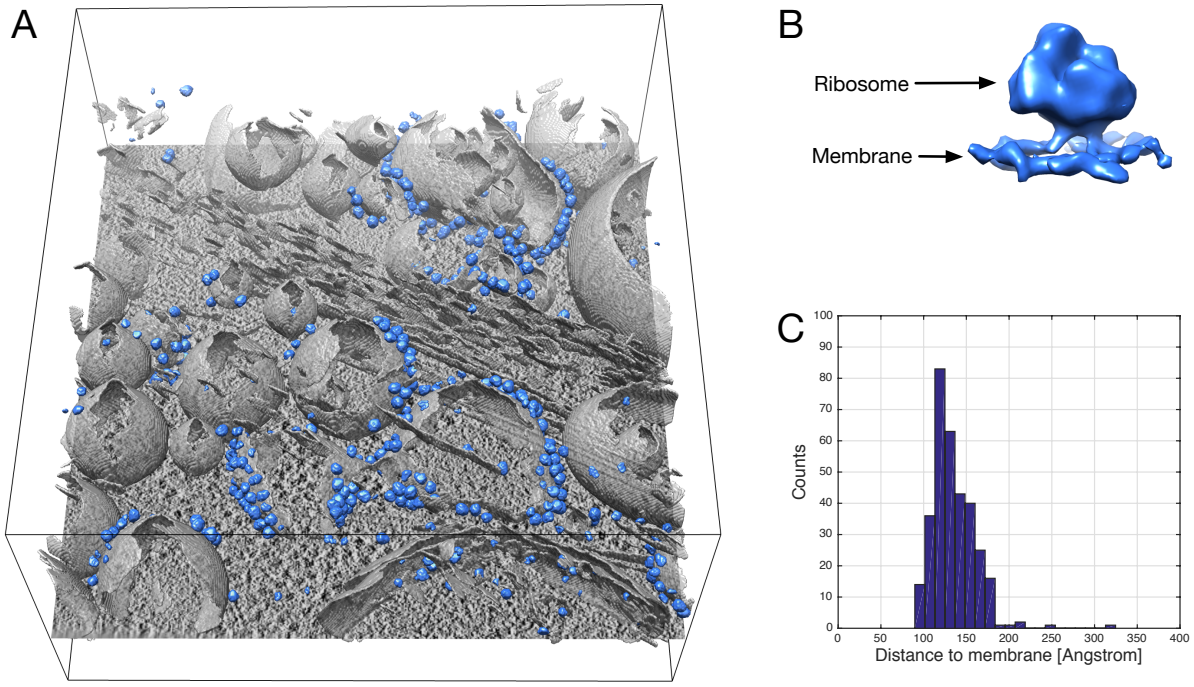

Figure 14: **Transfer learning on P19 cells**. DeepFinder has been trained on the Chlamydomonas (algae) dataset and applied on a tomogram (EMD-10439, `https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-10439`) of P19 cells (mouse). Although the ribosome has a different structure for the two species for a given voxel size (13.68Å), the structures are similar enough for DeepFinder to identify and localize *mb-ribo* particles in a P19 cell. (A) Tomographic slice with both the superimposed segmented cell membrane (gray) and *mb-ribo* particles (blue). (B) Average density of *mb-ribo* particles. (C) Histograms of distances of *mb-ribo* particles with respect to the nearest cell membrane. The histogram mode maximum is located at 136.8Å, which corresponds to the ribosome radius.