3D ConvNet improves macromolecule localization in 3D cellular cryo-electron tomograms

Emmanuel Moebel, Antonio Martinez, Damien Larivière, Julio Ortiz, Wolfgang Baumeister, Charles Kervrann

3D ConvNet improves macromolecule localization in 3D cellular cryo-electron tomograms

Emmanuel Moebel[1], Antonio Martinez[2], Damien Larivière[3], Julio Ortiz[2], Wolfgang Baumeister[2], Charles Kervrann[1]

[1] Serpico Project-Team, Inria-Rennes / CNRS-UMR 144
Inria, CNRS, Institut Curie, PSL Research University
Campus Universitaire de Beaulieu
35042 Rennes Cedex France

[2] Max Planck Institute of Biochemistry
Am Klopferspitz 18
Martinsried, Germany

[3] Fourmentin-Guilbert Scientific Foundation
2 Av du Pavé Neuf
93160 Noisy-le-grand France

Abstract

Cryo-electron tomography (cryo-ET) allows one to capture 3D images of cells in a close to native state, at sub-nanometer resolution. However, noise and artifact levels are such that heavy computational processing is needed to access the image content. In this paper, we propose a deep learning framework to accurately and jointly localize multiple types and states of macromolecules in cellular cryo-electron tomograms. We compare this framework to the commonly-used template matching method on both synthetic and experimental data. On synthetic image data, we show that our framework is very fast and produces superior detection results. On experimental data, the detection results obtained by our method correspond to an overlap rate of 86% with the expert annotations, and comparable resolution is achieved when applying subtomogram averaging. In addition, we show that our method can be combined to template matching procedures to reliably increase the number of expected detections. In our experiments, this strategy was able to find additional 20.5% membrane-bound ribosomes that were missed or discarded during manual annotation.

1 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 opposition to previous cell models where macromolecules where 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 a unique imaging technique capable of producing 3D views of large portions of a cell while having enough resolution to localize and identify macromolecules. Cryo-ET allows avoiding the use of markers, such as fluorescent probes used in light microscopy, which could perturb the cell. First, samples are vitrified in order to preserve both the spatial distribution and the structure of macromolecules in the cell during the image acquisition process. This technique enables the study of cells in a close to native state, and has thus the potential to create a molecular atlas from all the detected components (macromolecules, membranes) observed in cryo-tomograms. However, the analysis of such images is challenging due to poor signal-to-noise ratios and imaging artifacts caused by limited-angle tomography. As a consequence, cryo-ET analysis is heavily dependent on computational tools for interpreting the image content.

A well-established method for localizing macromolecules is template matching (TM) [Best et al., 2007], where a template containing the macromolecule of interest is used to explore a given 3D cryo-tomogram. While TM is efficient for localizing large macromoelecules such as ribosomes, it is necessary to apply several image post-processing and analysis methods to decrease the false positive (FP) rate (see Fig. 1). These methods include selection of regions of interest (e.g. areas in the cytoplasm), and thresh-

olding the TM score (e.g. correlation score) values. Additional difficulties also arise when TM is used to detect and localize several types of macromolecules that are structurally similar or specific macromolecule states, like binding states of ribosomes (e.g. membrane-bound vs cytoplasmic ribosomes). In general, TM is applied several times to detect each subclass of interest. Unfortunately, the score values are not selective enough to allow one to perform a satisfying classification, especially when the number of considered subclasses is high. Therefore the sub-volumes containing the detected macromolecules (also named particles) of interest are manually analyzed or automatically post-processed by sophisticated classification algorithms [Förster et al., 2008]. Such complex and time-consuming processing chains are routinely applied to accurately localize macromolecules and to identify the related native structure in the cell. Note that each TM and sub-volume classification round can each take 10 to 30 hours of computation on specialized CPU clusters.

In this paper, we propose an unified deep learning-based framework [Lecun et al., 2015] to jointly and fastly localize and classify macromolecules in cryo-ET. Deep learning (DL) is a set of machine learning techniques capable to produce state-of-the-art results in various fields (e.g. computer vision [Lecun et al., 2010], language processing [Hinton et al., 2012], super-resolution microscopy [Ouyang et al., 2018] and bioinformatics [**?**]). In particular, convolutional neural networks (CNN) are able to produce impressive results in image analysis, including image classification [Krizhevsky et al., 2012], segmentation [Long et al., 2014] and object recognition [Szegedy et al., 2013]. A neural network is generally composed of successive neuron layers, each transforming incoming data and transferring it to the next layer. The neurons can be seen as 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 (neighboring pixels have similar values). Thus, a neuron can be thought of as a filter, and a neuron layer as a filter bank. The role of a layer is to automatically extract features from the data. Applying sequentially the layers enables to progressively compute more abstract features, which results in a hierarchical representation of the data. The underlying idea is to learn high-level features from low-level features, which allows a computer to understand complex interactions from basic patterns. 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. circles, triangles), objects (e.g. eyes, ears), object ensembles (e.g. faces) and object conditions (e.g. face gender). Those powerful data representations are learned automatically from the data, and tend to be more efficient than conventional handcrafted representations, which require human ressources and are time consuming to design.

Deep learning has been recently investigated to learn high-level generic features in cryo-electron microscopy (cryo-EM). In [Wang et al., 2016], the authors proposed first a CNN architecture to au-

tomatically detect particles in single particle cryo-EM 2D micrographs. The computational method was designed to detect a unique object class in 2D images depicting stationary noisy backgrounds. In [Chen et al., 2017], a CNN architecture was used for the first time to analyse cryo-ET data; the authors proposed a DL framework especially dedicated to tomogram segmentation. They posed the segmentation of cryo-ET images as a set of N binary voxelwise classification problems.

An ensemble of N 2D CNN (one per class) is applied slice per slice on the 3D tomograms. The modeling and computation is clearlu sub-obtimal, but the proposed practical implementation (available in the EMAN2 package [Tang et al., 2007]) allows one to satistfyingly find several object classes such as cell membranes, microtubules and ribosomes in tomograms. While the authors also show that the proposed DL framework can be used to pick up ribosomes for subtomogram averaging, an additional post-classification step is necessary to get more satisfying reconstruction results. Unfortunately, no quantitative analysis is presented in [Tang et al., 2007] to assess the localization accuracy of detected particles and the actual resolution of macromolecule structures once subtomogram averaging is performed. Unlike [Wang et al., 2016, Chen et al., 2017], we consider a fully 3D CNN architecture in order to more accurately and reliably detect 3D particles in a native crowded cell environment as illustrated in Fig. 2. In addition, our network is also capable to handle multiple object subclasses at the same time. We especially demonstrate that manipulating jointly a higher number of object subclasses/classes is 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 previous approaches [Tang et al., 2007]. Besides, while training in [Chen et al., 2017] is faster (10 min per class) than with our method (12 hours), our processing time is at least twice as fast for one class and remains constant when the number of classes increases (see Fig. 3). In our experiments, we compared our 3D CNN architecture to TM in order to emphasize how the ways in which they were designed and their potentials differ. Unlike TM, our CNN framework is able to localize multiple types of macromolecule at once when applied to a given cryo-tomogram. We also show quantitatively for the first time how the CNN framework, when combined to TM guided procedures, can substantially improve the localization sensitivity of the structure of interest on real cryo-ET data.

The remainder of this paper is organized as follows. In Section 2, we present our deep learning framework. In Section 3, we explain our experimental setup focusing on ribosomes and the results. We show that DL outperforms TM on synthetic images and how TM and DL can be combined to improve the localization sensitivity in real data. In Section 4, we discuss the potential of DL in cryo-ET and future work.

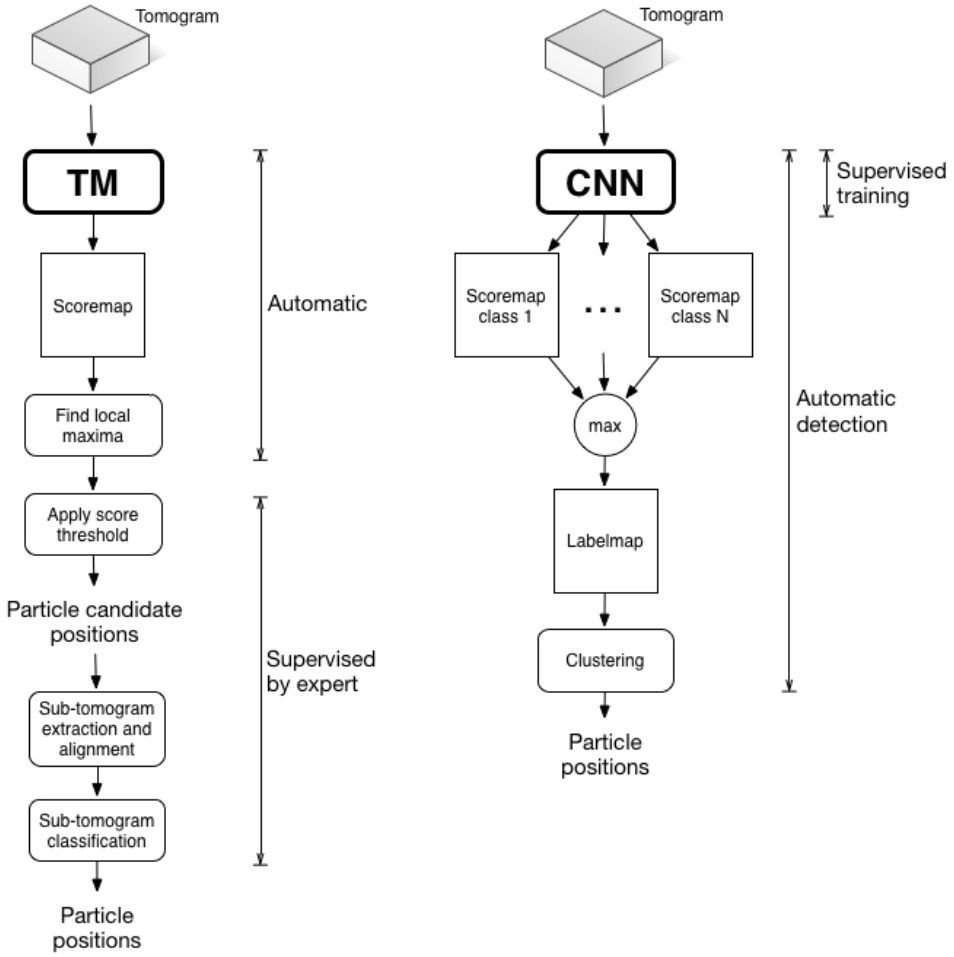

Figure 1: Comparing workflows: on the left, our DL framework and on the right, the common processing chain involving TM. Our approach is multi-class, whereas the TM processing chain needs to be applied once for each class.

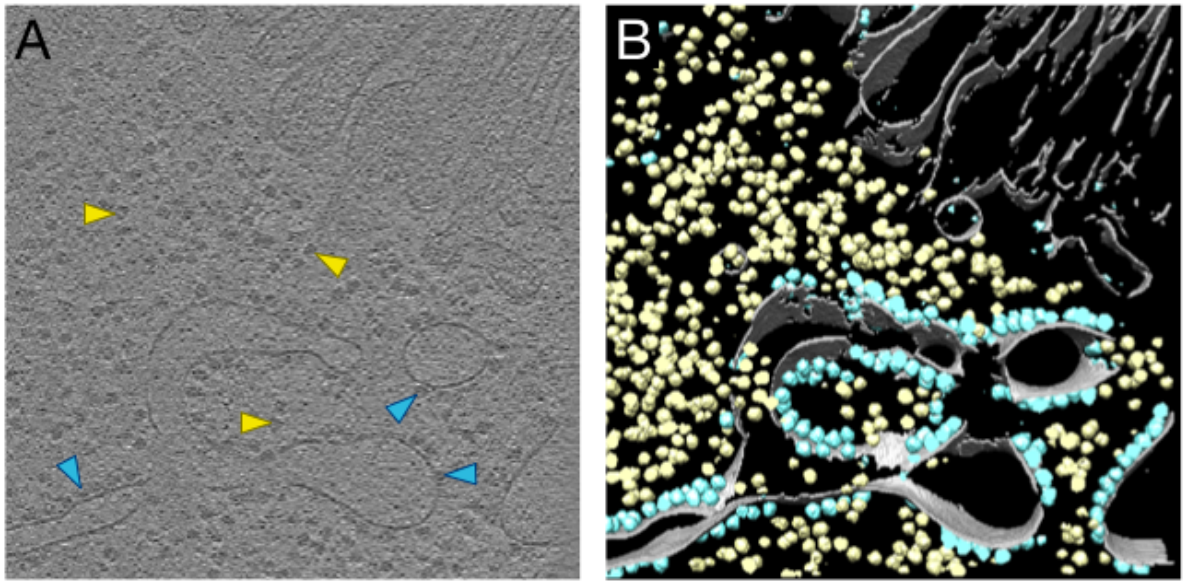

Figure 2: Chlamydomonas cell. (A) Tomogram slice; blue arrows indicate *mb-ribos*; yellow arrows indicate *ct-ribos*. (B) Corresponding voxelwise classification obtained by our 3D CNN, performed for 3 classes: *mb-ribos* (blue), *ct-ribos* (yellow) and *membrane* (gray).

2 Method

2.1 Localizing multiple objects with a CNN architecture

Given a training set of object classes, we propose a supervised 3D CNN-based method to classify the 3D tomogram voxels into several types of macromolecules or states of a given macromolecule. A clustering algorithm is then applied to aggregate voxels into clusters and to determine the position of particles (gravity center of clusters) in the volume (see Fig. 1). The detected particles are further exploited for subtomogram averaging.

2.1.1 Step #1: Multiclass voxelwise classification

Our objective is to provide a classification map for which each voxel in the 3D map is assigned an object class. The convolutional neural network is usually designed to produce a single output label: it exploits global information by progressively down-sampling the image layer by layer, in order to preserve only relevant information and reduce computation. The disadvantage is that by doing so, the network loses the local information needed for voxelwise classification tasks. While the network is able to reliably decide if an object is present or not in an image, it is not able to accurately estimate the

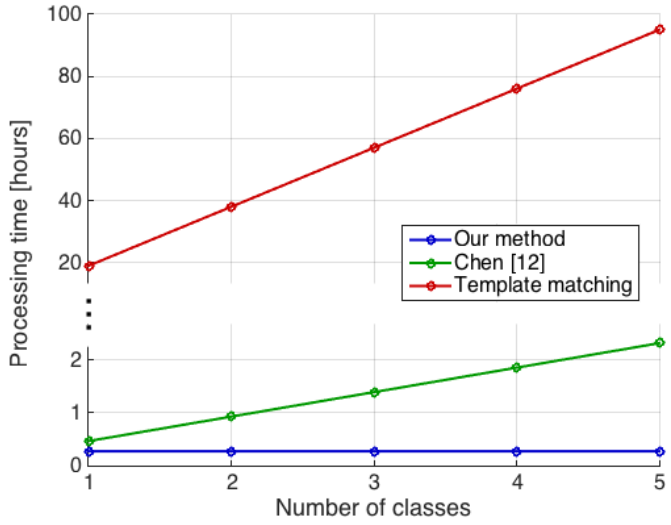

Figure 3: Time needed to process a tomogram of size $928 \times 928 \times 464$ voxels, w.r.t. the number of classes (once training is completed). We compare our method to [Chen et al., 2017] and to template matching. We do not consider post-processing in displayed time values (i.e. without clustering step for our method, and without connected component analysis and post-classification for [Chen et al., 2017]). We use a Tesla K80 GPU for our method and a 32-core CPU cluster for template matching, while in [Chen et al., 2017] the authors used a 12-core CPU workstation.

position of the object. As opposed to this conventional approach, fully convolutional networks (FCNN) [Long et al., 2014] overcome this difficulty by explicitly combining global and local information in order to provide high resolution label maps. FCNNs involve more interactions in the network and adapt conventional classification networks to perform more robust image voxelwise classification. This idea was exploited in [Ronneberger et al., 2015] and then adapted in [Milletari et al., 2016] for 3D image analysis. As described in [Ronneberger et al., 2015, Milletari et al., 2016], our architecture consists of a down-sampling path needed to generate global information and a up-sampling path used to generate high-resolution outputs, i.e. local information (see Fig. 4). Down-sampling is performed with max-pooling layers (factor 2) and up-sampling with up-convolutions [Long et al., 2014] (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. Unlike [Milletari et al., 2016], our architecture is not so "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 [Simonyan and Zisserman, 2015]. 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 an object 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 [Milletari et al., 2016]. Accordingly, we added two supplementary convolutional layers in the lowest stage of our architecture. To complete the description, we use rectified linear units (ReLU) [Krizhevsky et al., 2012] 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 mandatory in order to interpret the network outputs as probabilities for each class.

In summary, our proposed CNN architecture is capable of robustly classifying the cryo-ET tomogram into N subclasses/classes with a high accuracy. Given the voxelwise classification map, the next step consists in estimating the position of each individual object, as described in the next section.

2.1.2 Step #2: Clustering for macromolecule localization

Given the multiclass voxelwise classification map and classification errors, our objective is determine the position of each particle corresponding to a state of a given macromolecule or several types of macromolecules. The voxel labels should be ideally spatially well clustered into well distinct 3D connected components, each cluster corresponding to unique object/particle. Because of noise, non-

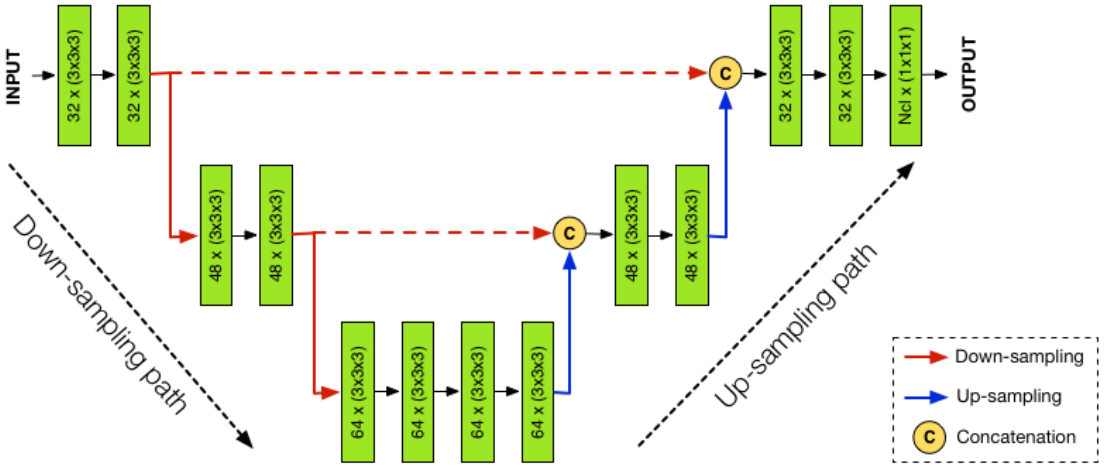

Figure 4: CNN architecture. In green convolutional layers labeled with (#filters × (filter size)). In the last layer, Ncl stands for the number of classes.

stationarities in the background, and artifacts in the tomogram, the CNN method generates isolated labels or very small groups of voxels, and groups that contain different label types. Post-processing is then necessary to assign an object class to a given position.

To address this issue, we apply a basic clustering approach. The clusters are built by aggregating neighboring voxels into objects and form 3D connected components. The smallest clusters are considered as false positives and are discarded. The cluster centroid is further used to estimate the object position. 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 or subclass/class is assigned to the detected particle. To address the computational issues, we used the popular mean-shift clustering algorithm [Comaniciu et al., 2002]. 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 object size. The K-means algorithm was not considered further since the number of clusters must be provided as input parameter by the user.

2.2 Training

In training step, the CNN is learned from pairs of tomograms and their corresponding voxelwise classification. In other words, the CNN needs every tomogram voxel annotated as member of the class of interest or as background. While voxelwise classification examples are naturally available for synthetic data, it is often not the case for experimental data. In our case, the experts accurately localized the

macromolecules of interest in several training tomograms. Actually, voxelwise classification cannot be performed manually for two reasons: i/ it is time consuming to label each voxel in hundreds of objects in 3D; ii/ the data is so noisy that the object borders are barely visible. To address this issue, we propose an original computational approach based on subtomogram averaging [Förster and Hegerl, 2007] to label voxels, only from the expert annotations corresponding to the spatial coordinates of macromolecules (see Fig. 5). Subtomogram averaging is a registration algorithm designed to obtain higher resolution structures by averaging thousands of aligned subvolumes containing the same structural unit. The labeled coordinates serve here as inputs to a subtomogram averaging procedure. The subtomograms around 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 of the macromolecule. From this density, 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 labeled position with the estimated 3D orientation. The resulting volume with well delineated macromolecules is then used as a target to train the parameters of the CNN architecture. It is worth noting that annotating the macromolecule with this semi-automatic approach saves time but may introduce "label noise" in the training. Indeed, we use an average shape to label the macromolecule, and we neglect structural macromolecule variability mainly localized in the object borders. Nonetheless, it has been shown that CNNs have a natural robustness to reasonable amount of "label noise" [Rolnick et al., 2017], which is also confirmed in our experiments.

Due to memory limitations, it is not feasible to load the whole tomogram set with the corresponding targets during training. Therefore, we randomly draw smaller 3D patches around macromolecules at each training iteration. The patch size should be large enough to capture sufficient context information; the macromolecule radius being 10 voxels, we choose a patch size of $56 \times 56 \times 56\times$ voxels. It is also common to use "data augmentation" when training a CNN; it allows to increase the training set artificially by applying geometric transform to the training images. In our approach, we implement "data augmentation" by applying a 180° rotation w.r.t. the microscope tilt-axis to each training example. Nevertheless, we do not use typical mirror operations or geometric deformations because the structure of expected objects is the principal clue in the detection problem. Also, we do not use random rotations because of the well-determined orientation of missing wedge artifacts, which is preserved when applying 180° rotations w.r.t. the tilt-axis. In our experiments, the CNN has been computationally trained for 6000 iterations with the ADAM algorithm, chosen for its good convergence rate [Kingma and Ba, 2014], using 0.0001 as learning rate, 0.9 as exponential decay rate for the first moment estimate and 0.999 for the second moment estimate. We use categorical cross-entropy as a loss function (see Fig. 6). The training has been performed on a Nvidia Tesla K80 GPU and took 12 hours of computation, which is reasonable knowing that for other tasks, CNN training can last

several days [Krizhevsky et al., 2012] [Simonyan and Zisserman, 2015]. In the next section, we will present the training datasets and the results of our CNN approach applied to both synthetic and real tomograms.

3 Results

The method has been evaluated and compared to TM on a synthetic and real datasets described below.

3.1 Description of data

3.1.1 Dataset #1: synthetic data

The data has been generated with the AV3 toolbox [Förster and Hegerl, 2007], using atomic densities from the PDB databank [rcsb.org]. Simulation parameters include a voxel size of 13.68 Å(Angströms), a defocus of $-6\mu m$, and several values of signal-to-noise (SNR) from 0.05 to 0.10, and tilt ranges from $\pm 50°$ to $\pm 70°$, with a tilt-increment of $2°$. In this experiment, nine classes of prokaryotic macromolecules have been chosen to depict varying amounts of inter-class similarity (see Fig. 7). The GroEL and ribosome classes are significantly different in size and shape, whereas different functional states of the proteasome and GroEL are structurally similar. In addition to these well-known macromolecules, we have introduced basic objects (ellipses, spheres, discs) in order to mimic not well-defined structures found in a cell, like membrane components, small molecules, and gold particles.

The training set consists of 510 macromolecules for each class, the validation set is composed of 240 macromolecules for each class, and the test set is made of 105 macromolecules for each class. The macromolecules and basic objects have been placed at random positions and orientations in the 3D volumes in order to simulate the crowded environment of a cell.

3.1.2 Dataset #2: experimental data

The second dataset is composed of 63 tomograms of Chlamydomonas Reinhardtii cells (see [Pfeffer et al., 2017] for details about data acquisition) and has been annotated for membrane-bound 80S ribosomes (denoted *mb-ribo* in the following) positions by experts. To get these annotations, the experts first used TM with a template generated from the dataset, using manually selected ribosomes and subtomogram averaging. Then they refined the TM results by applying subtomogram classification (CPCA [Förster et al., 2008]) and performing careful visual inspection (see Fig. 1). To reduce 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 $2°$.

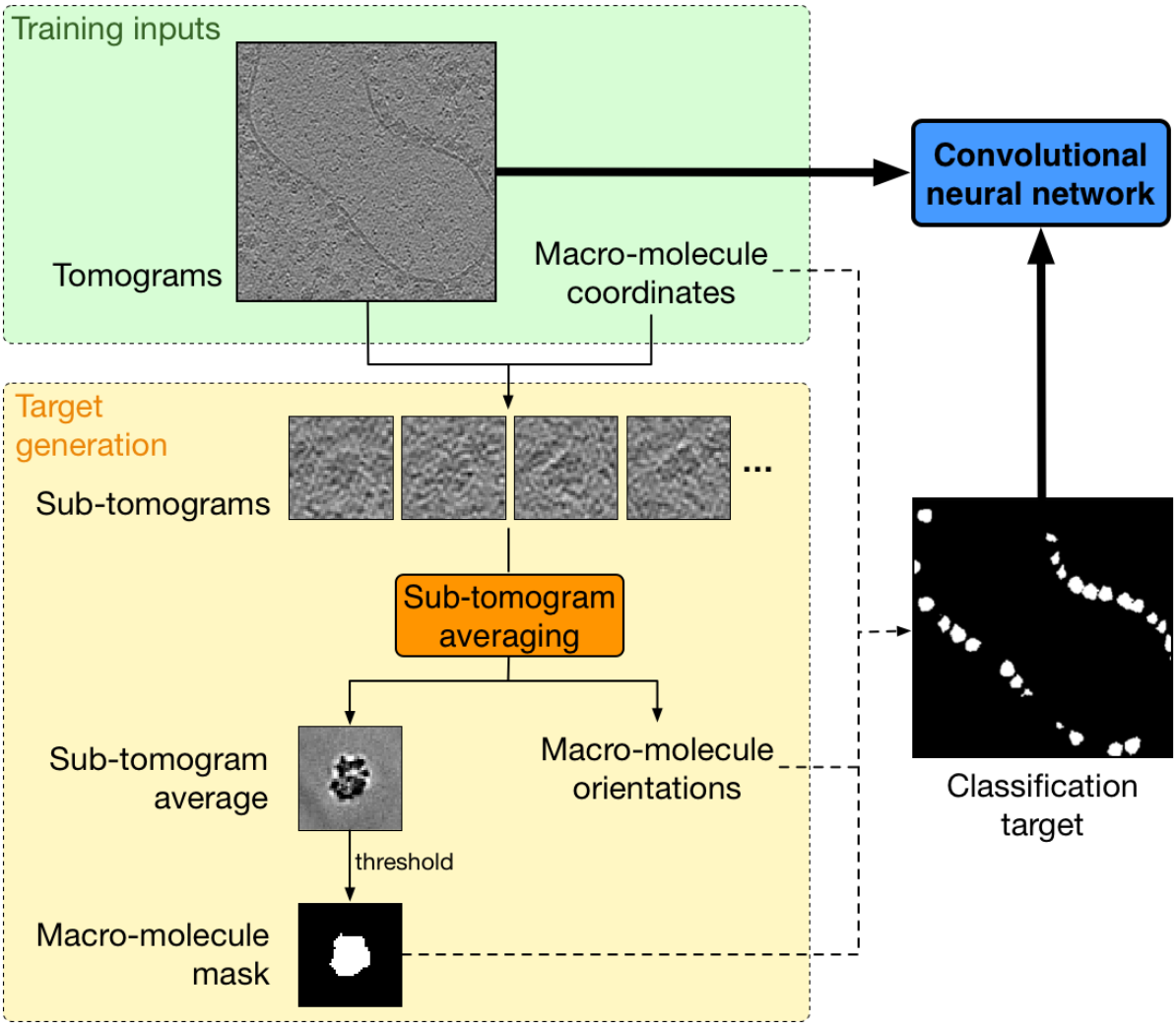

Figure 5: CNN training: this figure illustrates how to obtain voxelwise classification examples for training, using only position annotations.

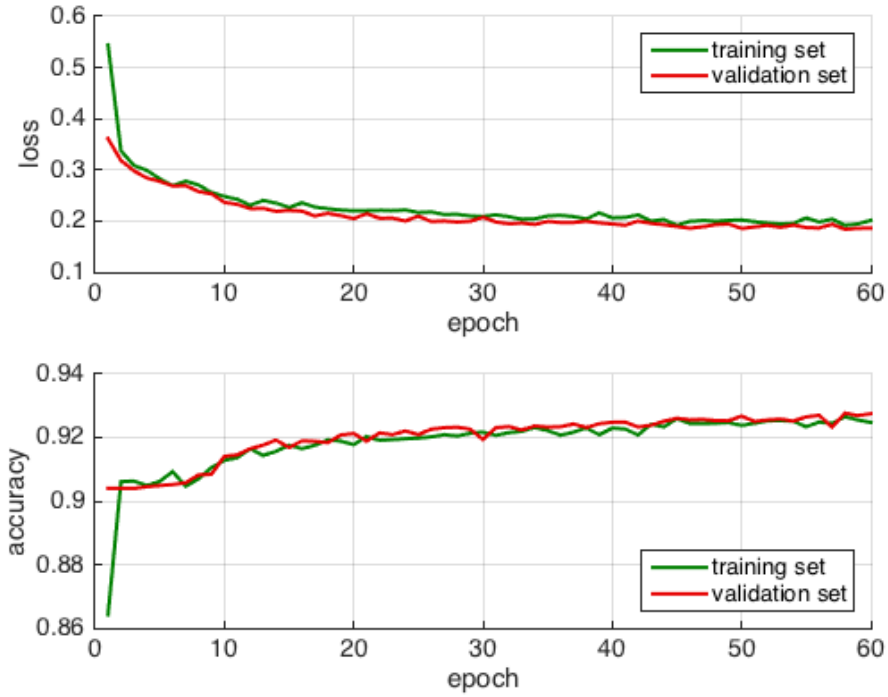

Figure 6: Evolution of loss and accuracy during training on dataset #2. These quantities are computed for the train set, as well as for the validation set, in order to estimate the generalization capabilities of our network. The curves for both sets should overlap, else it indicates overfitting (the network memorizes train samples instead of learning discriminating features).

In the end, the dataset has been annotated with 9487 *mb-ribos*. The subtomogram average computed from the total set of *mb-ribos* is the best model we have for 80S ribosomes in Chlamydomonas Reinhardtii cells, and is therefore used as a reference for subtomogram alignment. As this dataset was originally annotated and designed to study the (*mb-ribo*) class, we used computational tools to get examples corresponding to the cytoplasmic ribosome (denoted *ct-ribos*) class and the *membrane* class. First we added the *membrane* class by employing an algorithm dedicated to membrane segmentation [Martinez-Sanchez et al., 2014]. Meanwhile, we got (*ct-ribo*) examples by applying TM and selecting the most isolated candidates, located at a distance higher than 273.6Å (i.e. the ribosome diameter) to membrane components. The motivation behind adding new classes to the available annotations was twofold: on the one hand, our motivation was to demonstrate the multiclass ability of our method on real data; on the other hand, we noticed that the multiclass approach tends to improve the discriminating power of the network and the capacity to reliably detect *mb-ribos* in real tomograms. When trained by using only the *mb-ribo* examples, the network actually detects unwanted cytoplasmic 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 cytoplasmisc 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 in our protocol.

Dataset #2 has been arbitrarily split into training, validation and test sets. Training and validation sets have been sampled from 55 tomograms and consist of 5971 *mb-ribos* for training and 1493 *mb-ribos* for validation. The test set is composed of 8 tomograms annotated with 1736 *mb-ribos*.

3.2 Evaluation

The metric used to assess localization performance is the F1-score, as commonly used in detection problems. The F1-score 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):

$$R = \frac{\text{Number of TP}}{\text{Number of particles in the tomogram}}, \tag{1}$$

- Precision (also known as positive predictive value):

$$P = \frac{\text{Number of TP}}{\text{Number of localized particles}}. \tag{2}$$

The F_1 score defined as follows

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

allows one to evaluate performance by considering a single value.

In what follows, a detected particle is considered as a TP if is closer than 136.8Å (i.e. ribosome radius) to a ground truth object.

First, we quantitatively evaluated the performance of the multiclass localization method on dataset #1. Synthetic data are very helpful to objectively study the influence of SNR and tilt-ranges on results. To perform multiclass localization with TM (which is mono-class), TM has been applied once for each class with the corresponding templates. The templates have been computed from the atomic densities used to generate the dataset #1. A threshold applied to the TM score map is used to select the N best candidates for each class. The threshold is chosen so that the overlap with the ground truth is maximized. As recommended in [Best et al., 2007], if multiple templates compete for the same position, the class with the highest score is selected.

Dataset #2 was used to evaluate the performance of our method in real conditions. In our protocol, only the *mb-ribo* annotations have been provided by experts, and are therefore considered as the ground truth. Accordingly, we only quantify results for the *mb-ribo* class, while providing visual results for the *ct-ribo* and membrane classes. Unlike [Tang et al., 2007], 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) combined to the "0.143" threshold criterion.

3.3 Result analysis

3.3.1 Dataset #1: synthetic data

It turns out that DL outperforms TM for each class, in terms of F_1-score (see Fig. 7 A). An interesting result is that the scores achieved by DL are virtually perfect. In comparison TM scores are class dependent, ranging from 0.07 to 0.98. TM achieves good scores for the biggest macromolecules (ribosome 70S and FAS), but the performance is lower if the size of the macromolecule is small. Indeed, the confusion matrices (see Fig. 7 B) reveal that small targets like proteasomes and GroEL are often confused with background or decoy objects. In addition, TM has some difficulty to tackle inter-class similarity, especially for the three functional states of the proteasome (double bpa, single bpa and without bpa).

We evaluated the robustness of TM and DL methods by varying the signal-to-noise ratio (SNR) and tilt-range. Figure 8 plots the average F_1-score over the nine classes. As expected, we observe a performance drop for decreasing values of SNR and tilt-range. However, DL is remarkably stable and produces nearly constant scores. TM on the other hand loses 6% of F_1-score when the SNR decreases from 0.15 to 0.05, and 7% when the tilt-range decreases from $\pm 70°$ to $\pm 50°$. In all tested situations,

it turns out that DL is more robust to noise and missing-wedge than TM. In summary, the results on dataset #1 prove that DL is capable to provide better results than TM as soon as DL exploits ideal ground truth during training, that is with no "label noise". These results demonstrate that DL outperforms TM on synthetic noisy and MW corrupted data.

Notice that our DL approach is able to implicitly tackle the missing wedge (MW), while TM usually considers the so-called constrained cross correlation to handle the MW information. Actually, DL is capable to manage non-linear object deformations due to MW during training, without additional prior imposed by the experts.

3.3.2 Dataset #2: experimental data

Comparison of score values of TM and DL In the first part of experiments, we have carefully examined the scores produced by TM and DL methods. In Fig. 9, it is clear that the TM score map is much noisier than the maps generated by DL. Actually, the responses of TM based on the constrained cross correlation are very high at ribosome locations and at undesirable locations corresponding to highly contrasted structures with similar sizes (for instance see cell membrane in Fig. 9). TM tends to generate a lot of false positives in the cell. Consequently the experts need to apply post-processing techniques to select relevant information in order to exploit the TM results. Unlike TM, DL provides clean score maps and only depict high values in well-localized blobs. These results suggest that DL is more capable to properly learn the structure and geometry of complex macromolecules.

In order to further support this idea, we examined the distribution of local maxima values in each score map (see Fig. 9). A sharp mode (depicted in red in Fig. 9) can be regarded as an indicator to assess discrimination quality. As shown in Fig. 9, the mode for TM is weak, while for the DL *ct-ribo* class the mode is not very sharp either. However, for the DL *mb-ribo* class, the mode is more significant and sharp, suggesting a score less prone to ambiguity. This observation confirms the idea that the TM score is actually not very discriminating in general. More specifically, it is not a surprise if the score of the *ct-ribo* class is not as discriminating as the score of the *mb-ribo* class. This is probably related to the annotation quality used for learning, much more higher in the case of *mb-ribo* class. In our experiments, the *mb-ribos* annotations have been carefully performed by an expert.

In summary, DL produces more sharper scores than TM (constrained cross-correlation), when the training data is carefully labeled (we have a better performance for *mb-ribos* than for *ct-ribos*).

Evaluation of voxelwise multiclass classification Figure 10 (A) illustrates the ability of our DL approach to recognize objects and structures in experimental tomograms. Visually, the voxelwise multiclass classification makes sense: *mb-ribos* (in blue) are primarily located against cell membranes, whereas *ct-ribos* (in yellow) occupy the remaining space. We quantitatively measured for each class

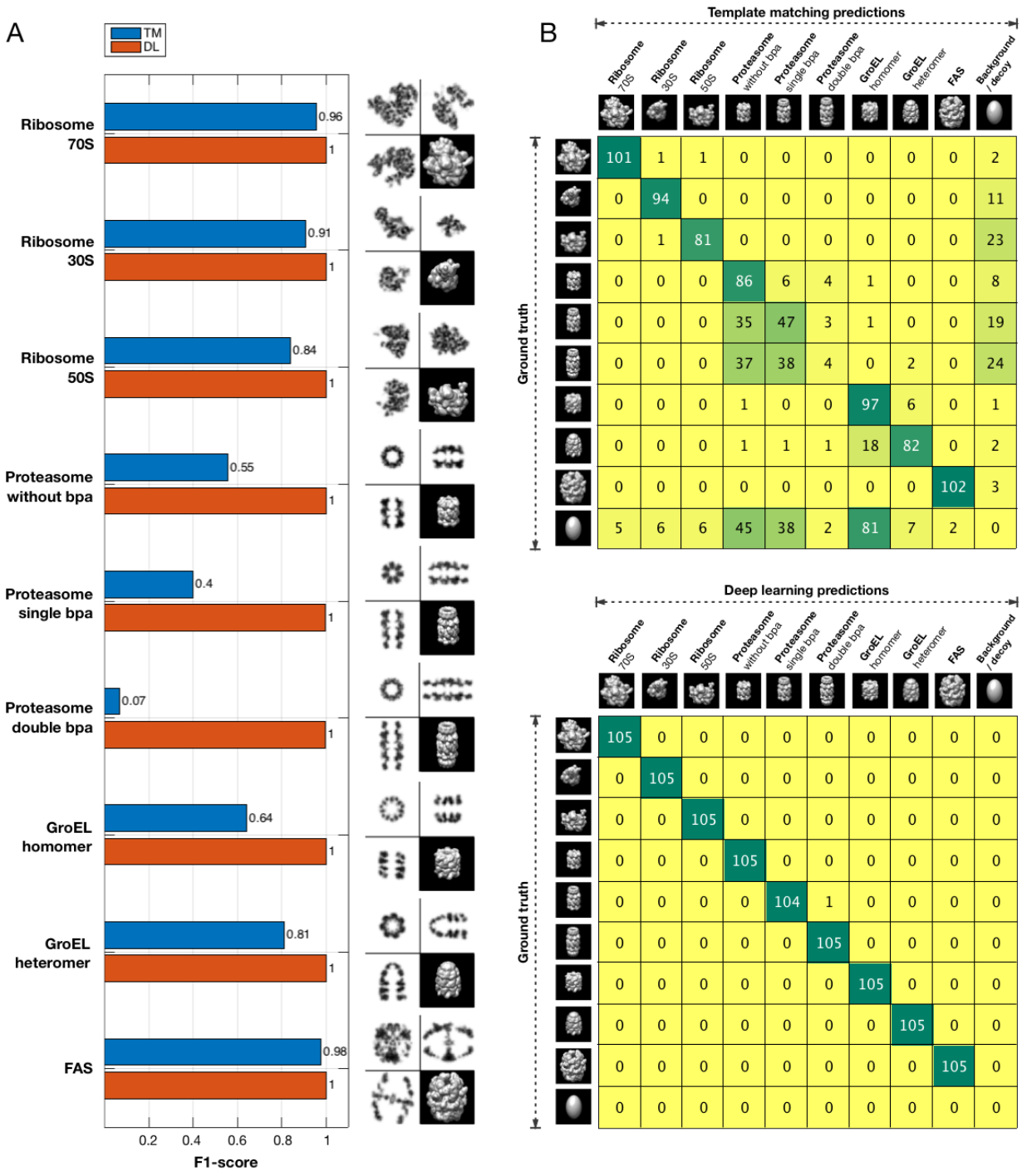

Figure 7: Dataset #1: comparing DL and TM performances per class. For this result we used SNR=0.1 and a tilt range of $\pm 60^\circ$. (A) displays the achieved F1-scores and (B) are obtained confusion matrices, illustrating the miss-classifications of both methods.

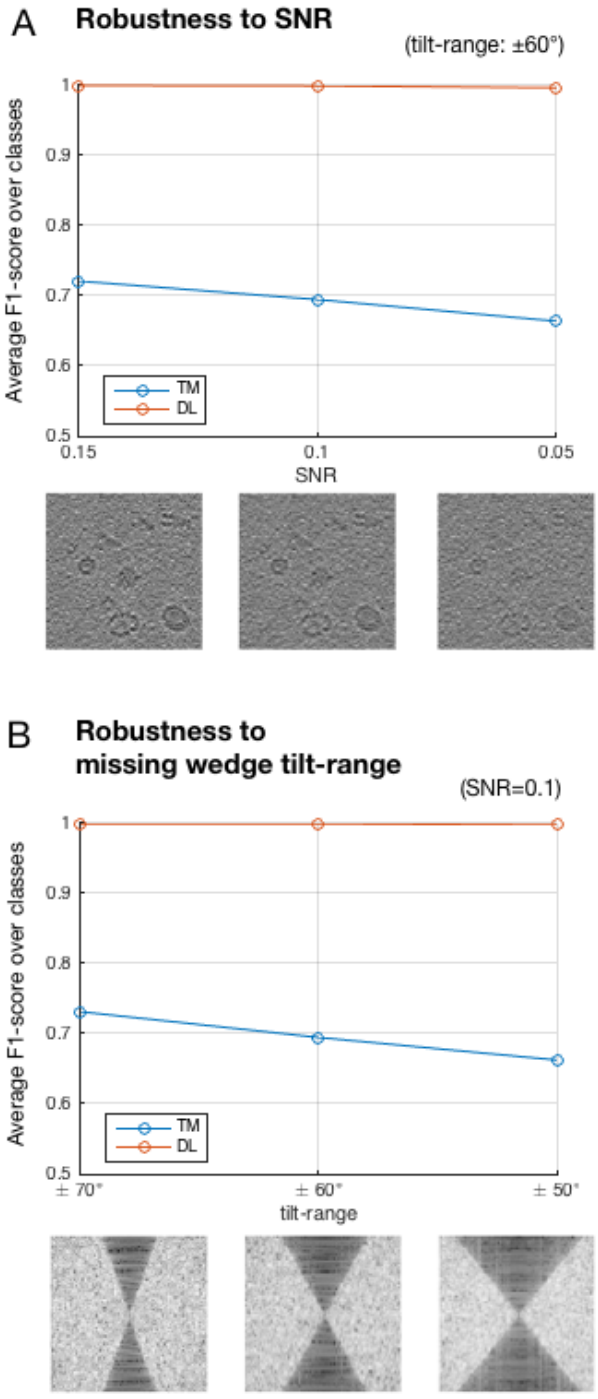

Figure 8: Dataset #1: average F1-score for each method, for varying SNR (A) and tilt-ranges (B). For (A), we consider a $\pm 60^\circ$ tilt-range and SNR values 0.15, 0.10 and 0.05. For (B), we consider a SNR of 0.10 and tilt-ranges $\pm 70^\circ$, $\pm 60^\circ$ and $\pm 50^\circ$. The images below the curves illustrate the effects of the varying parameters on a synthetic data sample (in image domain for (A), in spectral domain for (B)).

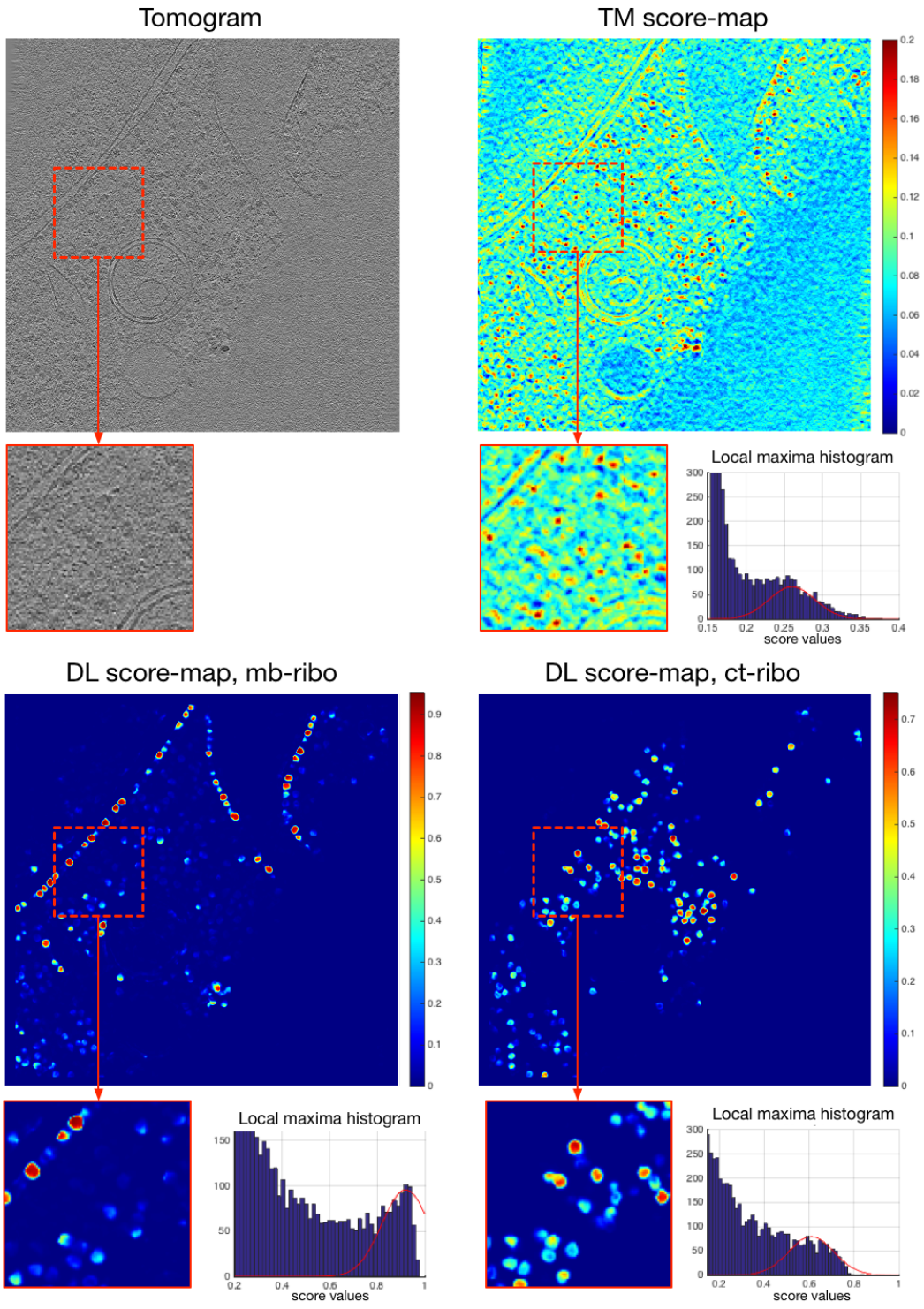

Figure 9: Dataset #2: comparing the score maps obtained from TM and our DL approach. On the score-maps bottom, zoomed-in windows and histograms of local maxima values.

the Euclidean distances between the ribosomes and the nearest membrane component (see Fig. 10 B). The distance histograms confirm the visual analysis: for *mb-ribos*, the distance histogram has a sharp mode at 136.8Å, which corresponds to the ribosome radius. Therefore a large majority of the voxels classified as *mb-ribos* are very close or linked to the membrane. For *ct-ribos*, we notice no sharp mode and the distance histogram looks Gaussian with a large variance. This confirms the heterogeneity of ribosomes freely floating in the cytoplasm. In conclusion, our method allows accurate multiclass object detection in cryo-ET.

Evaluation of overlap with annotations The next step of our evaluation process consists in measuring the overlap between the expert annotations and the *mb-ribos* found by DL as explained earlier. We compare the results to the TM outputs, that is before applying sophisticated post-classification methods. In Fig. 11, we plotted the Recall, Precision and F_1 score w.r.t. the DL and TM algorithm parameters. We focused on the thresholds used to detect objects (TM: threshold on score values; DL: object size threshold) (see Sec. 2.1.2). We obtained a F_1 score of 0.86 for DL and a F_1 score of 0.50 for TM. These numbers illustrates the ability of our DL approach to learn and bypass the expert processing chain. Moreover, the computation time of our DL approach is very small when compared to the TM algorithm as given in Figure 3. Now that it has been established that DL has a better overlap with the annotations than TM, in the remaining we focus our analysis DL detections.

We have also examined the complementarity between the two sets of *mb-ribo* macromolecules detected by the the experts (guided by TM) and the DL method. In what follows, we respectively denote S_E and S_{DL} the sets obtained by the experts and the DL method. While the overlap $S_E \cap S_{DL}$ between both sets is substantial (1516 particles), there is also a significant amount of particles belonging to $S_E \setminus S_{DL}$ (220 particles), i.e. the particles annotated by the expert but overseen by DL, and to $S_{DL} \setminus S_E$ (356 particles), i.e. particles found by DL but overseen by the expert. We can benefit from the two complementary object position estimations to improve overall validation rates. Actually, the union $S_E \cup S_{DL}$ of the two sets enables to increase the list of potential *mb-ribo* macromolecules, for which a confidence level can be assigned to each member depending on whether it belongs to $S_E \cap S_{DL}$, $S_{DL} \setminus S_E$ or $S_{DL} \setminus S_E$. Objects belonging to $S_E \cap S_{DL}$, i.e. found by both methods, are very likely to be true positives. Meanwhile the detected objects belonging to $S_E \setminus S_{DL}$ and $S_{DL} \setminus S_E$ 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-tomogram. From our analysis, it is possible to get a high overlapping rate with the expert annotations by using our DL approach, suggesting that DL is able to learn the expert analysis chain.

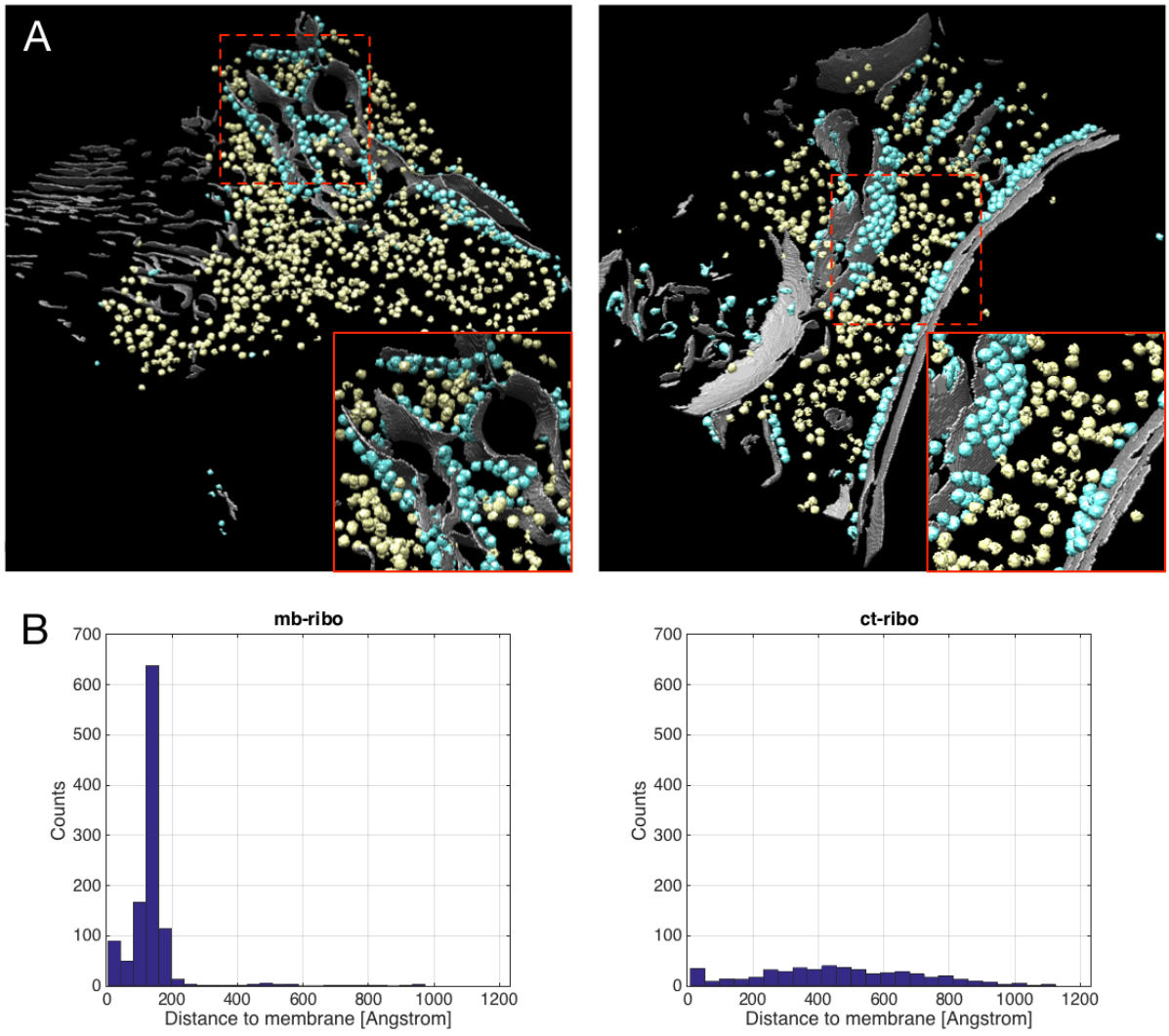

Figure 10: (A) 3D voxelwise classification of experimental tomograms, as obtained by our CNN. The classification displays cell *membrane* (in gray), membrane-bound ribosomes (blue), and cytoplasmic ribosomes (yellow). (B) Distance to membrane histograms of detected ribosomes, on the left for *mb-ribos* and on the right for *ct-ribos*.

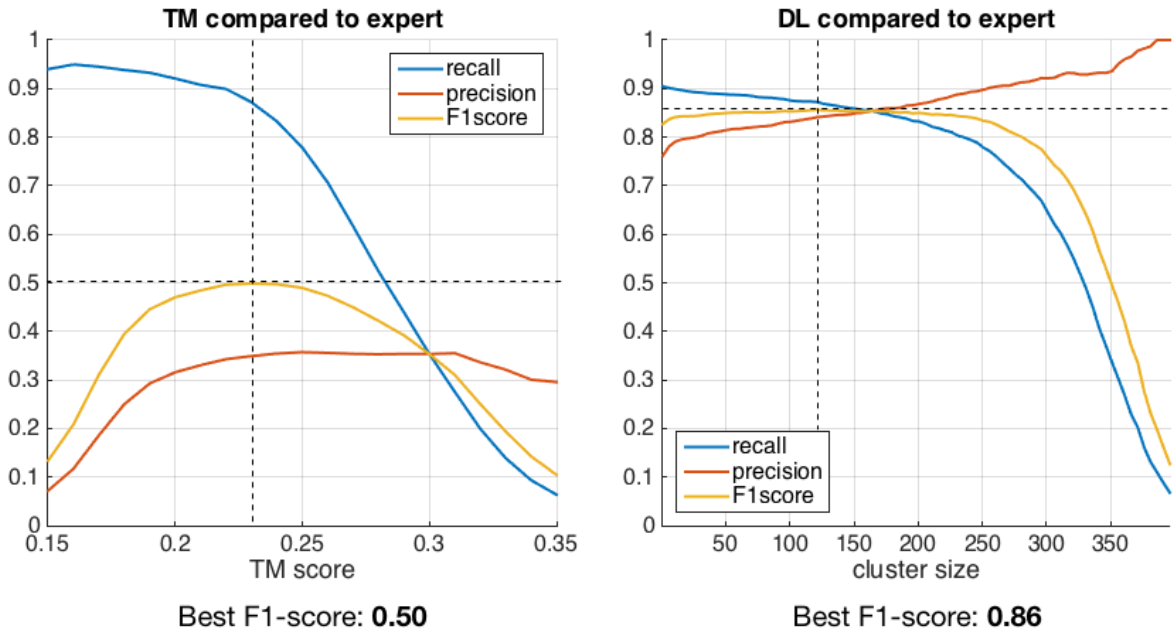

Figure 11: Overlap with expert annotation w.r.t. method threshold parameter: on the left for TM, on the right for our DL approach.

Analysis of subtomogram averaging results It is usually recommended in cryo-ET to analyse the 3D structure of macromolecules by using subtomogram averaging. Accordingly, we analysed the detected particles by computing subtomogram averages for each ribosome subfamily (see Fig. 12). In this way, we compare the subtomogram averages obtained from the DL detections and the expert annotations. The same process has been used to compute all averages, namely fast rotational matching [Chen et al., 2013] (see Sec. 3.4). The averages of *mb-ribo* are composed of two densities: the average ribosomal density and the average membrane density. The intensity level of the membrane density varies with the proportion of *mb-* and *ct-ribos* in the average; the higher the intensity is, the more the number of *mb-ribos* is high.

As expected, the average membrane density computed from the set S_{DL} of *mb-ribo* particles has a high intensity, while the average membrane density computed from the set of *ct-ribo* particles detected by our DL approach, is non-existent. Consequently, our method makes little to no confusion between *mb-* and *ct-ribos*. We notice that the average ribosomal density is similar in both ribosome subfamilies, the difference lying in the neighborhood of the link with the membrane. Therefore, DL is able to efficiently represent the geometric structure of the macromolecule, and at the same time to capture the local context and interactions of the macromolecule with the environment.

Next, we computed the Fourier shell correlation (FSC) for obtained subtomogram averages (see Fig. 12). According to Fig. 12, the best resolution has been obtained with the expert average: 24Å versus 24.7Å and 32.7Å for the DL subclasses *mb-ribo* and *ct-ribo* respectively. Our method therefore allows to achieve a resolution comparable to an expert. As for the *ct-ribo* average the resolution is lower, most likely because the *ct-ribo* annotations used for learning are of lower quality (see Sec. 3.1.2). Even though resolutions for the expert and DL *mb-ribo* averages are very close, the poorer resolution values of the DL averages can be caused by two main factors:

H_1 : The averages contain potential false positives, suggesting that our method probably made a few mistakes.

H_2 : The averages contain particles with a low quality, suggesting that our method found supplementary particles, missed or discarded during the annotation process.

In what follows, we performed further investigations to check the two hypotheses H_1 and H_2.

First, we decided to align and average the *mb-ribo* particles of the set $S_{DL} \setminus S_E$, i.e. found by DL but absent in the expert annotations (see Fig. 13). If no clear signature of ribosome density appears in the average, hypothesis H_1 is valid. On the contrary, if we observe a ribosome patterns in the average, we can conclude that our DL method found additional ribosome particles potentially discarded by the experts (hypothesis H_2). Nevertheless, note that the two hypotheses are not mutually exclusive.

In Fig. 13, we display the resulting subtomogram average denoted $\mathbf{A}_{DL}$ computed from 356 detected particles. Since it is not guaranteed that all the particles involved in the average are actual *mb-ribos*, we evaluated the difference between $\mathbf{A}_{DL}$ and the average $\mathbf{A}_{ref}$ computed from 356 true *mb-ribos* (expert annotations) randomly picked from the the set $S_E \cap S_{DL}$. Also, in order to check if $\mathbf{A}_{DL}$ is not biased by the reference template used for subtomogram alignment (see Sec. 3.1.2) as described in [Henderson, 2013], we computed another average denoted $\mathbf{A}^{\circ}_{DL}$ from 356 subtomograms picked from random positions. We compared $\mathbf{A}_{DL}$, $\mathbf{A}^{ref}_{DL}$, and $\mathbf{A}^{\circ}_{DL}$ visually and by estimating the underlying resolution (see Fig. 13). We notice that $\mathbf{A}_{DL}$ has a lower resolution (36.4Å) than $\mathbf{A}^{ref}_{DL}$ (33.5Å). It means that $\mathbf{A}_{DL}$ probably contains false positives and/or very noisy instances. Nonetheless, $\mathbf{A}_{DL}$ has a higher resolution than $\mathbf{A}^{\circ}_{DL}$ (48.6Å), suggesting that the reference template bias is not significant. Moreover, a ribosome pattern visually appears in $\mathbf{A}_{DL}$. This suggests that our DL approach has actually found *mb-ribos* that have been missed during the annotation process.

To be fair, we applied a similar comparison to $S_E \setminus S_{DL}$, i.e. the set of *mb-ribos* annotated by the expert but missed by DL. As before, we obtain $\mathbf{A}_E$ from the 220 objects belonging to $S_E \setminus S_{DL}$, $\mathbf{A}^{ref}_{DL}$ by randomly sampling 220 *mb-ribos* from $S_E \cap S_{DL}$, and $\mathbf{A}^{\circ}_{DL}$ from 220 random positions. The obtained resolutions are very similar to what is achieved with DL. Here again, $\mathbf{A}_E$ has a lower resolution (37.6Å) than the reference $\mathbf{A}^{ref}_E$ (34.2Å). It is noteworthy that in both cases, the *mb-ribos*

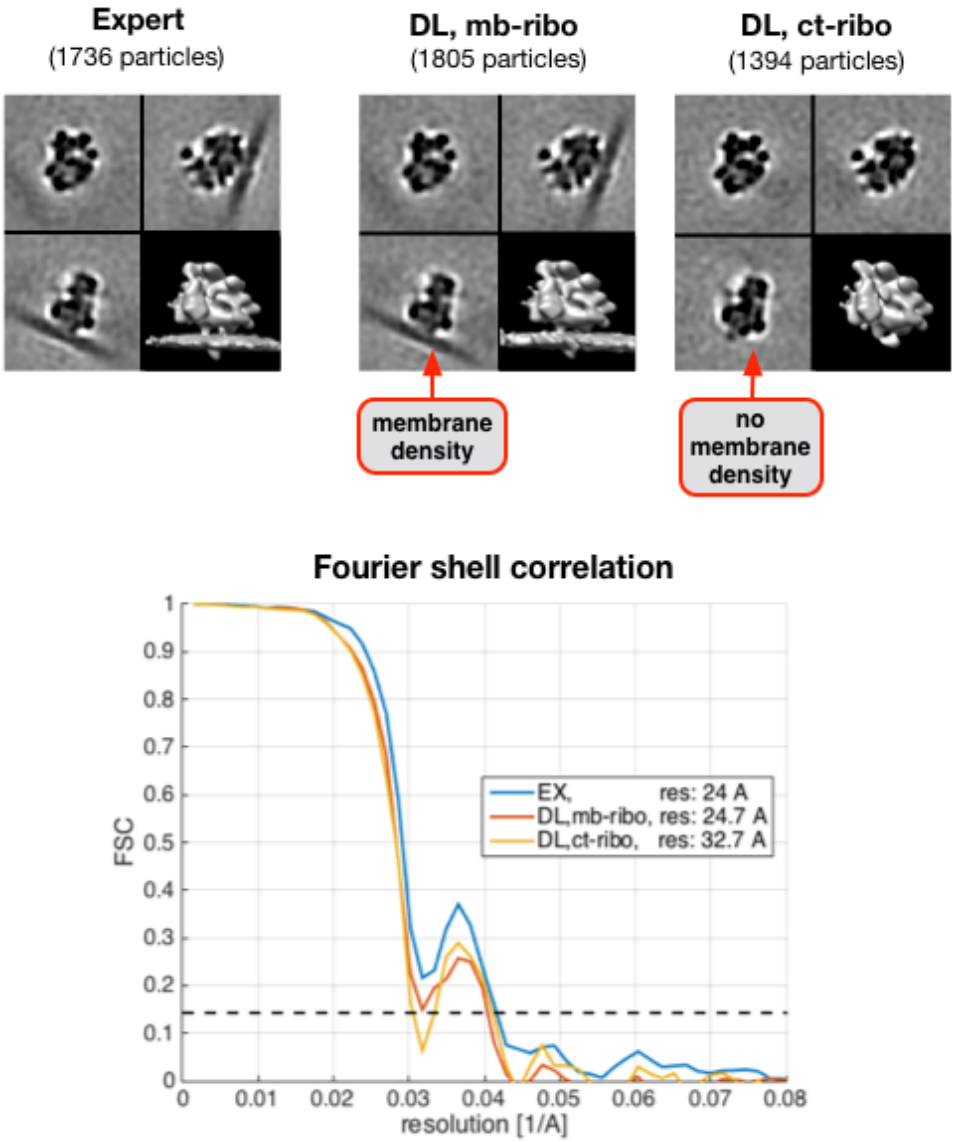

Figure 12: On top, the subtomograms obtained from, the expert annotations, the DL detections for *mb-ribos* and *ct-ribos*, respectively. All averages have been obtained with the same alignment procedure and parameters. For visualization purpose, the averages have been low-pass filtered at 40A resolution. At the bottom, the corresponding gold-standard FSC curves with estimated resolutions.

from $S_E \cap S_{DL}$ lead to a better resolution than the complement sets $S_E \setminus S_{DL}$ and $S_{DL} \setminus S_E$. This observation illustrates well what has been discussed earlier, namely that combining sets obtained from different methods allows to attribute confidence levels. The lower resolution of $\mathbf{A}_{DL}$ and $\mathbf{A}_E$ suggest that the sets $S_E \setminus S_{DL}$ and $S_{DL} \setminus S_E$ contain more heterogeneity, and thus potentially include rare conformations.

This set of results emphasizes that TM and DL can be combined to better investigate cryo-tomograms. The set of common objects found by the two methods enables to focus on detections exclusively found by the experts or by DL. In addition, we show that, while the FSC curves obtained with DL are below the curve obtained with expert annotations (see Fig. 12), it is risky to decide that the detected macromolecules are not ribosomes (see Fig. 13). The estimated resolutions are lower mainly because the particles involved in the subtomogram averaging are corrupted by noise and other sources of signal degradation. The supplementary noisy particles need to be further examined by experts since they may be valuable *mb-ribo* candidates. Finally, 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. In summary, DL found additional noisy *mb-ribos* that were missed or discarded during the annotation process.

3.4 Implementation details of the DL (3D CNN) software

To implement our 3D CCN method, we used Keras [keras.io], an open-source toolbox written in python and using the Tensorflow framework.

As to template matching and subtomogram averaging, we used the PyTom toolbox [Hrabe et al., 2012]. We used the in Pytom implemented fast rotational matching routine [Chen et al., 2013] for subtomogram alignment. The alignment has been performed with respect to a reference template (see Section 3.1.2).

For 3D visualizations, we used Chimera [Pettersen et al., 2004].

4 Discussion

We proposed a 3D CNN framework for voxelwise multiclass classification and particle localization in 3D cryo-ET. We showed on synthetic data that our DL method has superior localization performance than TM. On real data, our DL method is able to discriminate two binding states of ribosomes and to segment cell membranes. We achieved 86% of overlap with expert annotations. Also, when applying subtomogram averaging to the detections, we obtain a resolution comparable to the expert. The supplementary membrane ribosomes found by DL and missed during the annotation process, can be further inspected since the related set is small. While it is established that TM is widely applied in

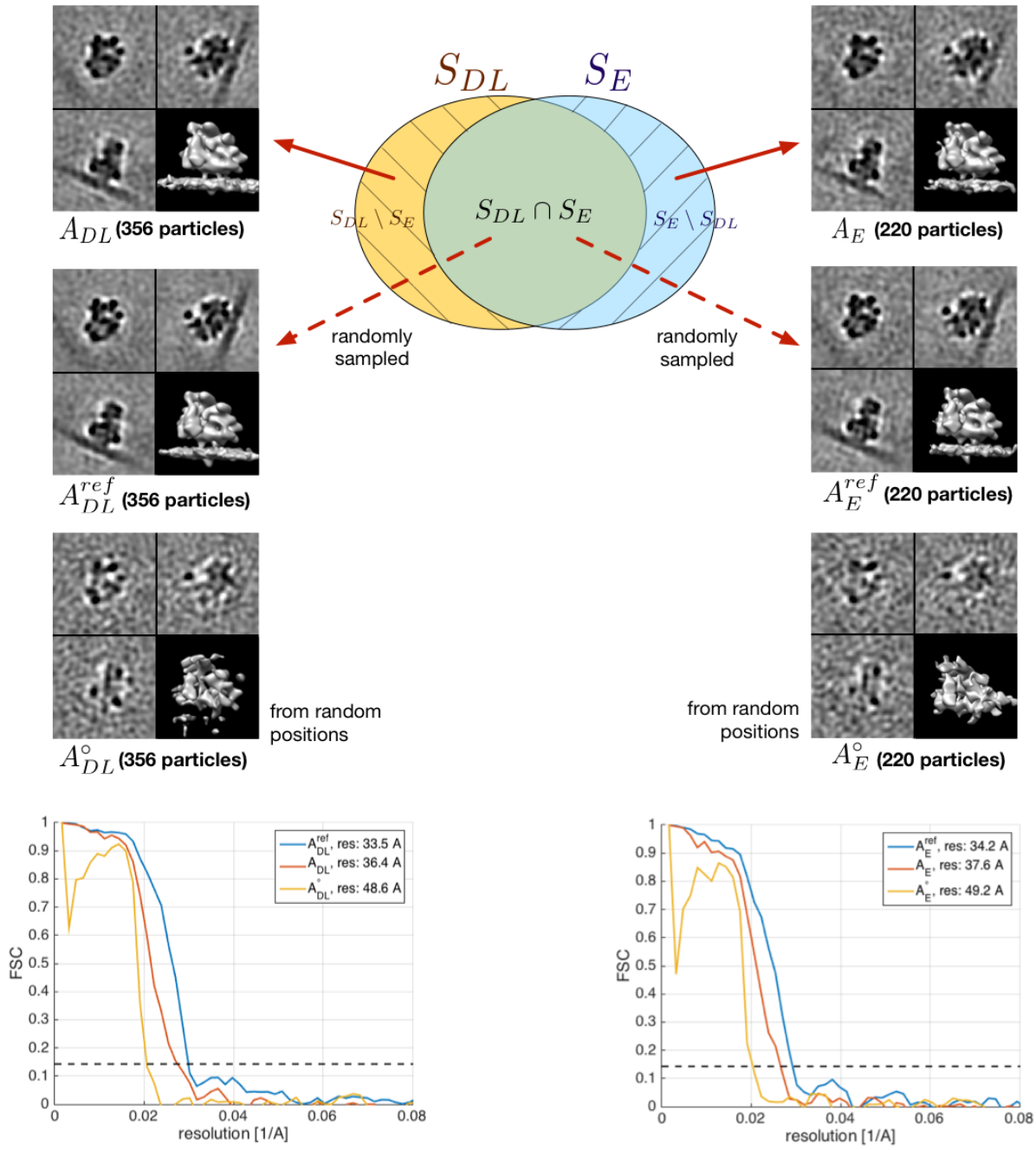

Figure 13: Top-middle: diagram representing the overlap between *mb-ribo* sets S_{DL} and S_E. The emanating arrows represent from which sub-set the displayed subtomogram averages originate. Bottom: FSC curves for each subtomogram average.

cryo-ET to successfully detect various macromolecular complexes, as of now it has mainly been used to detect relatively large macromolecules like ribosomes. It is worth noting that DL found ribosomes that where missed by the expert annotation process (assisted by TM), and conversely. Consequently both methods can be combined to investigate more efficiently cryo-tomograms. In particular, more effort can be made to analyze the particles detected by either DL or TM, that is when the algorithms are not in agreement. TM and DL can also be combined in a virtuous manner as follows: experts can perform a first round of detection with the routine tools, then train our method from the resulting detections; in a second round, our DL framework can be applied to check the missed objects and increase the size of databases.

To our knowledge, TM is not usually applied to detect more than two or three classes/subclasses on the same dataset. However, cryo-electron tomograms contain much more information. They offer 3D views of the whole macromolecular environment, albeit hardly discernable from noise and artifacts. Therefore more powerful pattern recognition and machine learning techniques are needed to analyse contents and extract information. DL has been shown to be able to handle the spectacular amount of 1000 classes [Krizhevsky et al., 2012]. Also, it is invariant to diffeomorphisms [Mallat, 2016] and then able to cope with non-rigid, elastic deformations within a class. A major disadvantage of TM is that it can only handle rigid views of an object, which is problematic knowing that many proteins (e.g. proteasomes) have a high structural variability. Thus, provided that DL has been trained with enough representative examples capturing shape variability, the CNN should be able to learn different conformations of the same macromolecule.

Acknowledgment

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").

Note

Through the LifeExplorer project, the Fourmentin-Guilbert 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 cryoelectron microscopy applied on frozen but intact cells. For

years, the Fourmentin-Guilbert Foundation has supported the Max Planck Institute of Biochemistry, headed by Wolfgang Baumeister whose team has been capable of delivering whole cryotomograms of E. coli cells at an unprecedented resolution.

The next big step, 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 "big" complexes like the ribosomes. Facing this challenge, the Fourmentin-Guilbert Foundation has solicited the research group headed by Charles 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 Proejct-Team has then developped 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

[Best et al., 2007] Best, C., Nickell, S., and Baumeister, W. (2007). Localization of protein complexes by pattern recognition. *Methods Cell Biol.*, 2007(79):615–638.

[Chen et al., 2017] Chen, M., Dai, W., Sun, S. Y., Jonasch, D., He, C. Y., Schmid, M. F., Chiu, W., and Ludtke, S. J. (2017). Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. *Nat. Methods*.

[Chen et al., 2013] Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J. M., and Förster, F. (2013). Fast and accurate reference-free alignment of subtomograms. *J. Struct. Biol.*, 182(3):235–45.

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

[Förster and Hegerl, 2007] Förster, F. and Hegerl, R. (2007). Structure determination In Situ by averaging of tomograms. In *Cell. Electron Microsc.*, volume 79, pages 741–767.

[Förster et al., 2008] Förster, F., Pruggnaller, S., Seybert, A., and Frangakis, A. S. (2008). Classification of cryo-electron sub-tomograms using constrained correlation. *J. Struct. Biol.*, 161(3):276–286.

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

[Hinton et al., 2012] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process. Mag.*, 29(6):82–97.

[Hrabe et al., 2012] Hrabe, T., Chen, Y., Pfeffer, S., Kuhn Cuellar, L., Mangold, A.-V., and Förster, F. (2012). PyTom: A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *J. Struct. Biol.*, 178(2):177–188.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. L. (2014). Adam: a method for stochastic optimization. *arXiv Prepr.*

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Conf. Neural Inf. Process. Syst.*, pages 1–9.

[Lecun et al., 2015] Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

[Lecun et al., 2010] Lecun, Y., Kavukcuoglu, K., and Farabet, C. (2010). Convolutional networks and applications in vision. In *IEEE Int. Symp. Circuits Syst.*, pages 253–256.

[Long et al., 2014] Long, J., Shelhamer, E., and Darrell, T. (2014). Fully convolutional networks for semantic segmentation. In *Conf. Comput. Vis. Pattern Recognit.*, pages 3431–3440.

[Mallat, 2016] Mallat, S. (2016). Understanding deep convolutional networks. *Philos. Trans. R. Soc. London A Math. Phys. Eng. Sci.*, 374(2065).

[Martinez-Sanchez et al., 2014] Martinez-Sanchez, A., Garcia, I., Asano, S., Lucic, V., and Fernandez, J.-j. (2014). Robust membrane detection based on tensor voting for electron tomography. *J. Struct. Biol.*, 186(1):49–61.

[Milletari et al., 2016] Milletari, F., Navab, N., and Ahmadi, S.-a. (2016). V-Net : fully convolutional neural networks for volumetric medical image segmentation. *arXiv Prepr.*, pages 1–11.

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

[Pettersen et al., 2004] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25(13):1605–1612.

[Pfeffer et al., 2017] Pfeffer, S., Dudek, J., Schaffer, M., Ng, B. G., Albert, S., Plitzko, J. M., Baumeister, W., Zimmermann, R., Freeze, H. H., Engel, B. D., and Förster, F. (2017). Dissecting the molecular organization of the translocon-associated protein complex. *Nat. Commun.*, 8:14516.

[Rolnick et al., 2017] Rolnick, D., Veit, A., Belongie, S., and Shavit, N. (2017). Deep learning is robust to massive label noise. *arXiv Prepr.*

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, volume 9351, pages 234–241.

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

[Szegedy et al., 2013] Szegedy, C., Toshev, A., and Erhan, D. (2013). Deep neural networks for object detection. In *Conf. Neural Inf. Process. Syst.*, pages 1–9.

[Tang et al., 2007] Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., and Ludtke, S. J. (2007). EMAN2 : An extensible image processing suite for electron microscopy. *J. Struct. Biol.*, 157:38–46.

[Wang et al., 2016] Wang, F., Gong, H., Liu, G., Li, M., Yan, C., Xia, T., Li, X., and Zeng, J. (2016). DeepPicker : A deep learning approach for fully automated particle picking in cryo-EM. *J. Struct. Biol.*, 195(3):325–336.