nature portfolio

https://doi.org/10.1038/s41592-021-01275-4

Supplementary information

Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms

In the format provided by the authors and unedited

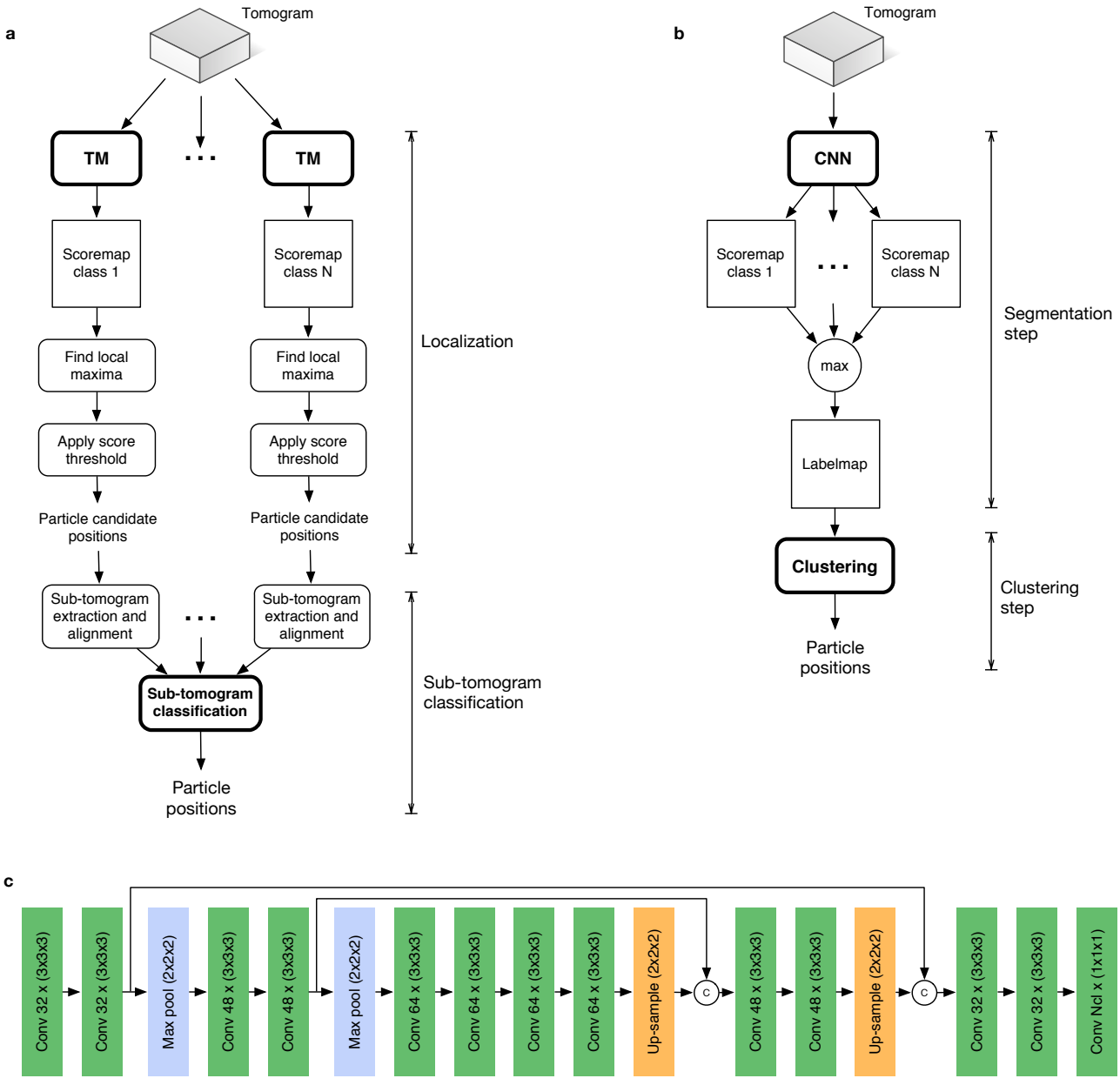
Supplementary information

Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms

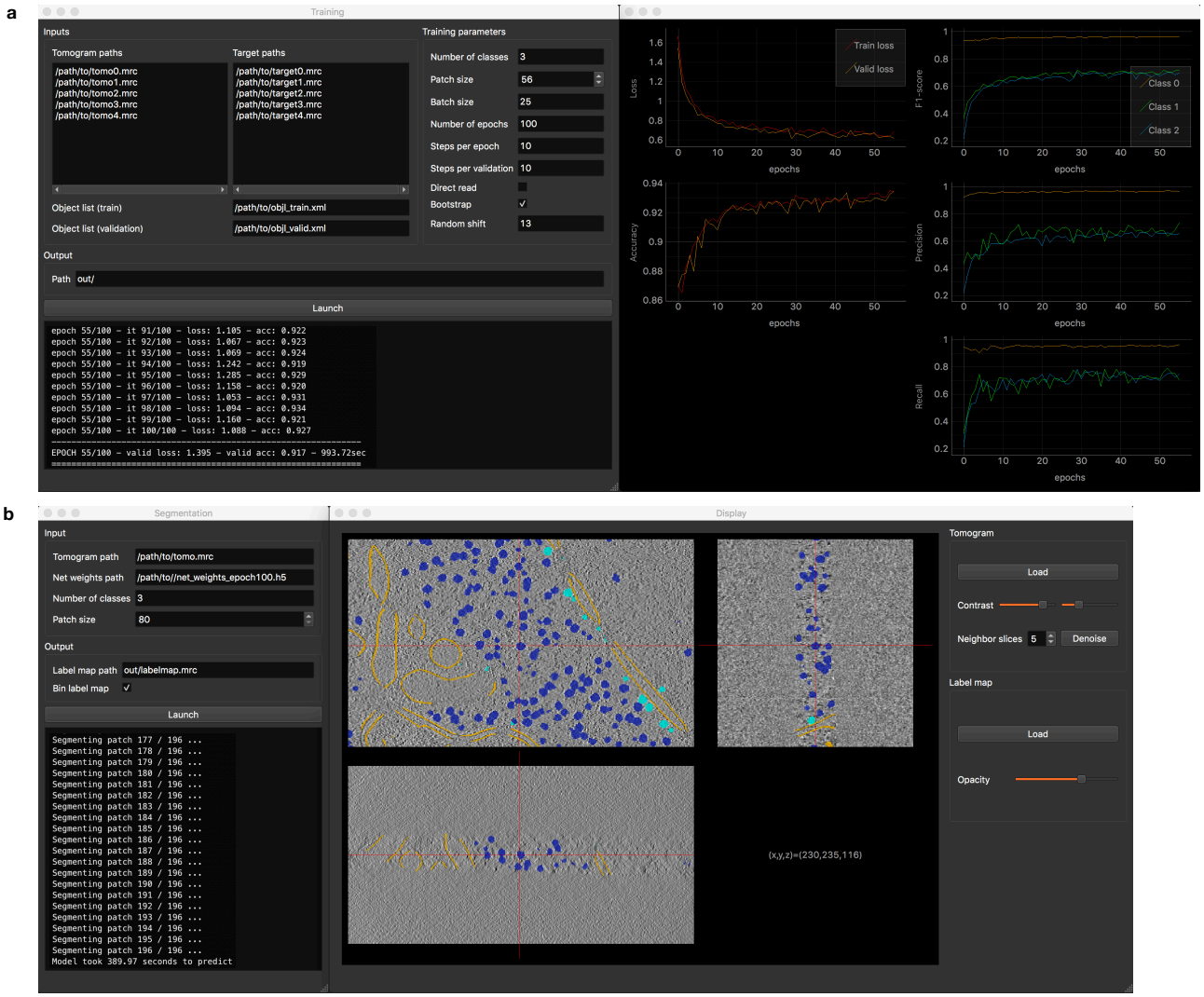
E. Moebel, A. Martinez-Sanchez, L. Lamm, R.D. Righetto, W. Wietrzynski, S. Albert, D. Larivière, E. Fourmentin, S. Pfeffer, J. Ortiz, W. Baumeister, T. Peng, B.D. Engel, C. Kervrann

Membranes		M1	M2	M3	M4	M5	M6	M7	M8	M9	**Global**
Mono-class PSII	F_1-score	0.557	0.465	0.533	0.625	0.789	0.516	0.571	0.588	0.286	**0.566**
	Precision	0.710	0.476	0.800	0.625	0.789	0.727	0.556	0.500	0.400	**0.643**
	Recall	0.458	0.455	0.400	0.625	0.789	0.400	0.588	0.714	0.222	**0.505**
Multi-class PSII	F_1-score	0.632	0.582	0.714	0.737	0.696	0.250	0.562	0.182	0.571	**0.619**
	Precision	0.638	0.485	0.625	0.636	0.593	0.750	0.600	0.250	0.800	**0.601**
	Recall	0.625	0.727	0.833	0.875	0.842	0.150	0.529	0.143	0.444	**0.638**
Template Matching PSII	F_1-score	0.400	0.200	0.355	0.258	0.350	0.279	0.318	0.000	0.286	**0.313**
	Precision	0.556	0.375	0.344	0.571	0.333	0.261	0.259	0.000	0.400	**0.353**
	Recall	0.312	0.136	0.367	0.167	0.368	0.300	0.412	0.000	0.471	**0.281**

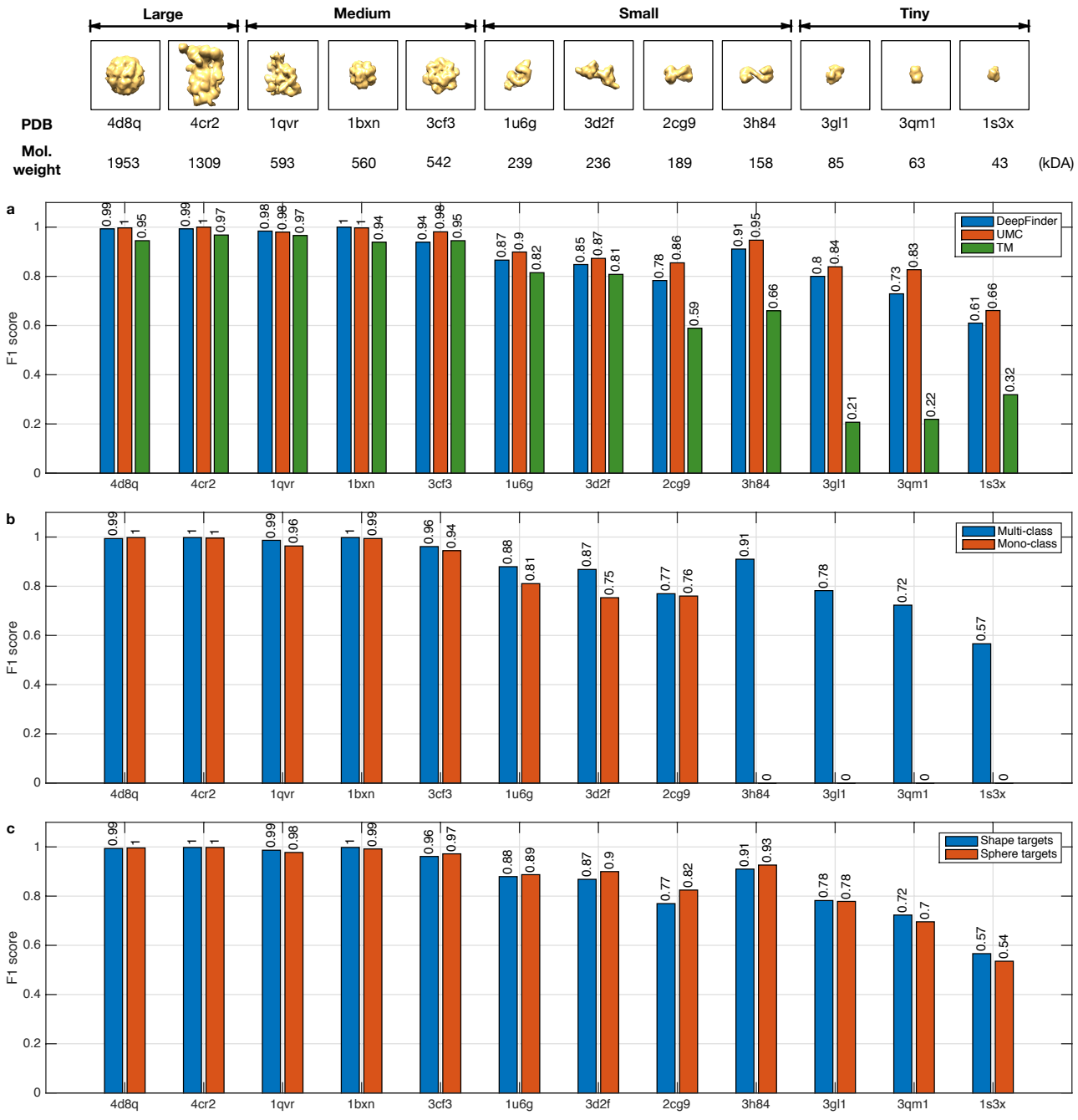
Supplementary Table 2: **Comparison of F_1-scores for the detection of PSII complexes embedded within native thylakoid membranes (Dataset #4)**. For the test tomogram, we ran the DeepFinder mono-class (470 particles) and multi-class (508 particles) strategies, as well as PyTOM template matching (508 particles). For an even comparison, we thresholded the template matching hits to match the number of picks from the DeepFinder multi-class approach.The scores were measured after masking the picks to different membranes (M1, M2...) of the test tomogram. These membranes vary in resolution and in the number of PSII complexes they host.



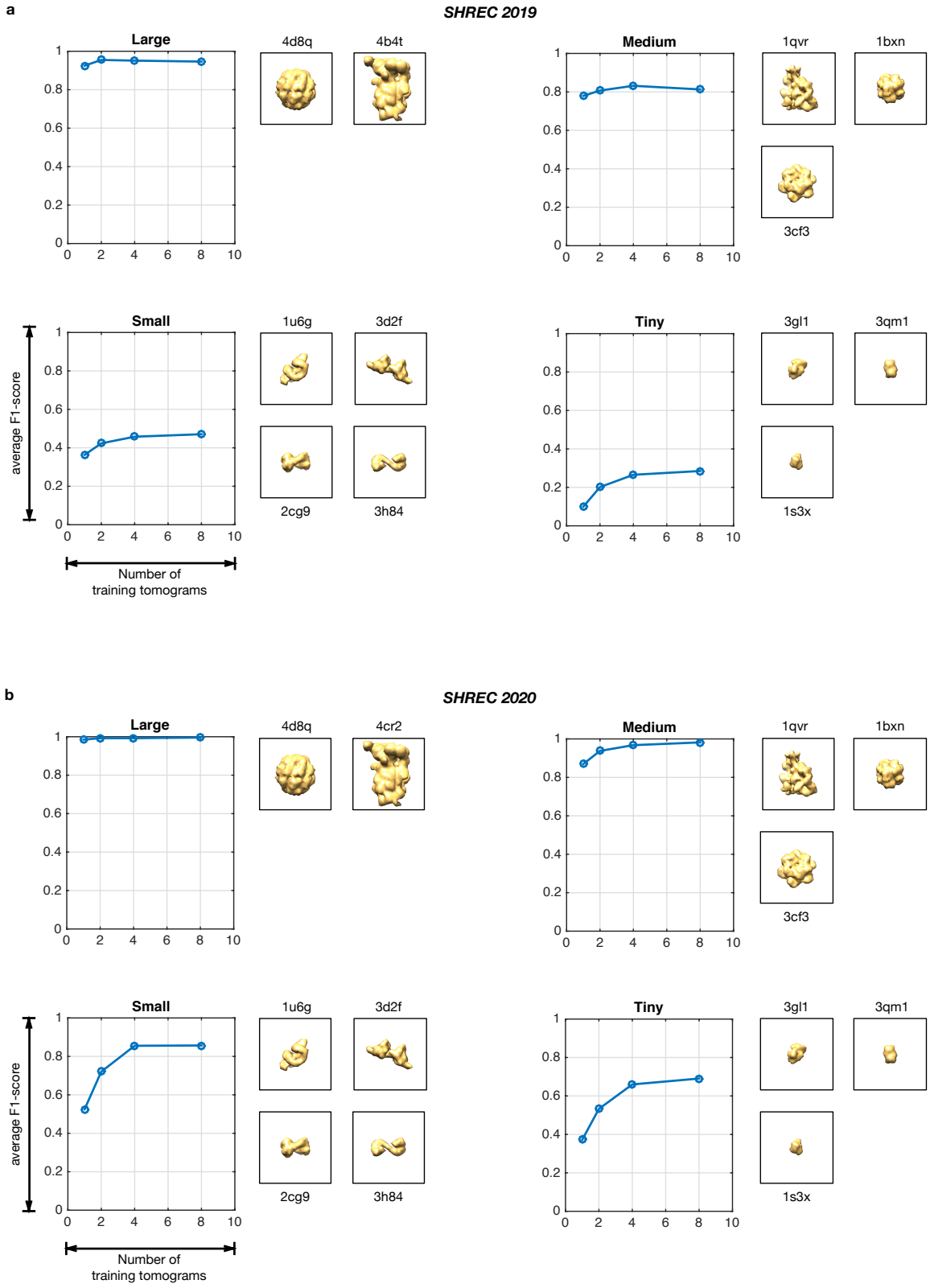
Extended Data Fig. 1: Two workflows for macromolecule localization in cryo-ET. a, Conventional processing pipeline based on template matching. **b,** DeepFinder (analysis stage): a multi-class approach able to localize particles of several different macromolecular species in one pass. **a,** and **b,** highlight why DeepFinder is more agile than Template Matching when several macromolecule classes need to be localized. **c,** CNN architecture used in DeepFinder and based on U-Net[7]. The architecture adopts the encoder-decoder paradigm, which produces an output volume with the same size as the input volume. Each green box represents a convolutional layer. The number of filters n and the filter size s is labeled as $n \times (s \times s \times s)$. All convolutional layers are followed by a ReLU activation function, except the last layer, which uses a soft-max function. The up-sampling is achieved with up-convolutions (also called "backward-convolution"). Combining feature maps from different scales is performed by concatenation along the channel dimension. In the end, the total number of architecture parameters is approximately 903k. More precisely, this number depends slightly on N_{cl}, the number of classes: $902,928 + N_{cl} \times 33$.



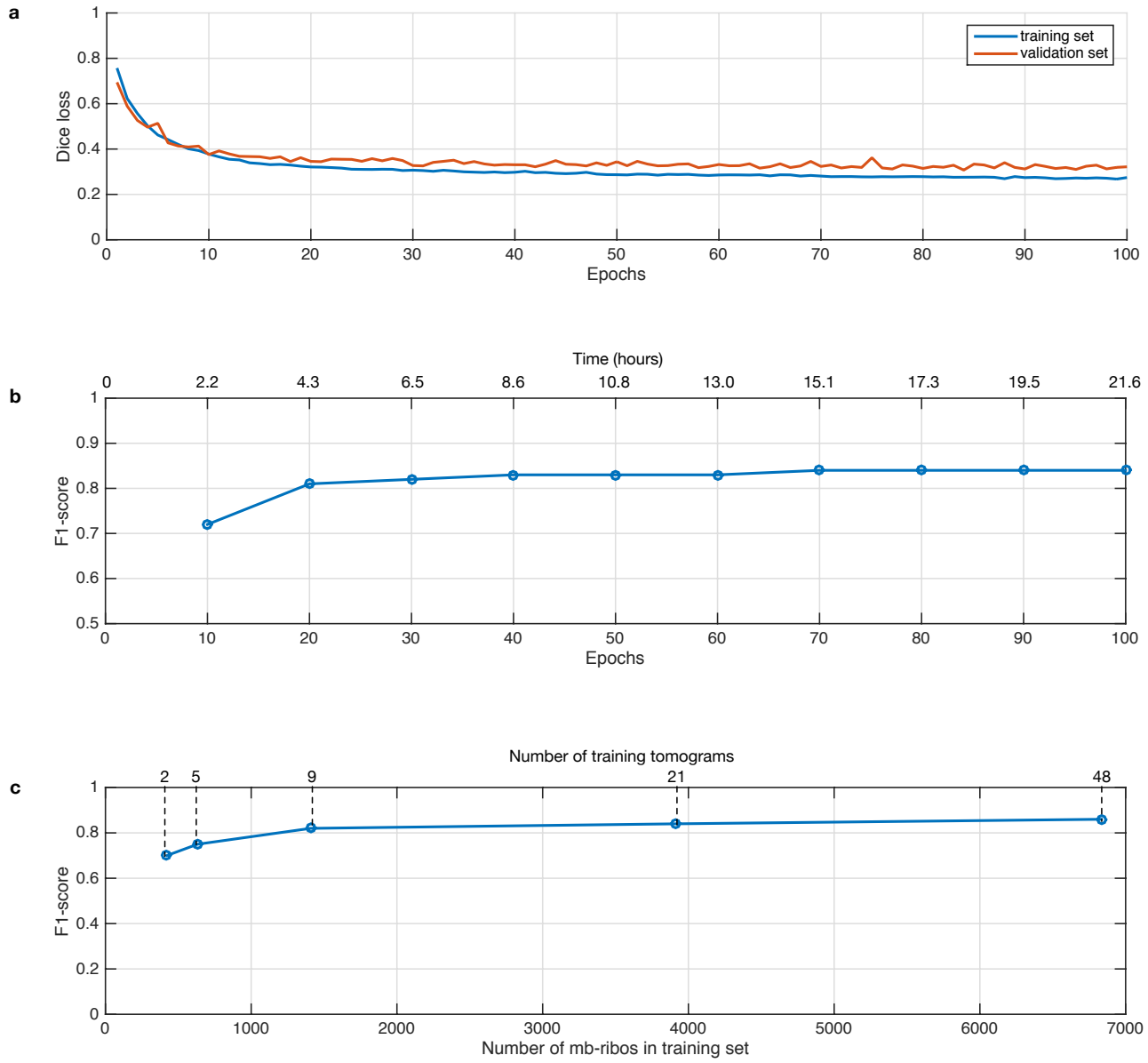
Extended Data Fig. 2: DeepFinder graphical user interface. a, Training interface composed of a first window for parametrizing the procedure and a second window for displaying the training metrics in real-time. **b,** Segmentation interface which also opens a data visualization tool. This tool allows the user to explore the tomogram with superimposed segmentations. 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).



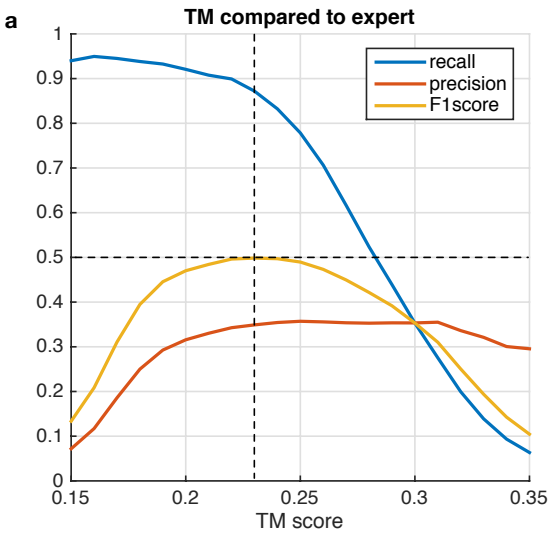
Extended Data Fig. 3: Analysis of algorithm performance on the synthetic dataset (SHREC'20 challenge). a, Performance (F_1-score) of DeepFinder, UMC and template matching algorithms and ability of algorithms to discriminate between 12 classes/subclasses of macromolecules. The highest (best) possible value of an F_1-score is 1.0 and the lowest (worst) possible value is 0. The scores of template matching were provided by the SHREC'20 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. This analysis used 8 tomograms for training, 1 tomogram for validation, and 1 tomogram for testing.

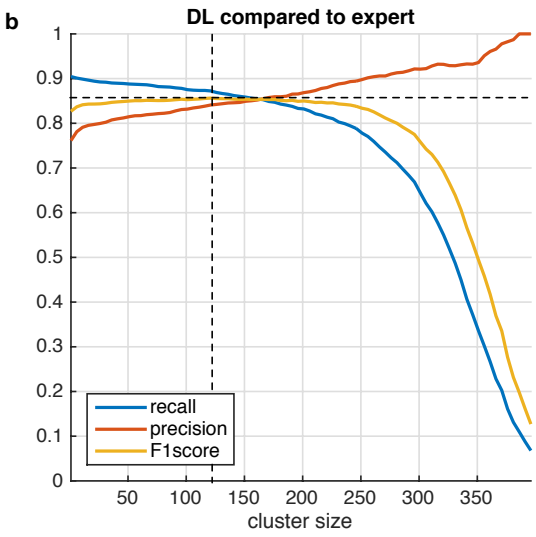


Extended Data Fig. 4: Evolution of F_1-scores with respect to sizes of the training sets (number of tomograms) on the synthetic SHREC dataset (12 classes). Scores are displayed for both the SHREC 2019 **a,** and 2020 **b,** editions. This figure gives an estimation of the amount of annotated data needed to identify macromolecules. This amount depends on the size of the target macromolecule: smaller targets require more annotations. Each tomogram contains in average 208 macromolecules per class. The macromolecules have been categorized into 4 groups (large, medium, small and tiny). This analysis used 8 tomograms for training, 1 tomogram for validation, and 1 tomogram for testing.

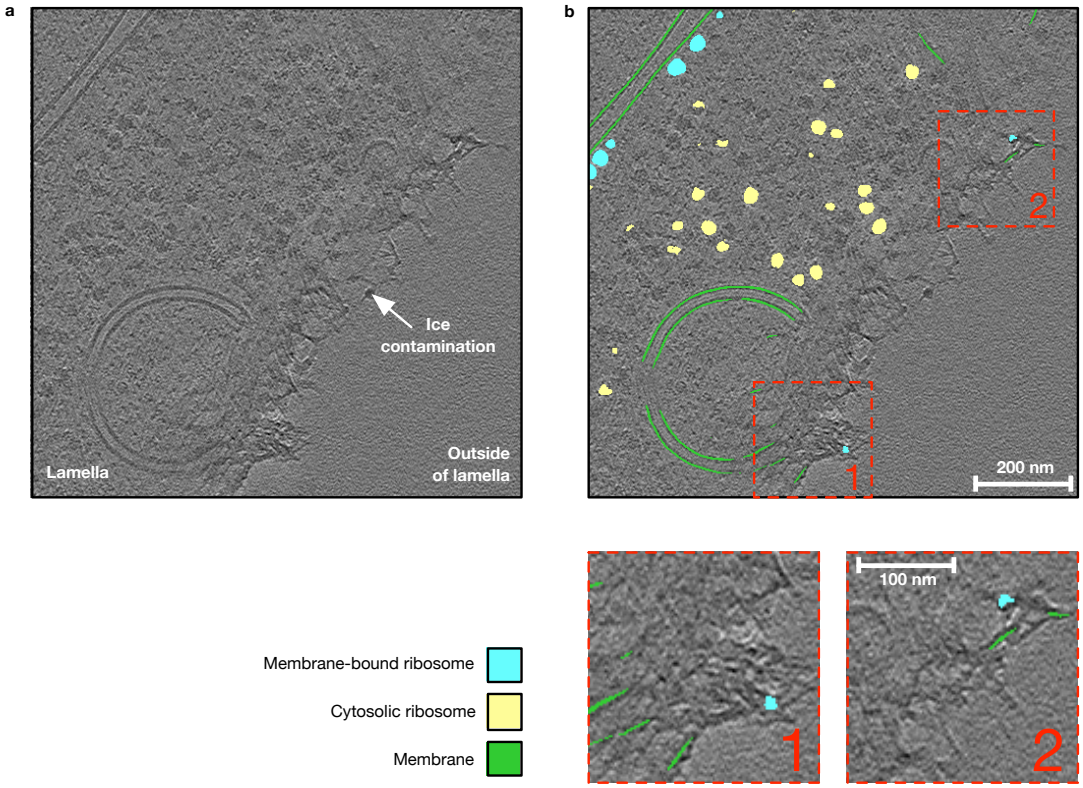


Extended Data Fig. 5: Evolution of F_1-score with respect to training iterations and training set size on real cryo-ET Dataset #2, *Chlamydomonas reinhardtii* (3 classes). a, The loss, which quantifies the segmentation quality, is computed for the training set, as well as for the validation set. Comparing both curves allows assessment of the generalization capabilities of DeepFinder. The curves for both sets should ideally overlap, otherwise it indicates overfitting (the network memorizes trained samples instead of learning discriminating features). One epoch equals 100 training iterations. **b,** The F_1-score, which quantifies the localization performance, computed on the test set. The F_1-score is obtained by comparing the membrane-bound ribosomes found by DeepFinder to expert annotations. The time axis has been obtained using a Tesla K80 GPU. The curve indicates that competitive particle picking results are obtained after 20 epochs, or 4.3 hours with the required GPU. This analysis used 21 tomograms for training, 1 tomogram for validation, and 8 tomograms for testing.**c,** In a similar fashion to Fig. 4, this curve provides an estimate of the quantity of training data required to achieve a competitive result. It appears that this quantity is 1400 ribosomes (9 tomograms), which is a typical size for a cryo-ET dataset. On first glance, this estimate seems to contradict the estimates in Fig. 4: the numbers do not coincide (the curve labeled "Large" estimates that quantity at 208 particles). Note that SHREC'19 is a synthetic dataset, composed of 12 classes. Here, we are dealing with a real cellular dataset consisting of 3 classes (membrane, membrane-bound ribosome and cytosolic ribosome). It appears that having a larger number of classes enables the use of smaller training sets. On the other hand, the case of real data is more difficult, notably because of the presence of "label noise" (errors due to the annotation pipeline) and other sources of signal corruption such as the missing wedge, the CTF and the low SNR (in part caused by increased molecular crowding inside cells). This analysis 1 tomogram for validation, and 8 tomograms for testing.

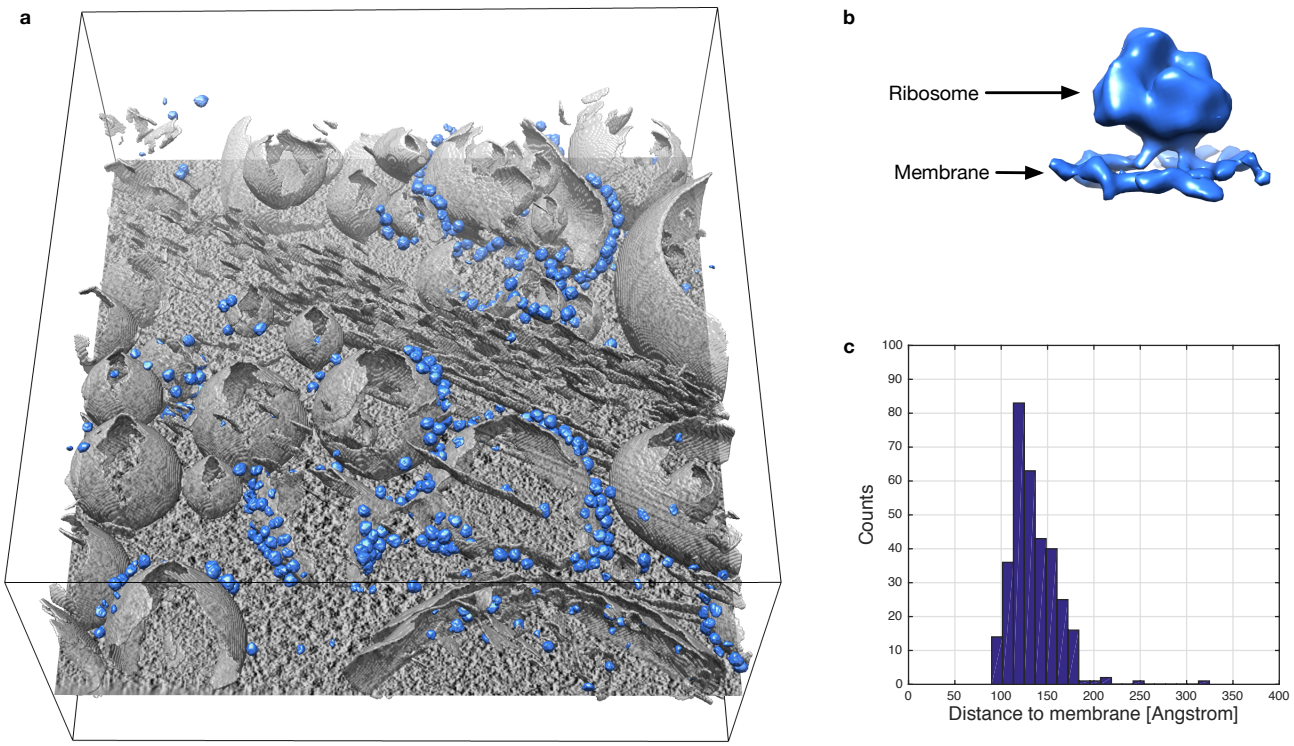




Extended Data Fig. 6: Quantitative analysis of overlap with expert annotations on cellular cryo-ET data (Dataset #2, *mb-ribos*). We varied the thresholds of template matching (**a**) and DeepFinder (**b**) to compute the *Recall* (ratio between the number of true positives (TP) and the number of particles in the ground truth), *Precision* (ratio between the number of TP and the number of detected particles) and F_1-score (2 $\times$ (Recall $\times$ Precision) / (Recall + Precision)) curves. The threshold parameter for template matching is the constrained correlation coefficient, and for DeepFinder it is the cluster size, which corresponds to the macromolecule volume (in voxels). We obtained a maximum F_1-score of 0.86 for DeepFinder and a maximum F_1-score of 0.50 for template matching (with no post-classification step, see Extended Data Fig. 1a). Template matching and DeepFinder both 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 from the synthetic dataset (SHREC'19 challenge). This analysis used 48 tomograms for training, 1 tomogram for validation, and 8 tomograms for testing.



Extended Data Fig. 7: **DeepFinder handles ice contamination on the lamella surface. a,** Tomogram slice depicting the border of a FIB-milled lamella. The lamella contains a *Chlamydomonas reinhardtii* cell, with a lamella surface suffering from ice contamination. **b,** Tomogram slice with superimposed DeepFinder segmentation. Most of the ice contamination artifacts have been correctly classified as "background". Nonetheless, some missclassifications exist, as can be observed in the zoomed-in boxes (in dashed red). In boxes 1 and 2, DeepFinder confuses some artifacts with membranes, and some features are wrongly classified as membrane-bound ribosomes. Such missclassifications can be filtered out, either by masking the boundaries of the lamella, or by rejecting segmented objects that are too small (using the "cluster size" attribute given by the clustering step of the DeepFinder analysis stage). This analysis used 48 tomograms for training, 1 tomogram for validation, and 8 tomograms for testing.

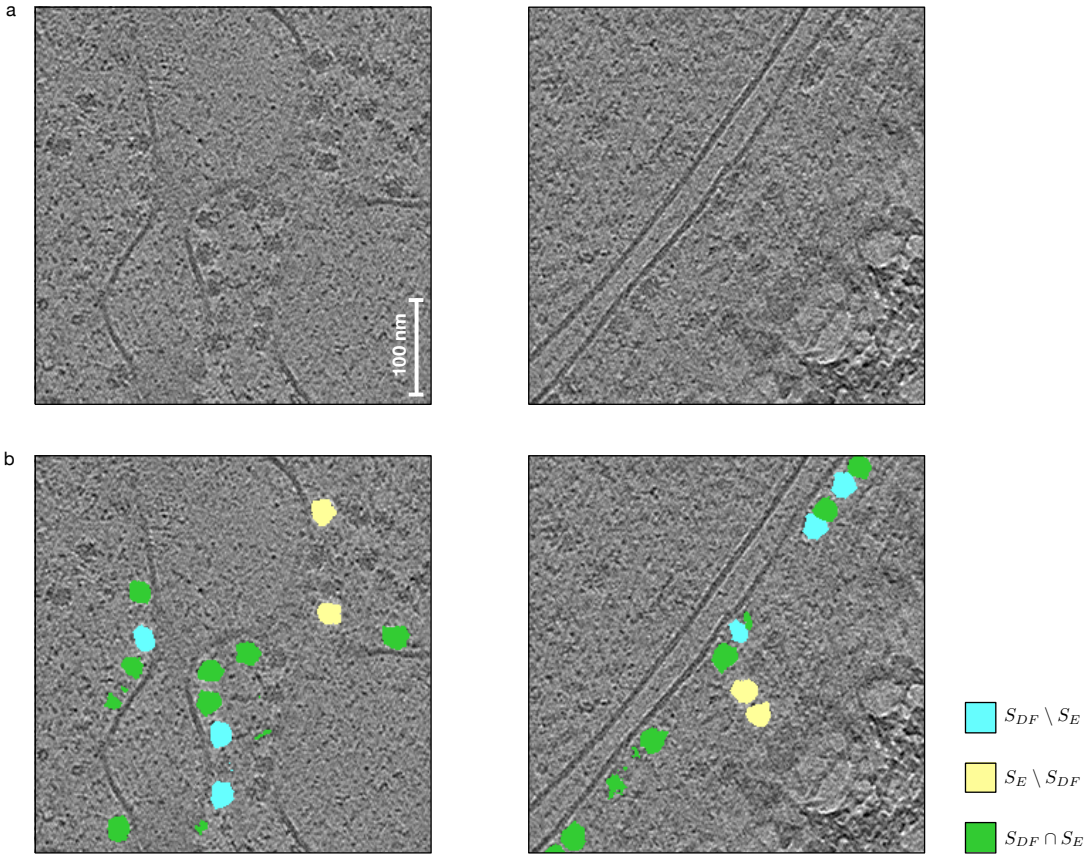


Extended Data Fig. 8: **The generalization potential of DeepFinder on P19 cells**. DeepFinder was trained on the Chlamydomonas (algae) dataset and then applied on a tomogram of mouse P19 cells (EMD-10439). Although the ribosome has a different structure for the two species, for a given voxel size (13.68 Å) 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 from 300 *mb-ribo* particles. **c,** Histogram of *mb-ribo* particle distance from the nearest cell membrane. In this histogram, the maximum mode is located at 136.8 Å, which corresponds to the ribosome radius. This analysis used 48 tomograms for training, 1 tomogram for validation, and 1 tomogram for testing.

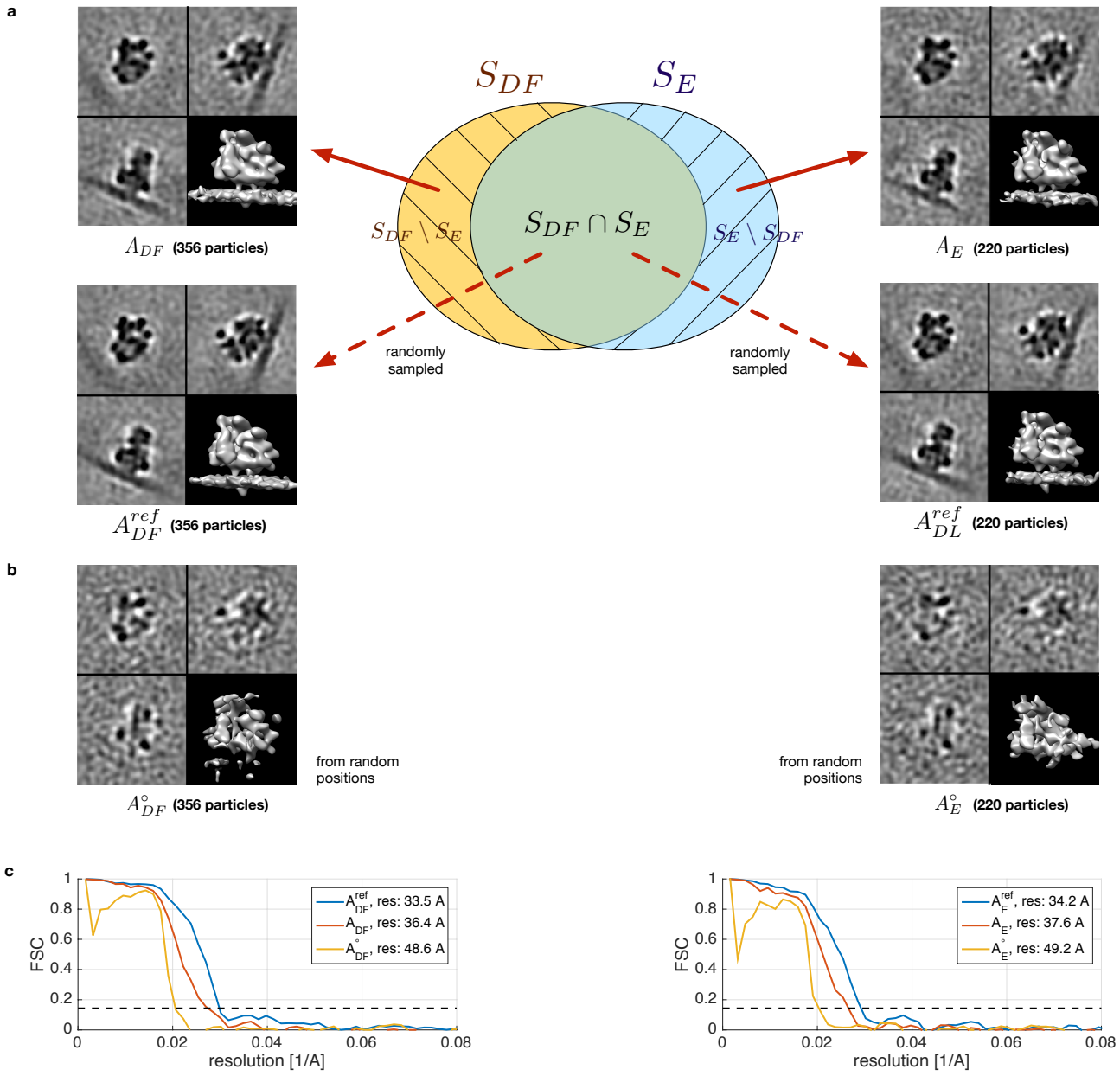
Supplementary Note 1

Analysis of consensus response

In this note, we examine the complementarity between the two sets of *mb-ribo* macromolecules found by the experts and DeepFinder. In the following analysis, we denote the sets obtained by experts and DeepFinder as S_E and S_{DF}, respectively. While the overlap $S_E \cap S_{DF}$ between both sets was substantial (1,516 particles), there was also a significant number 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. The union $S_E \cup S_{DF}$ of the two sets increases 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 require more investigation. These two non-union sets are relatively small, enabling assignment of the bulk of the high-confidence particles so the expert can focus on validating the remaining low-confidence particles. In this manner, it is possible to uncover inaccuracies in the expert annotations and refine the true-positive particle class, which can further improve the training performance of DeepFinder.



Analysis of localization consensus between DeepFinder and experts. **a,** Two tomogram slice ROIs depicting *Chlamydomonas reinhardtii* cells. **b,** Membrane-bound ribosomes mapped into the ROIs. The ribosomes found by DeepFinder but missed by the experts ($S_{DF} \setminus S_E$) are blue. The ribosomes found by the experts but missed by DeepFinder ($S_E \setminus S_{DF}$) are yellow. The ribosomes found by both DeepFinder and the experts ($S_{DF} \cap S_E$) are green. As expected, members of $S_{DF} \cap S_E$ constitute the majority of identified ribosomes. Members of $S_{DF} \setminus S_E$ tend to be found at locations where the membrane has less contrast (**b**, left) or where neighboring ribosomes are close (**b**, right). Members of $S_E \setminus S_{DF}$, which were obtained with the expert pipeline (template matching and CPCA clustering), may also be located at positions where membrane contrast is low (**b**, left). Nevertheless, it appears that this pipeline has a tendency of confusing membrane-bound and cytosolic ribosomes. The proximity of ice-contamination (**b**, right) also seems to be a factor responsible for missclassifications. This analysis used 48 tomograms for training, 1 tomogram for validation, and 8 tomograms for testing.



Analysis of consensus decisions and overlap sets (Dataset #2). a, The central Venn 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 $\mathbf{A}_{DF}$, $\mathbf{A}^{ref}_{DF}$, $\mathbf{A}_E$, $\mathbf{A}^{ref}_E$ indicate the particle subsets used 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 $\mathbf{A}^{\circ}_{DF}$ and $\mathbf{A}^{\circ}_E$ 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[7]. **c,** FSC curves for the above subtomogram averages. The averages $\mathbf{A}^{ref}_{DF}$ and $\mathbf{A}^{ref}_E$ 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 $\mathbf{A}^{\circ}_{DF}$ and $\mathbf{A}^{\circ}_E$, meaning that the impact of alignment bias is not significant. This analysis used 48 tomograms for training, 1 tomogram for validation, and 8 tomograms for testing.