Computers & Graphics (2020)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

SHREC'20 Benchmark: Classification in cryo-electron tomograms

Ilja Gubins[a], Marten L. Chaillet[b,2], Gijs van der Schot[b], Remco C. Veltkamp[a], Friedrich Förster[b], Yu Hao[c], Xiaohua Wan[c], Xuefeng Cui[d], Fa Zhang[c], Emmanuel Moebel[e], Xiao Wang[f], Daisuke Kihara[f,g], Xiangrui Zeng[h], Min Xu[h], Nguyen P. Nguyen[i], Tommi White[j], Filiz Bunyak[i]

[a]*Department of Information and Computing Sciences, Utrecht University, Netherlands*
[b]*Department of Chemistry, Utrecht University, Netherlands*
[c]*Institute of Computing Technology, Chinese Academy of Sciences, China*
[d]*School of Computer Science and Technology, Shandong University, China*
[e]*Inria Rennes Bretagne Atlantique, France*
[f]*Department of Computer Science, Purdue University, USA*
[g]*Department of Biological Sciences, Purdue University, USA*
[h]*Computational Biology Department, Carnegie Mellon University, USA*
[i]*Department of Electrical Engineering and Computer Science, University of Missouri, USA*
[j]*Department of Biochemistry, University of Missouri, USA*

ARTICLE INFO

Article history:

Keywords: Cryo-electron tomography, Computer vision, Pattern recognition, Protein classification, Benchmark

ABSTRACT

Cryo-electron tomography (cryo-ET) is an imaging technique that allows us to three-dimensionally visualize both the structural details of macro-molecular assemblies under near-native conditions and its cellular context. Electrons strongly interact with biological samples, limiting electron dose. The latter limits the signal-to-noise ratio and hence resolution of an individual tomogram to about 50Å (5nm). Biological molecules can be obtained by averaging volumes, each depicting copies of the molecule, allowing for resolutions beyond 4Å (0.4nm). To this end, the ability to localize and classify components is crucial, but challenging due to the low signal-to-noise ratio. Computational innovation is key to mine biological information from cryo-electron tomography.

To promote such innovation, we provide a novel simulated dataset to benchmark different methods of localization and classification of biological macromolecules in cryo-electron tomograms. Our publicly available dataset contains ten tomographic reconstructions of simulated cell-like volumes. Each volume contains twelve different types of complexes, varying in size, function and structure.

In this paper, we have evaluated seven different methods of finding and classifying proteins. Six research groups present results obtained with learning-based methods and trained on the simulated dataset, as well as a baseline template matching, a traditional method widely used in cryo-ET research. We find that method performance correlates with particle size, especially noticeable for template matching which performance degrades rapidly as the size decreases. We learn that neural networks can achieve significantly better localization and classification performance, in particular convolutional networks with focus on high-resolution details such as those based on U-Net architecture.

[1]*e-mail:* i.gubins@uu.nl
[2]Track organizers

1. Introduction

There is a resolution gap in knowledge of cellular life between the molecular level (obtained by techniques such as X-ray crystallography and cryo-electron microscopy single particle analysis) and the cellular level (typically obtained by light microscopy techniques) [1]. Cryo-electron tomography (Cryo-ET) has the potential to bridge this gap by simultaneously three-dimensionally visualizing the cellular context and the structural details of macromolecular assemblies [2]. This technique may offer insights into key cellular process, improve our understanding of essential life processes and the modes of action of drugs.

Cryo-ET is an application of transmission electron cryomicroscopy, in which samples are imaged as they are sequentially tilted, typically every 1 to 3 degrees from about -60° to $+60^\circ$. The resulting "tilt-series" of 2D projections are then combined in a 3D reconstruction. In cryo-ET samples are vitrified in their fully hydrated state by rapid cooling and imaged under cryogenic conditions. Rapid cooling allows imaging without dehydration or chemical fixation, which often disrupts and distorts biological samples [3].

Electron microscope's electrons strongly interact with biological samples, limiting signal-to-noise and as a result the resolution of individual tomograms to about 50Å (5nm), enough for the cellular context, but not for identifying structure of biomolecules in the sample. A common approach to increase resolution of the biomolecule of interest is to align and average copies of the same particle, introducing the challenge of correctly localizing and identifying those particles in low-resolution tomograms (Figure 1).

The core problem for this challenge is low signal-to-noise ratio of cryo-electron tomograms, often reaching extremely low values, closely followed by an incomplete reconstruction due to the limited tilt-series angles. Moreover, signal-to-noise in tomograms is strongly frequency-dependent. Multiplied by the large amount of volumetric data obtained during each imaging session, manual segmentation is rarely feasible and often provide subjective results. Instead, automated approaches are typically employed.

Particles of known structures can be found in the tomogram by template matching [4], a process of cross-correlating the template over tomogram to find peak locations and angles (i.e. location and angles where the template matches the most).

For particles with unknown structures, reference-free methods must be used. The most common approach is based on applying Difference of Gaussian (DoG) [5]: a band-pass filter that removes noisy high frequency components and homogeneous low frequency areas, obtaining edges of structures. Based on the edges, a subtomogram containing the particle can be extracted, aligned, averaged and refined with other copies of the particle present in the tomogram, allowing to obtain final, high-resolution structure of the particle.

In recent years, machine learning has seen successful application to cryo-ET. Classical support vector machines have been used for both detection and classification [6]. With ever increasing amounts of data captured by cryo-EM and -ET methods [7], deep learning methods are gaining popularity. Supervised methods were proposed for localization [8], classification [9], end-to-end segmentation [10] and joint localization and classification [11], providing faster and often more accurate results than template matching [12]. Moreover, methods based on clustering of representational features [13], segmentation by manually designed rules [14] and geometric matching [15] provide unsupervised and weakly-supervised alternatives, reducing the dependency on annotated data.

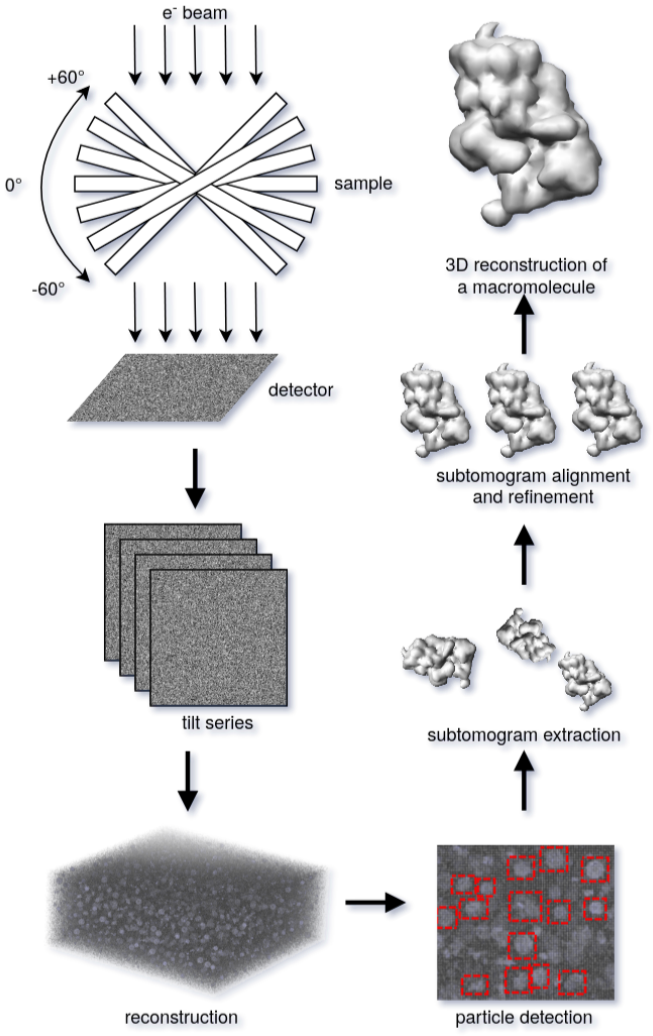

Fig. 1. The overall process of cryo-electron tomography from data collection to reconstruction and subtomogram averaging.

Each of the mentioned methods is validated on different tasks and different datasets, making it difficult to compare or draw conclusive results about their relative performance. With this paper, we aim to support researchers involved in developing new methods for localization and detection of biomolecular structures in cryo-electron tomograms. More specifically our contributions are as follows:

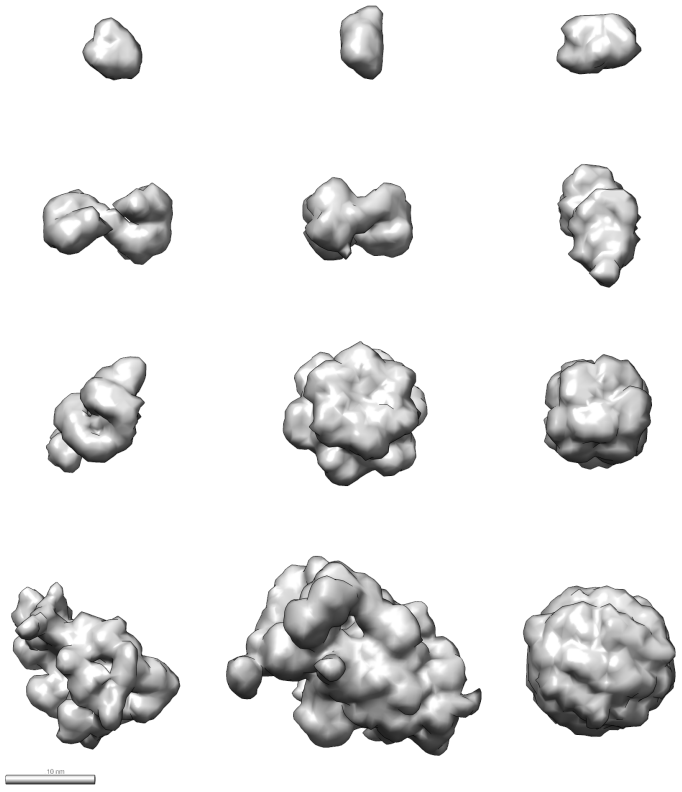

Fig. 2. 3D view of the macromolecular complexes that are present in the dataset. Sorted by their molecular weight, left to right, top to bottom: `1s3x`, `3qm1`, `3gl1`; `3h84`, `2cg9`, `3d2f`; `1u6g`, `3cf3`, `1bxn`; `1qvr`, `4cr2`, `4d8q`. Scalebar is $10nm$**.**

- We release a new, publicly available, fully-annotated simulated dataset that resembles experimentally obtained cryo-electron tomograms.
- We benchmark and conduct evaluation of six proposed learning-based methods against a strong, heavily-used baseline template matching.
- We experimentally confirm correlation between classification performance and molecular weight of a particle, highlighting the significant advantage of learning-based methods for such targets over template matching.

2. Benchmark

We propose a task of localization and classification of particles in the cryo-electron tomogram volume. A benchmark is conducted on a simulated cryo-electron tomogram populated with randomly positioned and oriented copies of structurally well-defined molecular complexes. In total, the volume contained 2782 particles of 12 different classes (Table 1). To facilitate application of learning-based methods, we also provide nine tomograms with similar protein distribution and ground truth data that was used for the simulation.

2.1. Dataset

Our dataset generation starts with creating the original density maps (grandmodels). First, to evaluate localization and classification for various size and shape proteins we chose 12 different proteins of known structure (Tables 1, Figure 2). To characterize their shape, we calculated sphericity, Ψ, a measure of how much the volume resembles a sphere:

$$\Psi = \frac{\pi^{1/3} \times (6V)^{2/3}}{A} \tag{1}$$

and effective radius, the radius of a sphere with the same surface area to volume ratio as the volume of interest:

$$r_eff = \frac{3V}{A} \tag{2}$$

where V is the volume and A is the surface area.

Next, we have generated their electron density maps at 3Å resolution with UCSF Chimera [16] and then resampled to 10Å resolution. Between 2400 and 2800 protein density volumes were placed in the "grandmodel" (ground truth sample volume) at random locations and in random $SO(3)$ orientations. The proteins were placed without overlapping each other but without limitations of how close to each other they can be, for a more realistic molecularly crowded environment (Figure 3a). For each protein volume we saved the class, the center coordinates and the Euler angles of its orientation (in ZXZ angle rotation notation). Moreover, we have saved various other ground truth artifacts: class masks (Figure 3b), occupancy maps (mapping from each voxel to corresponding particle), and their bounding boxes.

For ice simulation, we calculate the average charge density of embedding amorphous ice from a molecular water model [3] and obtain $0.15V/nm^3$. We then embed macromolecular complexes in an ice layer of $200nm$ and to encompass random variationin ice density, we add a random noise with $\sigma = 0.01$. Using our GPU affine transformation volumetric framework [4], each grandmodel was rotated over 40 evenly spaced tilt angles rang ing from $-60°$ to $+60°$ with cubic b-spline interpolation [17].

After rotation we added random structural noise, with standard deviation $\sigma = 0.04$, selected by comparison with experimental images. The structural noise varied between rotations, which modelled the sample deterioration due to the electron beam damage. To calculate the projection image for each rotation angle we implemented the multislice method [18]. This method models the defocus gradient through the ice layer by propagating the electron wave through slices of the model. We set the size of these slices to $5nm$. After calculating the wave propagation through the sample we obtain the exit wave in the image plane. To get the final projection image we multiplied the exit wave by the microscope's contrast transfer function (CTF) using a defocus of 3 micrometer, an acceleration voltage of $300kV$, amplitude contrast of 8%, and a Gaussian CTF decay of $0.4Å^{-1}$. Finally we applied CTF dependent noise and background noise with a signal-to-noise ratio of 0.004 to model

[3] NYU/ACF Scientific Visualization, Library of 3-D Molecular Structures: `http://www.nyu.edu/pages/mathmol/library/`

[4] Voltools: CUDA-accelerated NumPy 3D affine transformations, `https://github.com/the-lay/voltools`

PDB	Name	Mol. weight (*kDa*)	Volume (nm^3)	Area (nm^2)	Sphericity	Eff. radius (*nm*x)
1s3x	Hsp70 ATPase	42.75	104.1	122	0.877	2.56
3qm1	LJ0536 S106A	62.62	139.1	144.9	0.896	2.88
3gl1	Ssb1, Hsp70	84.61	207	202.6	0.835	3.065
3h84	GET3	158.08	375.3	399	0.631	2.822
2cg9	Hsp90-Sba1	188.73	394.2	380.5	0.683	3.108
3d2f	Sse1p, Hsp70	236.11	521.9	497.9	0.63	3.145
1u6g	Cand1-Cul1-Roc1	238.82	498.5	488	0.623	3.065
3cf3	P97/vcp	541.74	1123	805.7	0.648	4.181
1bxn	Rubisco	559.96	978.9	614.4	0.776	4.78
1qvr	ClpB	593.36	1255	1159	0.485	3.248
4cr2	26S proteasome	1309.28	3085	1971	0.52	4.696
4d8q	TRiC/CCT	1952.74	2152	1331	0.606	4.85

Table 1. Macromolecular complexes that are present in the dataset, sorted by their molecular weight.

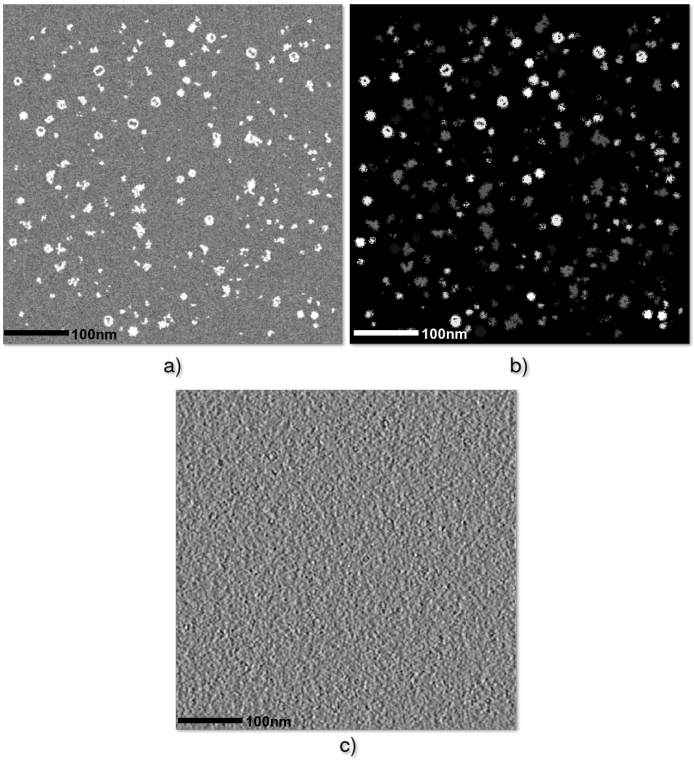

Fig. 3. Central slice of tomogram #1 in the generated dataset: (a) ground truth volume of the sample that was used for reconsutrction, (b) class mask, where each voxel is annotated by class, (c) tomographic reconstruction

the noise added by the detector's measurement. The final images were $512x512$ pixels with a pixel size of $1nm$. We did a weighted back-projection reconstruction to obtain the tomograms of $512x512x512$ with a sampling of $1nm/voxel$.

2.2. Evaluation

The main goal of the benchmark is to localize and classify biological particles in the tomographic reconstructions. The performance of the submissions has been evaluated solely on the test tomogram, the only tomogram for which ground truth is not available until after performing the test.

During evaluation, we parsed the submitted result and computed some commonly adopted performance metrics for classification and localization. The metrics are precision (Equation 3): percentage of results which are relevant; recall (Equation 4): percentage of total relevant results correctly classified; F1 score (Equation 5): harmonic average of the precision and recall; false negative rate also known as miss rate (Equation 6): percentage of results which yield negative test outcomes. We also record how far the predicted center was from the ground truth center and how many results refer to the same particles.

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{3}$$

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{4}$$

$$\text{F}_1\text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

$$\text{Miss rate} = 1 - \text{recall} \tag{6}$$

2.3. Comparison to an earlier benchmarks

Localization and classification in cryo-ET presents an open problem with major challenges due to the nature of imaging process and biological sample size (Section 1). Previous version of our benchmark [12] has already attempted to establish a comparison of the methods on a simulated, publicly available dataset, and highlight the most interesting research directions.

Since then, the dataset generation method has been considerably expanded. Multiple problems were addressed, most important of which is defective particle rotation that produced chopped particles with unrealistic, hard edges. The problem is particularly noticeable for smaller particles, where the cropping makes the number of available voxels for classification even smaller. Moreover, we have added following improvements:

- Instead of simple 2D projections, we now use multislice wave propagation algorithm [18] to better simulate electron microscope behavior.
- We now allow crowded simulations, where particles can be in direct contact with each other, instead of bounding box space limited particles.

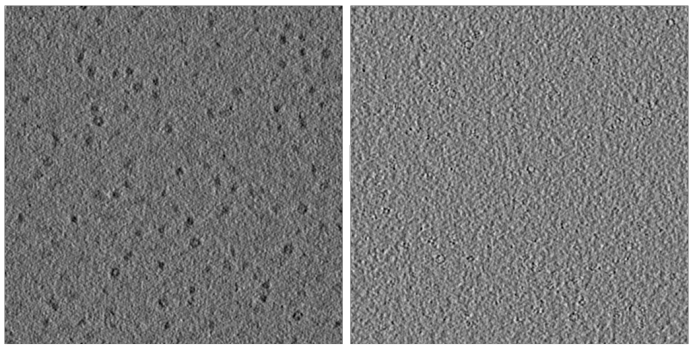

Fig. 4. Central slices of test tomograms in SHREC 2019 (left) and SHREC 2020 (right). In the previous benchmark largest classes particles are visible by eye without any processing.

- Noise model is now more precise and includes variations in ice thickness and detector measurements.
- We are able to provide more dataset generation artifacts, including class masks (voxel to class mapping) and occupancy masks (voxel to particle mapping). This makes it easier for participants to benchmark their methods and reduces the need to generate their own training data.

Another change compared to the previous dataset is the difference in the signal-to-noise ratio: 0.02 in 2019 vs. 0.004 in 2020 (Figure 4). Lower signal-to-noise ratio leads to a more challenging, but more realistic dataset, and allows to highlight methods that would generalize to experimental data the most.

3. Participants and methods

Six international research groups joined in the experimental comparison, applying seven different methods, obtaining eight output results.

3.1. Classification in cryo-electron tomograms with 3D MS-D network

By: Yu Hao, Xiaohua Wan, Xuefeng Cui, Fa Zhang

We designed a deep-learning based method to localize and classify the particles. We use 3D segmentation network to segment the tomogram. Then, the location and classification of particles are calculated by clustering algorithm.

First, the tomogram is cropped into cubic volumes with $64x64x64$ voxels. Then, our network conducts a voxel-wise classification on each cubic volume. Here, the voxels are classified from 0 to 13.

An example of a six layers 3D MS-D network with the dilation rate $\in [1, 3]$ is shown in Figure 5. 3D dilated convolution is introduced as our basic operation to reduce the number of trainable parameters, and the dense connection is applied to reuse all preceding feature maps. Our network has 64 dilated layers with dilation rate $\in [1, 16]$. It is implemented in PyTorch with CUDA acceleration and trained on 8 sets of tomograms. The model was trained for 200 epochs (spanning one week), in bathes of 125, using Adam [19] optimizer with learning rate of 0.0001 on five NVIDIA GeForce RTX 2080 Ti. The total inference time for a tomogram is 5 minutes.

We use mean-shift clustering to determine the central position of particles. In the segmented tomogram, each cluster can be regarded as a particle. To improve the localization, the centroids of 3D connected components are utilized as initial seeds to generate more precise clusters. In each cluster, the label that occurs most frequently is the classification result.

3.2. DeepFinder: Deep learning improves macromolecules localization and identification in 3D cellular cryo-electron tomograms

By: Emmanuel Moebel

DeepFinder [20] is a computational tool for multiple macromolecular species localization, based on supervised deep learning. This two-step procedure (Figure 6) first produces a segmentation map where a class label is assigned to each voxel. The classes can represent different molecular species (e.g. ribosomes, 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. The analysis of the obtained segmentation maps is achieved by clustering the voxels with the same label class, using the mean-shift algorithm. Hence, the detected clusters correspond to individual macromolecules and their positions can then be derived.

The 3D CNN architecture is trained with Adam [19] optimizer, 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. A Dice loss [21] is used to estimate the network parameters. The training took 50 hours on an Nvidia M40 GPU. For large and medium macromolecules, presented scores are reached after 22 hours; the additional time is necessary for having better performance with small macromolecules. The segmentation and clustering of a $512x512x200$ tomogram takes 20 minutes.

With feasibility in mind, we developed training strategies to assist the user in producing segmentation maps (needed for training the CNN) from tomogram annotations consisting of the spatial coordinates of macromolecules. DeepFinder is an open-source python package [5], with a graphical interface aimed towards non-computer scientist users.

3.3. Semantic segmentation using 3D ResNet with consensus checking

By: Xiao Wang, Daisuke Kihara

The method is based on 3D semantic segmentation of the tomogram data using deep learning. Given a voxel (cropped 3D region) from the tomogram, the proposed 3D-ResNet takes

[5] DeepFinder: `https://gitlab.inria.fr/serpico/deep-finder`

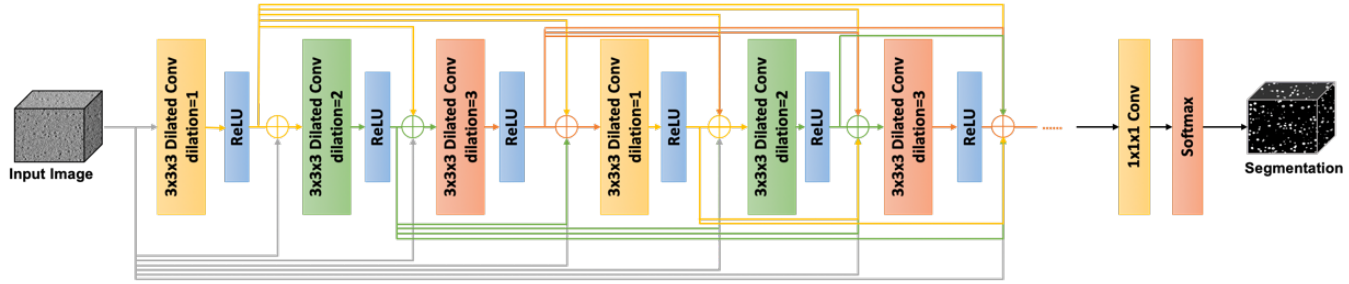

Fig. 5. 3D MS-D: Architecture of the network.

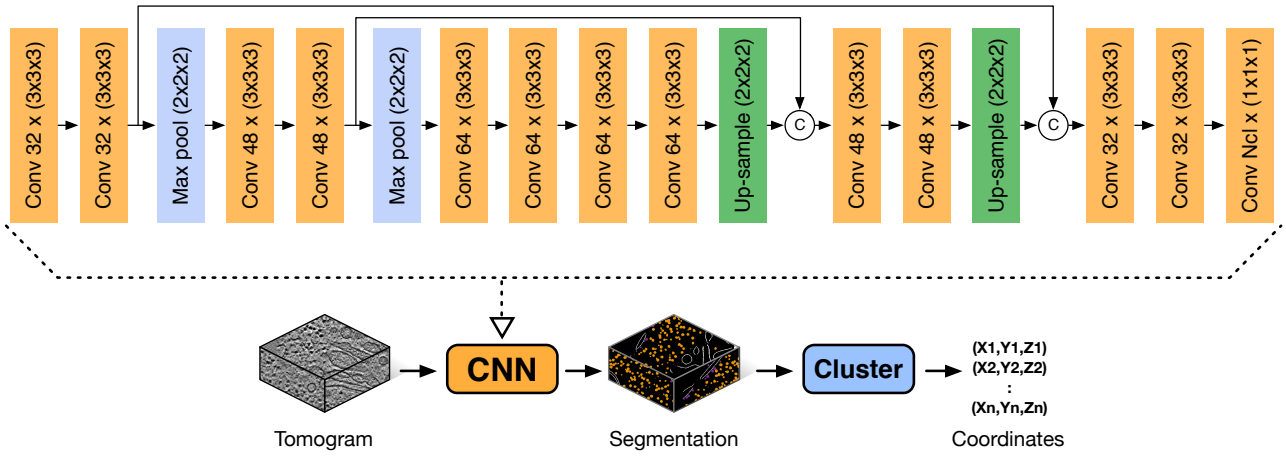

Fig. 6. Top: CNN architecture used in DeepFinder. All convolutional layers are followed by a ReLU activation function, except the last layer which uses a soft-max function. The up-sampling is achieved with up-convolutions (also called "backward-convolution"). Combining feature maps from different scales is performed by concatenation along channel dimension.
Bottom: workflow depicting how macromolecule coordinates are obtained from the segmentations generated by the CNN. A clustering algorithm (mean-shift) is applied on the segmentation map to differentiate individual macromolecules.

the voxel as input and outputs the 13 probability scores for 12 proteins and background. The size of each 3D input slice was selected to be $32x32x32$. To achieve better performance, we used ResNet of 20 layers [22, 23].

In order to train the proposed deep learning model, we first sample negative examples from the provided training tomogram data by extracting voxels with the center that is not closer than 16 grid units to any proteins. For each positive voxels (voxels that have a target protein at the center) and negative voxels, we randomly flipped and rotated before using it for training. We had in total $23,350$ positive voxels and $9,000$ negative voxels. In the training process, we used the Adam optimizer [19] with an initial learning rate of 0.002. The training took 5 hours on one NVIDIA GeForce GTX 1080Ti.

In the prediction, we use a stride of 2 to select a center point for an input voxel. When a protein label is assigned to the center of a voxel, we check labels assigned to points within the box of the size of the protein to examine the consistency of the label assignments. The box size of each target protein was provided by the organizers. We removed the prediction for the center point if its predicted label was different from the majority of the assigned labels in the box. We did not apply this majority checking protocol for small proteins (i.e. 1s3x, 3qm1, 3gl1). For each voxel with a predicted label, we counted the number of points with the same label in the protein box (majority count). Then, we clustered those points with the same label and computed the mean of their coordinates as the position of the protein. Finally, for each of the 12 target proteins, predicted voxel positions of the target protein were sorted by the majority count and the top N predictions was selected for submission, where N was decided based on the class ratio of the 12 proteins in the training set. The whole inference process takes 2 hours on one tomograms.

3.4. YOPO: one-step object detection for cryo-ET macromolecule localization and classification

By: Xiangrui Zeng, Min Xu

We formulate a novel one-step object detection framework specifically designed for cryo-ET data (Figure 7). Previous deep-learning-based works on detecting particles in cryo-electron tomograms are either two-step classification (extract potential structural regions as subtomograms and then perform classification) or segmentation methods.

Considering two important properties of subtomogram data: (1) the high-level structural details of a particle determine its function and identity and (2) the particle is of random orientation and displacement inside a subtomogram, we designed a convolutional neural network named YOPO (You Only Pool Once), which contains only one pooling layer (a global pooling

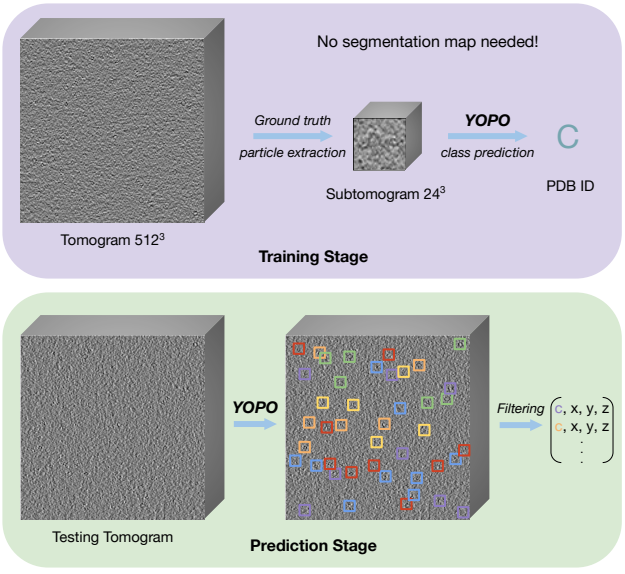

Fig. 7. YOPO: Flowchart of macromolecule detection.

layer) to retain discriminative high-level structural details and achieve the maximal transformation-invariance. The flowchart of macromolecule localization and classification using YOPO is illustrated in Figure 7). In the training stage, only particle location ground truth was used to train the YOPO network to predict the PDB ID of a subtomogram. In the testing stage, the trained YOPO network was applied on the tomogram level to directly predict the location and PDB ID of detected macromolecules.

From each training tomogram, we extract subtomograms of size 24^3 according to the ground truth particle location file. An additional 20000 subtomograms were extracted at random locations from the background. Therefore, there are $K = 13$ classes in total including the background class. Subtomograms from tomogram 0 - 7 were used as training data and subtomograms from tomogram 8 as validation data. The training took 8 hours on one NVIDIA GeForce Titan X GPU. The trained model predicted at every location by applying the learned model parameters on the whole testing tomogram. Locations with high confidence (probability > 0.99) to be one of the structural classes were kept. We then filtered the locations to ensure that the minimum distance between two detections was greater than 14 voxels.

As a one-step object detection method, the classification and localization tasks are unified in an end-to-end fashion. YOPO is an efficient cryo-ET macromolecule detection (localization + detection) framework: (1) the only ground truth information used for training is the particle locations and classes in ground truth particle location file; (2) YOPO performs prediction on a subtomogram level at every location, which is similar to the traditional template matching approach. However, the whole prediction on one tomogram took only about 40 minutes using one GPU instance.

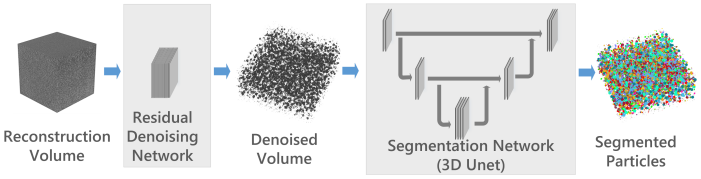

Fig. 8. Dn3DUnet: Cryo-electron volume particle detection and classification pipeline.

3.5. Cryo-electron tomogram particle localization and classification using 2D denoising network and 3D U-net pipeline

By: Nguyen P. Nguyen, Tommi White, Filiz Bunyak

In order to denoise the input tomograms and to improve the detection performance, we used DnCNN [24], a feed-forward denoising convolutional neural network utilizing residual learning strategies. The DnCNN network consists of 20 convolutional layers. The network is designed to predict residual image that is the difference between the noisy input image and the latent clean image. The network is trained using 2D XY slices from the tomogram volumes with Adam optimizer and initial learning rate of 0.0025. Data slices were split by ratio 0.8 : 0.1 : 0.1 for training, validation and test sets. We denoised the tomogram volumes slice by slice. The denoising step improved the average peak signal to noise ratio (PSNR) of the 3D volumes from 6 to 22, and the average Structural Similarity (SSIM) indices from 0.02 to 0.83 compared to the noisy input volumes. The training of the denoising network took 2 hours and 48 minutes.

The denoised tomogram volumes were fed to a modified 3D U-net [25] network, where we replaced the regular cross-entropy loss function with the general dice loss function described in [26, 21]. The network was retrained to perform semantic segmentation of the 3D tomograms patches into 13 classes (one background class and 12 classes of particles). The training was perfomed with Adam optimizer and initial learning rate of 0.001. The training of the segmentation network took 12 hours and 22 minutes. The test volumes were partitioned into non-overlapping 3D patches of size 104^3 voxels and fed to the 3D U-net.

Connected component analysis was performed to identify individual particle centroids and volumes. A post-processing step was used to filter-out spurious detections based on detection size. Detected particles with centroids within 5 voxels from the corresponding ground truth centroids were considered as detected. Detections having the same class labels as the corresponding ground truth particles were considered as correct classification. The result on average is obtained in 101 seconds (25s for denoising, 38s for segmentation, 38s for localization).

3.6. Deeply cascaded U-net for multi-task cryo-electron tomography processing

By: Ilja Gubins, Remco C. Veltkamp, Friedrich Förster

We used U-net Multi-task Cascade (UMC), a novel CNN architecture for multi-task learning (Figure 9). Inspired by U-net

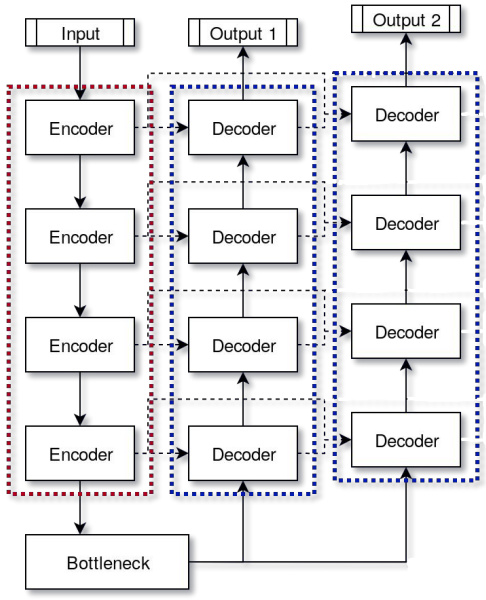

Fig. 9. UMC: Overview of U-net Multi-task Cascade architecture. Each decoder block accepts skip connections from encoder and previous decoders at the same depth level.

architecture [25], we extend it by an additional skip connection from each decoder block. Such outgoing connections allow us add multiple decoding pathways and connect them, forming deep cascades. Accordingly, UMC can be seen as a special case of a multi-task network cascade [27] where each cascade stage is a decoding pathway of U-Net. We hypothesize that connectivity between decoding pathways, facilitates inductive transfer between early and late stages of cascade. Moreover, the explicit parameter sharing acts as a form or regularization and reduces the risk of overfitting.

For cryo-ET volumes, we decided to use UMC with two output paths, one for denoising and the second for segmentation. Our hypothesis is that explicitly supervised denoising of reconstruction can help segmentation to produce better output. We use UMC with depth of 5 and following number of filters at each level: $16, 32, 64, 128, 256$, resulting in $8.82M$ of parameters. For denoising, we use mean squared error minimization objective between input reconstruction and provided ground truth grandmodel volume. For segmentation, we employ Tversky loss function [28] with $\alpha = 0.7$ and $\beta = 0.3$, targetting original provided class mask.

While developing, we used tomograms 0 to 7 for training and tomogram 8 for validation, but for the final model training we used all 9 available tomograms. We split each tomogram into patches of 64^3 voxels with 75% overlap and employ random horizontal flips for data augmentation. The model was trained for 25 epochs, in batches of 24, using Adam [19] optimizer with learning rate of 0.001. The training took 16 hours on a Tesla P100 GPU (Google Colab).

Using trained model, we segmentated the test tomogram. Then, we found connected components and filtered out components that have less than 10 voxels or have centroids less than 5 voxels away from another connected component. For a final predicted class of a particle, we took the most common occuring class in the connected component. The total inference time is 40 minutes.

3.7. Template matching

By: Gijs van der Schot, Ilja Gubins

We used the cryo-ET analysis framework PyTom [29] to conduct template matching using each of the twelve protein electron density maps as templates. The templates were modulated in the frequency domain using a standard ctf curve at $3um$ defocus. Frequencies beyond the first ctf-zero were set to 0. Spherical template masks with Gaussian smoothed edges based on the thresholded electron density were used for normalization for the cross-correlation value. We selected the top $2,000$ candidates with the highest cross-correlation score for each class and then used the candidate lists with the two following approaches:

1. Thresholded, where we take top N candidates per class going from the biggest class to the smallest one by one.
2. Filtered, where we take top N candidates per class as in the previous method, but we additionally filter out candidates that would overlap with already selected particles. To test for overlap, we calculate the distance between center of an existing particle to the center of the candidate and calculate whether the distance is smaller than the sum of their radii.

The exact number of particles in the test tomogram is unknown at the test time, only that it is a random number between 2400 and 2800 particles. Based on average of 2600 particles and 12 protein classes, we have selected $N = 217$. Template matching for 12 protein classes takes 27 hours 24 minutes on 16-core CPU (2 hours 17 minutes per class) or 1 hour 24 minutes on one NVIDIA GeForce GTX 1080 Ti (6 minutes 12 seconds per class).

4. Results

We have evaluated different metrics (Section 2.2) that allows comparison of localization (Table 3) and classification (Table 4) performance of the methods. For more convenient referencing, we have assigned following short names to the methods:

1. 3D MS-D (Section 3.1)
2. DeepFinder (Section 3.2)
3. 3D ResNet (Section 3.3)
4. YOPO (Section 3.4)
5. Dn3DUnet (Section 3.5)
6. UMC (Section 3.6)
7. TM-T and TM-F (Section 3.7)

The test tomogram has 2782 particles of the same 12 classes and same distribution as the training data (Table 2). To have a more detailed classification evaluation, we compare results with cumulative F1 score (Figure 10), as well as group proteins by their molecular weight (Table 5) and average F1 scores for an additional metric correlated with particle sizes (Table 6).

PDB	Quantity
1s3x	233
3qm1	241
3gl1	229
3h84	240
2cg9	228
3d2f	214
1u6g	217
3cf3	238
1bxn	245
1qvr	226
4cr2	231
4d8q	240

Table 2. Distribution of proteins in the test tomogram.

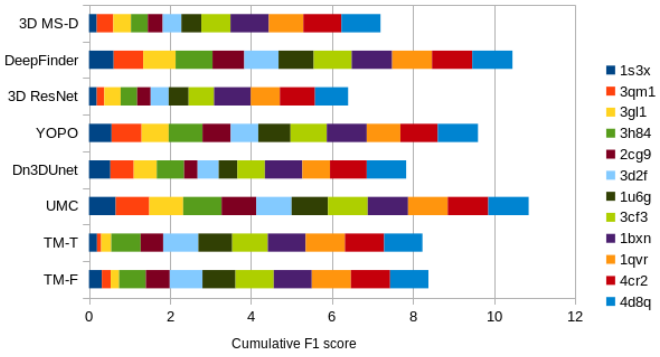

Fig. 10. Cumulative classification F1 scores of methods.

5. Discussion

Overall, the benchmark allows us to compare baseline and upcoming methods, as well as highlight current challenges in the novel cryo-ET localization and classification approaches.

Comparison with template matching. All but the template matching are learning-based methods using 3D convolutional neural networks. The results (Table 3, 4) show that learning-based methods can achieve higher performance than the traditional baseline template matching, heavily used in cryo-ET research this day. Compared to template matching, learning-based methods have the advantage of being significantly more robust to noise perturbations than cross-correlation. Another significant advantage is computational time (Table 7). The table highlights the range of possible training and inference timings of the methods. All learning-based methods require significantly less time than the traditional CPU-based template matching, often even including the training stage. Advanced GPU template matching shows significant speedup compared to CPU time, still takes longer than almost all learning-based methods.

Method performance correlates with size. Results (Table 6, Figure 12) show that there is a correlation between macromolecular complex size and classification performance for all methods. For a better overview, we have plot method classification performance vs. molecular weight (Figure 11). The performance of all methods is consistent, with similar performance dips and spikes (i.e. 1bxn at 560kDa and 2cg9 at 188kDa). The results suggest that template matching provides comparable results for finding large and medium particles, but rapidly falls behind as the size decreases. This shows potential of learning-based methods for smaller particles.

Neural network architectures. Smaller particles were found especially well by UMC (Section 3.6, and is closely follower by DeepFinder (Section 3.2). Both methods are variations of U-Net network [25] architecture and use similar overlap-based loss functions, however UMC has noticeably higher number of parameters (more filters and higher depth) and also uses additional supervision for denoising. One of the main features of U-Net architecture are skip connections that give the network higher control over feature map combination, preservation of information despite of downsampling between network levels and subsequently leads to a higher resolution, and it was first used for biomedical semantic segmentation, where high accuracy is critical. Dn3DUnet (Section 3.5) also uses U-Net for segmentation, however there is a pre-segmentation denoising done with a separate DnCNN [24] network. Using not connected networks might induce loss of information between pipeline stages and that might explain lower performance compared to other U-Net inspired methods.

Other methods also draw inspiration from neural network architectures designed for image processing. 3D MS-D (Section 3.1) using a densely connected convolutional network [30] and 3D ResNet (Section 3.3) using a residual network [22] conducted semantic segmentation of the tomograms. Alternatively, YOPO (Section 3.4) does not rely on semantic segmentation and still achieves top-3 performance.

Supervised training. All of the learning-based methods featured in the paper are supervised, requiring a training dataset with data distribution closely related to the data the method will be used on. For cryo-ET, this is a highly limiting factor that can prevent the wide adoption of deep learning, especially with semantic segmentation models that require voxel-level annotations. DeepFinder provides users with a GUI to generate such annotations. However, various approaches try to use labels are less difficult to obtain. For example, YOPO requires only coordinates and the class of a particle location, making it significantly more accessible for cryo-ET researchers.

Results compared to an earlier benchmark. One of the compared methods, DeepFinder, was also benchmarked on an earlier version of the benchmark. Previously, it has obtained localization F1 score of 0.791 and average classification F1 score of 0.565. Compared with this year performance (0.924 on localization and 0.871 on classification), the significant improvement can suggest that the new dataset is less challenging and therefore less realistic.

We have decided to conduct a baseline template matching using the same approach as described in Section 3.7. The results (Table 8, 9) show that localization in 2019 is less challenging while classification is noticeably harder. This is consistent with

Submission	RR	TP	FP	FN	MH	RO	AD	Recall	Precision	Miss rate	F1 Score
3D MS-D	2663	2523	139	259		0	2.05	0.906	0.947	0.094	0.926
DeepFinder	2594	2485	107	297	2	0	2.166	0.893	**0.957**	0.107	0.924
3D ResNet	2864	1983	611	799	246	0	3.501	0.712	0.692	0.288	0.702
YOPO	2821	2543	240	239	37	0	2.104	0.914	0.901	0.086	0.907
Dn3DUnet	2598	2340	146	442	112	0	2.807	0.841	0.9	0.159	0.869
UMC	2781	**2642**	**68**	**140**	68	0	**1.873**	**0.949**	0.95	**0.051**	**0.949**
TM-T	2604	1898	20	884	412	0	1.528	0.682	0.728	0.318	0.704
TM-F	2604	2267	331	515	6	0	1.767	0.814	0.87	0.185	0.841

Table 3. Results of localization evaluation. ***RR*****: results reported;** ***TP*****: true positive, unique particles found;** ***FP*****: false positive, reported non-existant particles;** ***FN*****: false negative, unique particles not found;** ***MH*****: multiple hits: unique particles that had more than one result;** ***RO*****: results outside of volume;** ***AD*****: average euclidean distance from predicted particle center;** ***Recall*****: uniquely selected true locations divided by actual number of particles in the test tomogram;** ***Precision*****: uniquely selected true locations divided by RR;** ***Miss rate*****: percentage of results which yield negative results;** ***F1 Score*****: harmonic average of the precision and recall. The best results in each column are highlighted.**

Submission	1s3x	3qm1	3gl1	3h84	2cg9	3d2f	1u6g	3cf3	1bxn	1qvr	4cr2	4d8q
3D MS-D	0.192	0.408	0.437	0.416	0.368	0.461	0.492	0.719	0.948	0.851	0.942	0.964
DeepFinder	0.61	0.729	0.8	0.911	0.783	0.848	0.866	0.939	**1**	**0.984**	0.993	0.993
3D ResNet	0.193	0.185	0.405	0.407	0.334	0.445	0.491	0.628	0.906	0.719	0.868	0.817
YOPO	0.558	0.741	0.67	0.834	0.696	0.682	0.795	0.896	0.987	0.83	0.923	0.993
Dn3DUnet	0.529	0.577	0.569	0.674	0.332	0.523	0.462	0.676	0.925	0.684	0.907	0.974
UMC	**0.661**	**0.827**	**0.839**	**0.947**	**0.855**	**0.873**	**0.899**	**0.981**	0.997	0.98	**1**	**0.997**
TM-T	0.2	0.102	0.248	0.727	0.555	0.869	0.835	0.88	0.934	0.97	0.968	0.945
TM-F	0.319	0.219	0.207	0.66	0.589	0.808	0.815	0.945	0.939	0.966	0.968	0.945

Table 4. Results of classification evaluation for all classes. The values correspond to F1 score achieved by methods on specific classes. The best results in each column are highlighted.

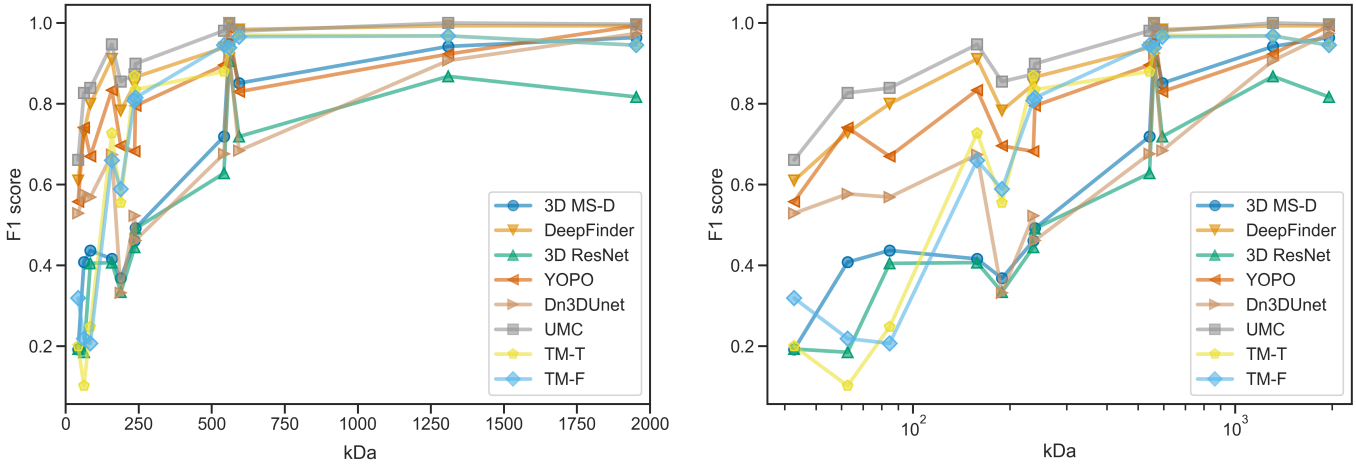

Fig. 11. Method classification performance plot against particle molecular weight. X-axis in the right plot is in logarithmic scale.

Group	Weight	Proteins
Small	<200	1s3x, 3qm1, 3gl1, 3h84, 2cg9
Medium	200 - 600	3d2f, 1u6g, 3cf3, 1bxn, 1qvr
Large	600	4cr2, 4d8q

Table 5. Grouping of macromolecular complexes by their molecular weight in *kDa*

both higher signal-to-noise ratio (making it easier to find particles, Figure 4) and the previously mentioned in Section 2.3 rotation bug (making it harder to classify found particles). Different signal-to-noise ratios make it hard to compare relative difficulty of the datasets. However there is no doubt that the current version of the benchmark is more realistic due to fixed bugs and improved simulation.

Future work. Our dataset and benchmark provides cryo-ET researchers with a baseline and highlights potential research directions. However, additional work can be done to make the comparison between algorithms stronger. First and most im-

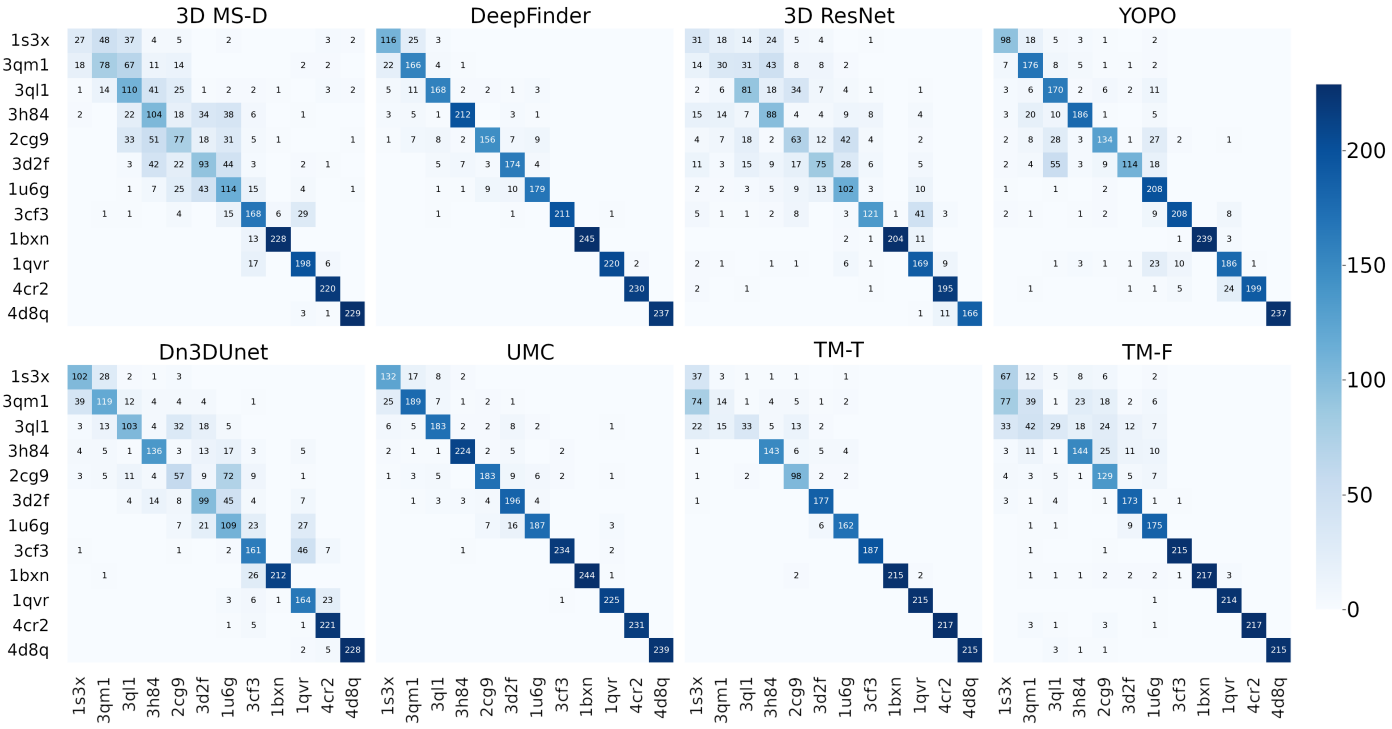

Fig. 12. Classification confusion matrices of the compared methods. The particles are ordered by molecular weight. The colorbar indicates the number of correct classifications.

Submission	Small	Medium	Large
3D MS-D	0.364	0.694	0.953
DeepFinder	0.767	0.927	0.993
3D ResNet	0.305	0.638	0.843
YOPO	0.7	0.838	0.958
Dn3DUnet	0.536	0.654	0.941
UMC	**0.826**	**0.946**	**0.999**
TM-T	0.366	0.898	0.957
TM-F	0.399	0.895	0.957

Table 6. F1 scores of each submission for size group defined in Table 5. The best results in each column are highlighted.

Method	Training stage	Inference stage
3D MS-D	168h	5m
DeepFinder	50h	20m
3D ResNet	5h	2h
YOPO	8h	40m
Dn3DUnet	15h 10m	1m 41s
UMC	16h	42m
TM-T/TM-F GPU	N/A	27h 24m
TM-T/TM-F CPU	N/A	1h 24m

Table 7. Reported training and inference stages timings. Template matching results (last two rows) are reported for both CPU and GPU processing, for all 12 classes.

portantly, the simulator has not been quantitatively validated with experimental data, leading to the question of how well the simulation captures realistic data. Next, the results show that learning-based methods achieve better performance than traditional template matching. At the same time, validation of the template matching itself is not trivial, but can be done with our simulated dataset. Such inspection can provide an insight on strengths and weaknesses of the most widely used method in cryo-ET. Finally, the benchmark should reflect the performance on the experimental data, so the simulation process can be further improved to improve transfer to the experimental domain, for example with defocus gradient and motion blur.

Acknowledgments

This work was supported by the European Research Council under the European Union's Horizon2020 Programme (ERC Consolidator Grant Agreement 724425 - BENDER) and the Nederlandse Organisatie voor Wetenschappelijke Onderzoek (Vici 724.016.001 and 741.018.201).

References

[1] Beck, M, Baumeister, W. Cryo-electron tomography: can it reveal the molecular sociology of cells in atomic detail? Trends in cell biology 2016;26(11):825–837.

[2] Yahav, T, Maimon, T, Grossman, E, Dahan, I, Medalia, O. Cryo-electron tomography: gaining insight into cellular processes by structural approaches. Current opinion in structural biology 2011;21(5):670–677.

[3] Huang, BQ, Yeung, EC. Chemical and physical fixation of cells and tissues: an overview. In: Plant microtechniques and protocols. Springer; 2015, p. 23–43.

Submission	RR	TP	FP	FN	MH	RO	AD	Recall	Precision	Miss rate	F1 Score
TM-T-2020	2604	1261	140	1521	475	0	1.934	0.453	0.484	0.546	0.468
TM-F-2020	2604	1787	813	995	4	0	2.433	0.642	0.686	0.357	0.663
TM-T-2019	2496	982	113	1558	559	0	2.975	0.386	0.393	0.613	0.389
TM-F-2019	1987	1664	291	876	31	0	2.9	0.655	0.837	0.344	0.735

Table 8. Localization results of template matching done on SHREC 2019 and SHREC 2020 datasets. *RR*: results reported; *TP*: true positive, unique particles found; *FP*: false positive, reported non-existant particles; *FN*: false negative, unique particles not found; *MH*: multiple hits: unique particles that had more than one result; *RO*: results outside of volume; *AD*: average euclidean distance from predicted particle center; *Recall*: uniquely selected true locations divided by actual number of particles in the test tomogram; *Precision*: uniquely selected true locations divided by RR; *Miss rate*: percentage of results which yield negative results; *F1 Score*: harmonic average of the precision and recall.

Submission	3qm1	1s3x	3h84	3gl1	2cg9	3d2f	1u6g	3cf3	1bxn	1qvr	4cr2	4d8q
TM-T-2020	0.046	0.128	0.34	0.064	0.178	0.429	0.28	0.579	0.683	0.583	0.937	0.94
TM-F-2020	0.064	0.169	0.364	0.053	0.207	0.492	0.34	0.653	0.696	0.611	0.91	0.945
TM-T-2019	0.083	0.038	0.038	0.075	0.109	0.251	0.035	0.234	0.676	0.107	0.679	0.82
TM-F-2019	0.188	0.07	0.14	0.158	0.133	0.426	0.191	0.21	0.308	0.149	0.704	0.668

Table 9. Classification results of template matching done on SHREC 2019 and SHREC 2020 datasets. The values correspond to F1 score achieved by methods on specific classes.

[4] Frangakis, AS, Böhm, J, Förster, F, Nickell, S, Nicastro, D, Typke, D, et al. Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. Proceedings of the National Academy of Sciences 2002;99(22):14153–14158.

[5] Voss, N, Yoshioka, C, Radermacher, M, Potter, C, Carragher, B. Dog picker and tiltpicker: software tools to facilitate particle selection in single particle electron microscopy. Journal of structural biology 2009;166(2):205–213.

[6] Chen, Y, Hrabe, T, Pfeffer, S, Pauly, O, Mateus, D, Navab, N, et al. Detection and identification of macromolecular complexes in cryo-electron tomograms using support vector machines. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI). IEEE; 2012, p. 1373–1376.

[7] Baldwin, PR, Tan, YZ, Eng, ET, Rice, WJ, Noble, AJ, Negro, CJ, et al. Big data in cryoem: automated collection, processing and accessibility of em data. Current opinion in microbiology 2018;43:1–8.

[8] Wang, F, Gong, H, Liu, G, Li, M, Yan, C, Xia, T, et al. Deeppicker: a deep learning approach for fully automated particle picking in cryo-em. Journal of structural biology 2016;195(3):325–336.

[9] Che, C, Lin, R, Zeng, X, Elmaaroufi, K, Galeotti, J, Xu, M. Improved deep learning-based macromolecules structure classification from electron cryo-tomograms. Machine vision and applications 2018;29(8):1227–1236.

[10] Chen, M, Dai, W, Sun, SY, Jonasch, D, He, CY, Schmid, MF, et al. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. Nature methods 2017;14(10):983.

[11] Li, R, Zeng, X, Sigmund, SE, Lin, R, Zhou, B, Liu, C, et al. Automatic localization and identification of mitochondria in cellular electron cryo-tomography using faster-rcnn. BMC bioinformatics 2019;20(3):132.

[12] after peer review, A. Available after peer review. Available after peer review 2019;.

[13] Xu, M, Singla, J, Tocheva, EI, Chang, YW, Stevens, RC, Jensen, GJ, et al. De novo structural pattern mining in cellular electron cryotomograms. Structure 2019;27(4):679–691.

[14] Xu, M, Alber, F. Automated target segmentation and real space fast alignment methods for high-throughput classification and averaging of crowded cryo-electron subtomograms. Bioinformatics 2013;29(13):i274–i282.

[15] Zeng, X, Xu, M. Gum-net: Unsupervised geometric matching for fast and accurate 3d subtomogram image alignment and averaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, p. 4073–4084.

[16] Pettersen, EF, Goddard, TD, Huang, CC, Couch, GS, Greenblatt, DM, Meng, EC, et al. Ucsf chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry 2004;25(13):1605–1612.

[17] Ruijters, D, Thévenaz, P. Gpu prefilter for accurate cubic b-spline interpolation. The Computer Journal 2012;55(1):15–20.

[18] Vulović, M, Ravelli, RB, van Vliet, LJ, Koster, AJ, Lazić, I, Lücken, U, et al. Image formation modeling in cryo-electron microscopy. Journal of structural biology 2013;183(1):19–32.

[19] Kingma, DP, Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980 2014;.

[20] Moebel, E, Martinez-Sanchez, A, Lariviere, D, Fourmentin, E, Ortiz, J, Baumeister, W, et al. Deep learning improves macromolecules localization and identification in 3d cellular cryo-electron tomograms. bioRxiv 2020;.

[21] Sudre, CH, Li, W, Vercauteren, T, Ourselin, S, Cardoso, MJ. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer; 2017, p. 240–248.

[22] He, K, Zhang, X, Ren, S, Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–778.

[23] Hara, K, Kataoka, H, Satoh, Y. Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017, p. 3154–3160.

[24] Zhang, K, Zuo, W, Chen, Y, Meng, D, Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Transactions on Image Processing 2017;26(7):3142–3155.

[25] Ronneberger, O, Fischer, P, Brox, T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015, p. 234–241.

[26] Crum, WR, Camara, O, Hill, DL. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE transactions on medical imaging 2006;25(11):1451–1461.

[27] Dai, J, He, K, Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, p. 3150–3158.

[28] Salehi, SSM, Erdogmus, D, Gholipour, A. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging. Springer; 2017, p. 379–387.

[29] Hrabe, T, Chen, Y, Pfeffer, S, Cuellar, LK, Mangold, AV, Förster, F. Pytom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. Journal of structural biology 2012;178(2):177–188.

[30] Huang, G, Liu, Z, Van Der Maaten, L, Weinberger, KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 4700–4708.