Practical Analysis of Macromolecule Identity from Cryo-electron Tomography Images using Deep Learning

Mostofa Rafid Uddin
Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA, USA
mru@andrew.cmu.edu

Ajmain Yasar Ahmed
Dept. of Computer Science and Engineering
Bangladesh University of Engineering & Technology
Dhaka, Bangladesh
yasar199510@gmail.com

Kafi Khan
Independent Researcher
Dhaka, Bangladesh
kafikewu@gmail.com

Md Shahrar Fatemi
Dept. of Computer Science and Engineering
Bangladesh University of Engineering & Technology
Dhaka, Bangladesh
shahrar007@gmail.com

Xiangrui Zeng
Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA, USA
xiangruz@andrew.cmu.edu

Min Xu
Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA, USA
mxu1@cs.cmu.edu

2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) | 978-1-6654-2471-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/AIPR52630.2021.9762209

***Abstract*—Cellular electron cryo-tomography (cryo-ET) has made possible the systematic 3D visualization of the near-native structures and spatial-organizations of large macromolecules (represented as subtomograms) and their interactions with organelles inside single cells. It has emerged as a major tool for in situ structural biology. However, the systematic identification of such macromolecules from cryo-ET images is very difficult due to structural complexity and imaging limits. In particular, conventional methods are too slow to process millions of highly structurally heterogeneous macromolecules fastly imaged using cryo-ET. Since 2017, supervised deep learning has become an important tool for facilitating high-throughput analysis of cryo-ET data. However, supervised learning based approaches depends on manual data annotation by biologists, which is an extremely time-consuming and burdensome process. Therefore, none of these methods are practical to use. In order to facilitate deep learning for practical identification of macromolecules from cryo-ET images, in this paper, we demonstrate the pathway towards unsupervised learning for fast and high-throughput identification of macromolecules from cryo-ET images. To this end, we demonstrate the use of three selected recent macromolecule identification methods on several commonly used benchmark cryo-ET datasets.**

***Index Terms*—Bioimage informatics, Image classification, Cryo-electron tomography, Unsupervised learning.**

I. Introduction

Cryo-electron tomography (cryo-ET) is a revolutionary 3D imaging technology that enables *in situ* visualization of macromolecular structures inside a single cell [1]. Without hampering the cell, it can image the spatial organization of macromolecules inside a cell in near native and near atomic scale [2]. Unlike cryo-electron microscopy (cryo-EM), cryo-ET does not purify the samples to be imaged and preserves their native condition [3]. Cryo-ET first vitrifies the whole cellular sample and then takes 2D projection images from different view angles. The 3D view of the sample is then reconstructed from these 2D projections [4]. Here, instead of a single type of macromolecule, every structure inside a single cell is reconstructed while preserving their spatial organizations.

However, extracting information about macromolecular structures from cryo-ET images is non-trivial and requires extensive computational processing. Due to crowded cytoplasmic environment and imaging artefacts, cryo-ET images are extremely noisy with a very low signal to noise ratio. In addition, due to spatial anisotropy, the sample can not be imaged from full ±180 tilt angle range. This limitation create missing values in the cryo-ET images, which is known as missing wedge effect [5]. Due to low SNR and missing wedge effect, cryo-ET images are hard to analyze with traditional image processing algorithms.

One of the crucial task in cryo-ET image analysis is identifying macromolecules from cryo-ET reconstructed 3D tomograms. A traditional approach for macromolecule identification is template matching [6], [7]. Given a structural template of a known macromolecule resolved by high-resolution techniques such as single-particle cryo-EM or X-ray crystallography, all possible orientations of the template is generated and scanned through the tomograms. At each location, the highest correlation among all orientations is taken to generate template-matching score map. Then a cut-off is applied to determine the locations of the target structure. However, the major drawback of template matching is that it is extremely slow. In addition, template matching is subject to template specific bias and can only detect known structures with available templates.

Due to being an extremely slow process, template matching is not practical to be used for high-throughput analysis of cryo-ET data at large scale. To this end, supervised deep learning based methods have become popular recently thanks to their speed and accuracy. In these approaches, small subvolumes of cryo-ET images each containing a macromolecule are extracted using template-free DoG particle picking. These small subvolumes are called subtomograms. Hence macromolecule identification essentially becomes a subtomogram classification problem [5]. The supervised methods use subtomograms annotated by biologists to train deep subtomogram classification methods and obtains high-throughput classification results with high accuracies. However, the efficacy of supervised learning methods is dependent on availability of annotated training data, which is very hard to obtain in cryo-ET image analysis. Annotating 3D cryo-ET images is tedious and subject to biases. Moreover, supervised approaches can only detect macromolecules with known structures. Therefore, they can not leverage cryo-ET's potential of detecting novel macromolecules inside single cells, which is one of the major advantages of cryo-ET imaging method.

To solve the above-mentioned problems, we have described gradual efforts to practically efficient identification of macromolecules from cryo-ET subtomograms in a fully unsupervised manner. To this end, we first describe a few-shot learning based method [8], which is able to conduct subtomogram classification on unseen macromolecules with few (or even one) labeled subtomograms from each kind of these structures. To achieve fully unsupervised classification, we first highlight an autoencoder based method [9] where k-means clustering on intermediate latent space is used to identify macromolecules from subtomograms. However, the performance of that method was very limited. Very recently, Zeng et al [10] have developed a high-throughput deep iterative subtomogram clustering approach (DISCA) [10] that can perform highly accurate subtomogram classification in a fully unsupervised way. However, these works have demonstrated results on their different subtomogram datasets and a comparative analysis of such methods against some common benchmark datasets is still missing. In this paper, we have provided a comparative analysis of the methods on several common experimental and realistically simulated datasets. Our analysis demonstrate the gradual path towards practical unsupervised identification of macromolecules from subtomograms using deep learning based methods and indicates some interesting future research directions that can engage the community.

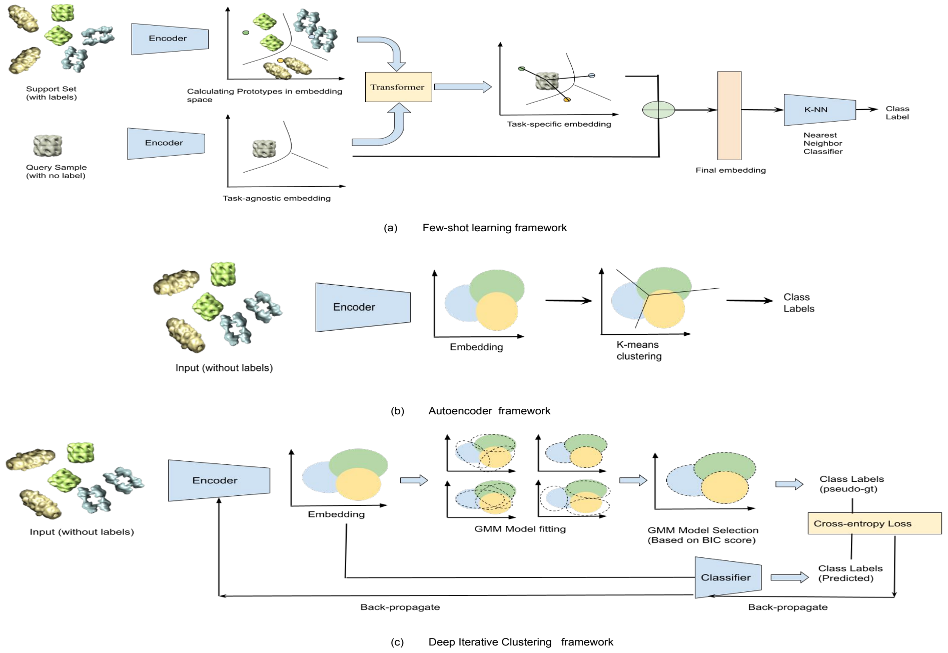
II. Related Works

Template Matching

Template matching/search has been the most widely used method for identifying known macromolecules of interest from a template. These methods calculate the structural correlation between a subtomogram or a recovered structure with a known structural template. The structural template is usually obtained by expensive high resolution methods like X-ray crystallography, NMR microscopy, or single particle cryo-EM. The correlation score is obtained by correlation or convolution operations. However, a simple correlation score cannot often conclude the template matching fully as templates are subject to reference dependent bias. To address this issue, rigorous statistical tests need to be carried out. Wang et al. [6] recently proposed a Monte Carlo sampling hypothesis testing framework based on generative adversarial network modeling for assessing template matching results. They create a generative adversarial network by using known structures to generate the structural distribution of macromolecules. First, the structure generator is trained to the extent that the discriminator cannot distinguish between a known structure and a pseudo one. Second, a large number of pseudo macromolecules are generated from the learned structural distribution in a Monte Carlo sampling fashion. Finally, the subtomogram or recovered structure of interest is compared to the known structure and pseudo structure to assess the statistical confidence of template matching. This method computes not only a correlation score of template matching but also the P-value of whether the structure is significantly close to the template. Such a statistical assessment provides rigorous evidence of template matching and reduces its false-positive rate. Though the robustness of template matching can be ensured by these approaches, template matching remains unpractical to be used for high-throughput classification of subtomograms due to computational inefficiency. While template matching, all possible orientation and shift of macromolecules need to be generated, which is a long time-consuming process. For instance, rotating a ribosome structural template at a 10° interval takes $(360^\circ/10^\circ)^3 * 161s = 87$ days on one CPU [10]. Moreover, template matching can never identify novel or unseen structures from cryo-ET subtomograms, and thus can not utilize one of the major advantage provided by cryo-ET.

Supervised macromolecule detection methods

To cope up with the exponentially increasing accumulation of cryo-ET images, high-throughput subtomogram classification methods have become a necessity. To this end, a plethora of supervised deep learning based methods [11]–[13] to identify macromolecules from subtomograms have emerged. Subtomogram classification and segmentation is held as an individual contest in 3D Shape Retrieval Challenge [12] every year. State of the art supervised subtomogram classification method includes deep convolutional neural networks (CNN), deep recurrent neural networks (RNN) and its variants, deep attention based networks, etc. These deep models can perform high-throughput and highly accurate classification of subtomograms thanks to the efficacy of deep models. However, such supervised methods suffer from a common drawback. Their performance are highly dependent on the quality of annotated data for training. In cryo-ET, good quality annotated data is very difficult to obtain as the annotation process is highly burdensome and time-consuming. Moreover, the annotation is often subject to biases. A tentative solution for scarcity of annotated data is to generate realistically simulated data to train supervised deep models. However, it has been observed that models trained on simulated data performs very poor on real data due to domain shift existent between simulated and real data [14]. To this end, some domain randomization and adaption methods have been introduced [15], [16]. Nev-



(a) Few-shot learning framework

(b) Autoencoder framework

(c) Deep Iterative Clustering framework

Fig. 1: Schematic diagram of the methods discussed in our paper. (a) shows the architecture of our few-shot learning method [8] (b) shows the architecture of the autoencoding classifier method [9](c) depicts the schematic diagram of deep iterative subtomogram clustering approach (DISCA) [10]

ertheless, these methods are usually very time consuming and yet to ensure negligible domain shift between simulated and real data. Though reducing domain shift is still an interesting open research problem, this approach introduces several extra step in subtomogram classification and makes it a very time consuming task. Therefore, a more feasible solution is to develop high-throughput and highly accurate unsupervised subtomogram classification methods.

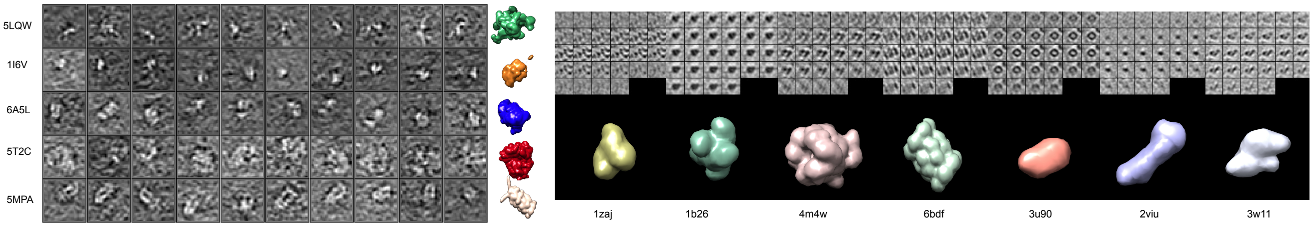
III. Methods

A. Few-shot learning

First, we describe a few-shot learning based method developed by Li et. al. [8] to conduct subtomogram classification on unseen macromolecular structures with few (or even one) labeled subtomograms from each class of these structures. In a few-shot learning task, there exists a training set consisting of considerable amount of labeled data for providing prior knowledge and a test set consisting of samples from new classes of structures which do not appear in the training set. The test set is further divided into two subsets: a support set with a few labeled samples from each class and a query set with unlabeled samples from the same class with the support set. The task is to make predictions about unlabeled samples in the query set based on few labeled samples in the support set and knowledge gained from the training set. Therefore, an X-way Y-shot classification task in few-shot learning indicates taking X classes with Y labeled samples from each class in the support set. The same sampling strategy is applied during training as well, where the training set is randomly subsampled as minibatches called episodes [17].

The few shot learning subtomogram classification approach [8] focuses on learning an embedding for each class that maintains essential features of the data and so that simple classifiers like the nearest neighbor classifier can be applied in the embedding space. Following this idea, one of the major components in this approach emerges from prototypical network (ProtoNet) [18]. The structure of ProtoNet is tailored to propose ProtoNet3D method for cryo-ET data in this work. In the embedding space learnt from ProtoNet, a prototype for each class is computed and the nearest prototype to a particular sample is the class that the sample belongs to. However, the embedding obtained through this method is a universal embedding learnt from the entirety of the training data which is, basically, a task-agnostic embedding. To extract useful information from the classification tasks, the embedding should be more targeted. For that purpose, inspired by [19], a transformation step with self-attention mechanism is added in this approach, thus obtaining a task-specific embedding. The proposed method by Li et. al [8], called ProtoNet-CE (ProtoNet with Combined Embedding) method combines both types of embedding for subtomogram classification (Figure 1 (a)).

Instance embedding based on ProtoNet3D: The architecture ProtoNet is based on a basic assumption that there exists an embedding space where each sample cluster around a class-specific prototype. Thus, in this embedding space, we can find the nearest prototype and also the class for each sample



(a) (b)

Fig. 2: (a) Sample 2D central slice subtomogram images with corresponding noise-free isosurface representations and pdb ids of macromolecules for simulated (SNR 0.1) dataset. Ten sample subtomograms per macromolecule class are visualized similarly. (b) Sample 2D slice images of subtomograms with corresponding noise-free isosurface representations and pdb ids of macromolecules for air-water interface noble single particle dataset. One sample subtomogram per macromolecule class is visualized in such way.

through a nearest neighbor classifier [18]. Because the input data are 3D gray scale images, we design a ProtoNet3D model by replacing the 2D filters with 3D filters in the ProtoNet model.

Embedding adaptation via transformer: The embedding described in the previous section is simply obtained from all training samples, regardless of the classification task in the testing set. Inspired by FEAT [19], an adaptation step to extract task-specific features via a transformer is added.

Combination of the two embeddings: In order to consider the task-specific features together with the task-agnostic features, the distances calculated in both embedding spaces are combined as the final classification criteria.

Implementation details: The original embedding function is implemented through a convolutional neural network (CNN) architecture and a 3D variant of the original ProtoNet is proposed for few-shot subtomogram classification denoted as ProtoNet3D. It contains 4 ConvBlock modules where a 3D convolutional layer with 64 parallel 3D filters are combined along with batch normalization layers, ReLu activation, and max pooling layers. The parallel 3D filters are designed to extract different features from subtomograms and the max pooling layer is used for feature selection and dimension reduction. The ConvBlocks are followed by a Flatten layer which ensures that features are integrated into a one-dimensional embedding. The transformer is implemented with an attention block concatenating three fully connected layers as the learnable weight matrices, followed by a softmax layer. Then another fully connected layer is designed to obtain the weighted average of the outputs of the attention block which is then added to the original embedding.

B. Autoencoder based fully unsupervised learning

The few-shot learning method shows great promise towards a fully unsupervised subtomogram classification approach. However, it still requires annotated data for training. As an initial effort towards fully unsupervised subtomogram classification method, we next describe an auto-encoder based method that can classify subtomograms with limited precision. In this method, the input is converted into a lower dimensional vector representation and reconstructed into another higher dimensional representation in order to characterize image features. The proposed method is termed 'Autoencoder3D', which can be divided into 3 types for explanation. Encoder3D, Decoder3D and EDSS3D Network.

The Encoder3D network encodes an small input subvolume into a 32 dimensional vector. For example, a 3D subvolume is represented as a 3D array of $m \times n \times p$ size. It's encoded vector v is a vector of $\mathbb{R}^{32}$. The Encoder consists of two 3D convolutional layers and two max pooling layers with layers and one fully connected layer for outputting a vector of length 32. The Decoder3D network is a mirror reflection of the encoder part. All hidden layers and the encoder's fully connected layers are equipped with Rectified Linear (Relu) activation whereas the convolutional output of the decoder part has linear activation. To improve the performance of autoencoder, L1 norm regularization is used to encourage sparsity in the encoded features.

The encoded feature vector is used in K-means clustering algorithm to distribute the dataset into structurally heterogeneous groups of subvolumes. Before that, the learned vectors are plotted and the user is asked to select a group of clusters of interest. The selected clusters are referred as positive clusters and later used as input for a 3D Encoder-Decoder Semantic Segmentation (EDSS3D) network for subtomogram segmentation. Zeng et al [9] demonstrated that this simple strategy can successfully cluster globule and surface subtomograms into separate clusters.

C. Deep iterative Subtomogram clustering

Now, we discuss the most recent and successful unsupervised subtomogram classification method developed by Zeng et al [10], called Deep Iterative Subtomogram Clustering Approach (DISCA). This is a high-throughput template-and-label-free deep clustering approach based on a generalized Expectation Maximization framework. The model is trained to distinguish sets of homogeneous structures from the 3D spatial features and their distribution.

a) ***Transformation-invariant and noise-robust feature extraction****:* In order to ensure the maximum possible availability of depth-wise information, the feature extraction in DISCA is done using a special CNN based network named **YOPO** (You Only Pool Once). The YOPO architecture model consists of one Gaussian Dropout layer followed by a series of Convolutional Neural Network (CNN) layers. A Global Max Pooling layer is used againts the output of the CNN layers. Further, a fully connected layer is added with softmax activation to generate a class label as output. Max or average pooling is not used in YOPO to avoid drastic information losses. The global max pooling layer after the convolutional layers preserve structural details of subtomograms very well. Moreover, this pooling operation takes up feature information of similar characteristics irrespective of the coordinates of the area of interest. Therefore, it ensures translation invariance for input subtomograms by YOPO.

To make YOPO rotation invariant, a randomly rotated copy of the subtomogram was fed as input alongside the original input subtomogram at each iteration. The empty space for the rotated copy was filled up with gaussian white noise (0 mean, 1 standard deviation). To preserve the robustness to noise, a Gaussian Dropout Layer was introduced, randomly silencing 50% nodes and introducing 1-Centered Gaussian noise with standard deviation of 1. This strategy was introduced to apply some denoising in the dataset, ensuring robustness to noise.

*b) **Statistical Modelling of Feature Space**:* Second order statistics is much suitable for detecting differences of visual features than generic clustering algorithms like K-means or hierarchical clustering. So, to calculate the feature covariance completely, the learnt feature vectors from YOPO are modeled for each representative structural pattern as multivariate Gaussian distributions in the feature space.

In short, if P is the dimensionality of feature space and $x_n \in \mathcal{R}^p$ is the extracted feature for a subtomogram s_n, x_n is modeled as a mixture of K multivariate Gaussian distributions. Then, the probability distribution function for x_n is calculated using the following equation,

$$f_g(x_n;\phi,\mu,\Sigma,K)=\sum_{k=1}^{K}\phi_k g(x_n;\mu_k,\Sigma_k)$$

Here, ϕ_k is the prior probability of sampling x_n from the k^{th} component, which is a multivariate Gaussian Distribution g with mean μ_k and co-variance matrix Σ_k .So, the equation for figuring out the posterior of sampling x_n for the k^{th} component would be as follows:

$$\rho_k(x_n)=\frac{\phi_k g(x_n;\mu_k,\Sigma_k)}{\sum\limits_{i=1}^{K}\phi_i g(x_n;\mu_i,\Sigma_i)}$$

. Solving the model for the first equation provides the probability $\rho_k(x_n)$. Consequently, $\hat{k} = \text{argmax}_{k\in[K]}\,\rho_k(x_n)$ is regarded as the class label output for input subtomogram x_n.

*c) **Automated Selection of** K:* A challenge in such approach is to settle with a specific number of K. As this is an unsupervised learning approach, the number cannot be exactly determined beforehand. An automatic estimation approach for solving this become necessary, which itself is a hard and challenging problem. Nevertheless, the estimation of K is significant in a sense that, neither a too small K nor a too large K is expected because both of them would lead to poor subtomogram averaging by either putting a lot of heterogeneous structures in the same subset or causing over-partition between subsets.

Some approaches for finding appropriate K would be predicting K or running the algorithm repeatedly using bagged samples, or using measures like Silhoutte Coefficient [20]. However, these approaches have poor scalability and high time complexity. To overcome these, DISCA partakes a statistical model selection approach. The number of model parameters increases with K, so this might increase likelihood as well as create overfitting. DISCA balances likelihood and number of parameters among a set of models with different Ks and selects one among a set of fitted models. DISCA uses Bayesian Information Criterion (BIC) [21]. At each iteration of DISCA, the feature space for input subtomograms are fitted with multivariate Gaussian distributions across a set of candidate K values. The model with lowest BIC score is kept and the corresponding class label output $\hat{k}$ for each subtomogram is regarded labels for next iteration.

*d) **Iterative Dynamic Labelling**:* At each EM iteration in DISCA, pseudo-ground truth labels are generated for each input subtomogram, which are used as training labels at following iterations. The labels at first step is estimated by random initialization of YOPO architecture and are used as labels for second iteration. The whole process is revisited in the next iterations until convergence.

However, with such approach, mislabeling might occur during the initial iterations and can be propagated until convergence. To address this issue, DISCA uses label smoothing regularization technique by followed the equation, $l_{ls} = (1-\alpha)\times l_{hot}+\frac{\alpha}{K}$, where K is the number of clusters in that iteration, l_{hot} is the one-hot encoding of the labels gained from the previous iteration, and α is a smoothing factor. K is also kept dynamic in different iterations in this process.

Another issue is that the number of labels (K) is not fixed from the first iteration until convergence in DISCA. As a solution, when a new K is found in the current iteration, the last layer (classification layer) is replaced with a new one with K number of nodes.

*e) **Matching Inter-step Clustering Labels**:* During whole training phase, K stays the same most of the time. Nevertheless, it could certainly happen that one group of samples get differently labelled in consecutive iterations despite the number of labels (K) stay the same. To address this issue, DISCA directly matches clustering labels between consecutive steps by formulating this as a maximum weighted bipartite graph matching problem. If the clusters of a solution are considered as vertices of a graph, the vertices of the current solution and the vertices of the previous solution creates two disjoint sets. The edges represent that there are common samples between two vertices from different clustering and the edge weight is defined by the number of common samples shared between those clusters. DISCA uses Hungarian algorithm [22] to determine a matching in this graph where no two edges share a common vertex and the total edge weight is maximized. Furthermore, DISCA rearranges the current clustering labels according to the solution.

IV. Results

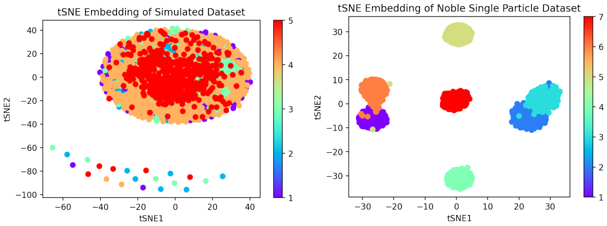
A. Datasets

We have used five realistically simulated and two experimentally obtained real datasets to perform comparative analysis of the above-mentioned methods. A brief description of the datasets are as follows:

a) Simulated Datasets: We have used five realistically simulated datasets with varying SNR and tilt angle range ±60 (Missing Wedge Angle 30°). The dataset contains five representative macromolecular complexes: spliceosome (PDB ID: 5LQW), RNA polymerase-rifampicin complex (PDB ID: 1I6V), RNA polymerase II elongation complex (PDB ID: 6A5L), ribosome (PDB ID: 5T2C), and capped proteasome (PDB ID: 5MPA). Each type of macromolecule are present in 1000 subtomograms counting to 5000 subtomograms in total. Among the five simulated datasets, one is relatively clean (SNR 100) and four are with SNR close to the experimental conditions (0.1, 0.05, 0.03, and 0.01). Each subtomogram is of size 32^3 with voxel size 1.2 nm. 2D central slice images of 10 sample subtomograms per macromolecule class along with the corresponding noise-free isosurface representations are provided in Figure 2 (a).

b) Experimental Air-water interface noble single particle dataset: We used a single particle dataset from EMPIAR that contains 2800 subtomograms of 7 distinct types of macromolecular structures. Among these seven types, rabbit musle aldolase (PDB ID: 1ZAJ) was collected from EMPIAR 10130 and 10131, glutamate dehydrogenase (PDB ID: 1B26) were obtained from EMPIAR 10133, DNAB helicase-helicase (PDB ID: 4M4W) from EMPIAR 10135, T20S proteasome (PDB ID: 6BDF) from EMPIAR 10143, apoferritin (PDB ID: 3U90) from EMPIAR 10169, hemagglutinin (PDB ID: 2VIU) from EMPIAR 10172 and insulin-bound insulin receptor (PDB ID: 3W11) were collected from EMPIAR 10173. Each of these types have 400 subtomograms in the entire dataset. The subtomograms have a SNR of 0.5 and missing wedge angle of 30°. Each subtomogram is of size 28^3 with voxel size 0.94 nm. Stacked 2D slice images of one sample subtomogram per macromolecule class along with the corresponding noise-free isosurface representations are provided in Figure 2 (b).

c) Experimental rat neuron culture dataset: This dataset contains 18419 subtomograms extracted from rat neuron culture tomogram by expert annotation and template matching [23]. Among them 15167 subtomograms are false positive and contains no macromolecule, 1095 subtomograms contain ribosome, 1043 subtomograms contain double capped proteasome, 386 subtomograms contain single capped proteasome, 460 contains membrane, and 268 subtomograms contain TriC structures. Each subtomogram is of size 40^3 with voxel size 1.368 nm and −50° to +70° tilt angle range (30° missing wedge). The subtomograms have a SNR of 0.01.



(a)

(b)

Fig. 3: (a) tSNE embedding plot of simulated (SNR 0.1) dataset. tSNE plots demonstrate the approximate ‘hardness’ of the classification problem for the particular dataset. (b) tSNE embedding plot of noble single particle experimental dataset

B. Experimental Design

1) Few-shot learning:

Classification on real datasets: Due to the smaller number of available classes in the real datasets, we removed the validation set and randomly divided them into training and testing sets. For noble single particle dataset, 4 classes were kept in training set and the remaining 3 classes were kept in testing set. And for rat neuron dataset, 3 classes were kept in training set and the remaining 3 classes were kept in testing set. The best model for evaluating against testing set was, thus, picked according to the performance on the training set. The classification accuracy of the model was calculated through 100 episodes, each with randomly sampled 3 classes, for obtaining the mean classification accuracy. For each randomly sampled class in an episode, 5 samples were picked in support set for 5-shot analysis and 1 sample was picked in support set for 1-shot analysis along with 15 samples in query set in both analyses. The experimentation was conducted with the ProtoNet3D method. The results for real datasets are shown in II.

Classification on simulated datasets: Due to small number of available classes in the simulated datasets, we removed the validation set and randomly divided them into training and testing sets. Among the available classes, 3 classes were kept in training set and the remaining 2 classes were kept in testing set. This division of classes remained consistent among different SNR levels. The best model for evaluating against testing set was, thus, picked according to the performance on the training set. The classification accuracy of the model was calculated through 100 episodes, each with randomly sampled 2 classes, for obtaining the mean classification accuracy. For each randomly sampled class in an episode, 5 samples were picked in support set for 5-shot analysis and 1 sample was picked in support set for 1-shot analysis along with 15 samples in query set for both analyses. The experimentation was conducted with the ProtoNet3D method. The results for simulated datasets are shown in I.

2) Autoencoder: As the autoencoder approach is fully unsupervised, there was particularly no need for train-test split. The model was trained with the entirety of the datasets. Furthermore, since the labels were available with the datasets, after clustering, accuracy was calculated by comparing the clusters with the labels. While training the autoencoder, learning rate and learning rate decay were set to 0.001, 2×10^{-6} respectively based on cross-validation results. Finally, the number of clusters, K, was set to the number of classes for each dataset as it provided the optimum results.

3) DISCA: As DISCA is also a fully unsupervised method like autoencoder, the entire dataset were used to train the model. To do so, cross-validation was performed to tune the hyperparameters and the learning rate, iteration, and reg cover set to 1×10^{-5}, 20, and 1×10^{-5} yielded the optimum results. Furthermore, the number of clusters, K, was set to the number of classes for each dataset. Increasing or decreasing K led to a relatively higher DDBI, and therefore, was avoided. Finally, the same hyperparameters were used to train all the datasets.

C. Results

Subtomogram classification results on the simulated datasets for the discussed methods are demonstrated in Table I. Corresponding results on the two experimental datasets are provided in Table II.

We can observe that accuracy of all methods drops with the decline of SNR in our simulated datasets. To provide an estimate of the 'difficulty' of clustering, we provide a 2-dimensional tSNE (t-Distributed Stochastic Neighbor Embedding [24]) scatter-plot of simulated SNR 0.1 dataset in Figure 3 (a) and color the embedding based on macromolecule identity. The tSNE plot testifies that separating out the macromolecules into separate homogeneous structure classes is a not an easy task. However, the corresponding tSNE plot of air-water interface dataset (Figure 3 (b)) shows a different pattern and it can be inferred that subtomograms are more easily separable in noble dataset compared to simulated SNR 0.1 dataset. Consequently, the accuracies for noble datasets were higher on average.

In our evaluation of rat neuron dataset subtomogram clustering, we assumed false positive subtomograms as a separate class. Due to its high presence, the methods were able to separate it easily from true positive subtomograms. Such evaluation significantly uplifted the clustering accuracy of the methods, however, the comparison among methods still remain fair.

In almost all the datasets (except noble single-particle), 3-way-5-shot few shot learning achieved the highest accuracy followed by DISCA. Only in noble single particle dataset, 3-way-1-shot accuracy is slightly higher than that of DISCA. The highest accuracy of 3-way-5-shot in all scenarios is expected as it includes 5 labeled sample per macromolecule class in test set as support, which introduces label specific inductive bias in the model. On the contrary, the fully unsupervised methods- Autoencoder and DISCA does not use any label associated information. Though Zeng et. al. [9] demonstrated that applying k-means on autoencoder features can distinguish between surface and globule subtomograms, it can not distinguish between finer details of macromolecule classes. This fact is clearly evident from the autoencoder results against the datasets.

Nevertheless, DISCA, the deep iterative clustering based approach, can successfully separate macromolecules of different classes into distinct clusters. In other words, DISCA can classify subtomograms in completely unsupervised manners with high accuracy; bypassing the need of exhaustive data annotation and reference dependent templates.

TABLE I: Results on Simulated Dataset

Simulated Dataset	Method	Accuracy
SNR 100	Few Shot	$\mathbf{0.963 \pm 0.008}$ (2-way 5-shot)
		0.892 ± 0.021(2-way 1-shot)
	Autoencoder	0.215
	DISCA	*0.912*
SNR 0.1	Few Shot	$\mathbf{0.867 \pm 0.014}$ (2-way 5-shot)
		0.745 ± 0.023(2-way 1-shot)
	Autoencoder	0.214
	DISCA	*0.814*
SNR 0.05	Few Shot	$\mathbf{0.798 \pm 0.018}$ (2-way 5-shot)
		0.657 ± 0.028(2-way 1-shot)
	Autoencoder	0.212
	DISCA	*0.806*
SNR 0.03	Few Shot	$\mathbf{0.776 \pm 0.015}$ (2-way 5-shot)
		0.677 ± 0.022(2-way 1-shot)
	Autoencoder	0.203
	DISCA	*0.781*
SNR 0.01	Few Shot	$\mathbf{0.570 \pm 0.016}$ (2-way 5-shot)
		0.525 ± 0.019(2-way 1-shot)
	Autoencoder	0.20
	DISCA	*0.560*

TABLE II: Results on Experimental Datasets

Dataset	Method	Accuracy
Air-water interface noble single particle	Few Shot	$\mathbf{0.994 \pm 0.002}$ (3-way 5-shot)
		0.962 ± 0.012(3-way 1-shot)
	Autoencoder	0.162
	DISCA	0.86
Rat neuron culture	Few Shot	$\mathbf{0.992 \pm 0.003}$ (3-way 5-shot)
		0.892 ± 0.016(3-way 1-shot)
	Autoencoder	0.441
	DISCA	*0.95*

V. Discussion and Future Directions

The experimental results ensures the efficacy of unsupervised approaches for macromolecule identification from subtomograms, overcoming the hurdles caused by traditional template matching and supervised approaches. However, there is still a large room for improvements. For instance, in very low SNR settings (SNR < 0.03), the classification is close to random guess (50%). This can be mitigated with effort towards joint denoising and classification of subtomograms. Developing methods for learning noise-invariant features in extremely low SNR conditions is also an interesting direction to explore. Moreover, the classification accuracy difference between fully unsupervised DISCA and few-shot learning with very low supervision is higher in experimental datasets than simulated ones. This can be due to many 'extra' complexities in experimental data that are not present in realistically simulated data. Consequently, learning 'good' features is easier in simulated data than real ones. As unsupervised methods solely depend on encoded features to distinguish subtomograms, the feature quality and complexity largely affects the performance. Therefore, developing methods to learn 'better' features that are less sensitive to the 'extra' complexities present in experimental data is also a direction to investigate in future. Nonetheless, unsupervised subtomogram clustering methods can overcome many issues caused by traditional supervised approaches and can lead to ground-breaking advantage of cryo-ET image analysis domain.

VI. Conclusion

Cryo-electron tomography (cryo-ET) is a revolutionary imaging technology that enables *in situ* identification of macromolecules in 3D images of cellular samples. With the growing accumulation of cryo-ET images, fast and high-throughput methods that accurately identifies macromolecules from such images has become a necessity. The traditional template matching based approaches are extremely slow and low-throughput and the fast supervised learning approaches are dependent on extremely time consuming annotation process, which is subject to biases. In this work, we have described a pathway towards fully unusupervised identification of macromolecules from cryo-ET subtomograms as a solution to such problems. We have performed a comparative evaluation of three subtomogram classification methods against common benchmark datasets, which shows the promise of unsupervised learning for bias-free high-throughput identification of macromolecules from subtomograms. However, we also indicated that the methods are still far from perfect, specially in very low SNR settings. To this end, we have discussed some interesting directions that the corresponding research community may explore in future.

Acknowledgment

We thank Dr. Qiang Guo for sharing with us experimental rat-neuron tomograms that we make use of as real data.

Funding

This work was supported in part by U.S. National Institutes of Health (NIH) grants R01GM134020 and P41GM103712, U.S. National Science Foundation (NSF) grants DBI-1949629 and IIS-2007595, AMD COVID-19 HPC Fund, and the Mark Foundation For Cancer Research 19-044-ASP. X.Z. was supported in part by a fellowship from Center of Machine Learning and Health at Carnegie Mellon University.

References

[1] F. K. Schur, "Toward high-resolution in situ structural biology with cryo-electron tomography and subtomogram averaging," *Current opinion in structural biology*, vol. 58, pp. 1–9, 2019.

[2] M. Beck and W. Baumeister, "Cryo-electron tomography: can it reveal the molecular sociology of cells in atomic detail?," *Trends in cell biology*, vol. 26, no. 11, pp. 825–837, 2016.

[3] F. R. Wagner, R. Watanabe, R. Schampers, D. Singh, H. Persoon, M. Schaffer, P. Fruhstorfer, J. Plitzko, and E. Villa, "Preparing samples from whole cells using focused-ion-beam milling for cryo-electron tomography," *Nature protocols*, vol. 15, no. 6, pp. 2041–2070, 2020.

[4] M. Chen, J. M. Bell, X. Shi, S. Y. Sun, Z. Wang, and S. J. Ludtke, "A complete data processing workflow for cryo-et and subtomogram averaging," *Nature methods*, vol. 16, no. 11, pp. 1161–1168, 2019.

[5] A. Bartesaghi, P. Sprechmann, J. Liu, G. Randall, G. Sapiro, and S. Subramaniam, "Classification and 3d averaging with missing wedge correction in biological electron tomography," *Journal of structural biology*, vol. 162, no. 3, pp. 436–450, 2008.

[6] K. Wang, X. Zeng, X. Liang, Z. Huo, E. P. Xing, and M. Xu, "Image-derived generative modeling of pseudo-macromolecular structures – towards statistical assessment of electron cryotomography template matching," in *British Machine Vision Conference (BMVC)*, 2018.

[7] X. Wu, X. Zeng, Z. Zhu, X. Gao, and M. Xu, "Template-based and template-free approaches in cellular cryo-electron tomography structural pattern mining," in *Computational Biology*, ch. 11, pp. 175–186, Australia: Codon Publications, 2019.

[8] R. Li, L. Yu, B. Zhou, X. Zeng, Z. Wang, X. Yang, J. Zhang, X. Gao, R. Jiang, and X. Min, "Few-shot learning for classification of novel macromolecular structures in cryo-electron tomograms," *PLOS Computational Biology*, 2020.

[9] X. Zeng, M. R. Leung, T. Zeev-Ben-Mordehai, and M. Xu, "A convolutional autoencoder approach for mining features in cellular electron cryo-tomograms and weakly supervised coarse segmentation," *Journal of structural biology*, vol. 202, no. 2, pp. 150–160, 2018.

[10] X. Zeng, A. Kahng, L. Xue, J. Mahamid, Y.-W. Chang, and M. Xu, "Disca: high-throughput cryo-et structural pattern mining by deep unsupervised clustering," *bioRxiv*, 2021.

[11] I. Gubins, G. v. d. Schot, R. C. Veltkamp, F. Förster, X. Du, X. Zeng, Z. Zhu, L. Chang, M. Xu, E. Moebel, A. Martinez-Sanchez, C. Kervrann, T. M. Lai, X. Han, G. Terashi, D. Kihara, B. A. Himes, X. Wan, J. Zhang, S. Gao, Y. Hao, Z. Lv, X. Wan, Z. Yang, Z. Ding, X. Cui, and F. Zhang, "Classification in Cryo-Electron Tomograms," in *Eurographics Workshop on 3D Object Retrieval* (S. Biasotti, G. Lavoué, and R. Veltkamp, eds.), The Eurographics Association, 2019.

[12] I. Gubins, M. Chaillet, G. van der Schot, R. Veltkamp, F. Förster, Y. Hao, X. Wan, X. Cui, F. Zhang, E. Moebel, X. Wang, D. Kihara, X. Zeng, M. Xu, N. Nguyen, T. White, and F. Bunyak, "SHREC'20 Benchmark: Classification in cryo-electron tomograms," *Computers & Graphics*, 2020.

[13] M. Chen, W. Dai, S. Y. Sun, D. Jonasch, C. Y. He, M. F. Schmid, W. Chiu, and S. J. Ludtke, "Convolutional neural networks for automated annotation of cellular cryo-electron tomograms," *Nature methods*, vol. 14, no. 10, pp. 983–985, 2017.

[14] J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer, *Dataset shift in machine learning*. Mit Press, 2009.

[15] L. Yu, R. Li, X. Zeng, H. Wang, J. Jin, G. Yang, R. Jiang, and X. Min, "Few shot domain adaptation for in situ macromolecule structural classification in cryo-electron tomograms," *Bioinformatics*, 2020.

[16] H. Bandyopadhyay, Z. Deng, L. Ding, S. Liu, M. R. Uddin, X. Zeng, S. Behpour, and M. Xu, "Cryo-shift: Reducing domain shift in cryo-electron subtomograms with unsupervised domain adaptation and randomization," *arXiv preprint arXiv:2111.09114*, 2021.

[17] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[18] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4077–4087, 2017.

[19] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Learning embedding adaptation for few-shot learning," 2019.

[20] S. Aranganayagi and K. Thangavel, "Clustering categorical data using silhouette coefficient as a relocating measure," in *International conference on computational intelligence and multimedia applications (ICCIMA 2007)*, vol. 2, pp. 13–17, IEEE, 2007.

[21] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[22] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[23] Q. Guo, C. Lehmer, A. Martínez-Sánchez, T. Rudack, F. Beck, H. Hartmann, M. Pérez-Berlanga, F. Frottin, M. S. Hipp, F. U. Hartl, *et al.*, "In

situ structure of neuronal c9orf72 poly-ga aggregates reveals proteasome recruitment," *Cell*, vol. 172, no. 4, pp. 696–705, 2018.

[24] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.