Towards unsupervised classification of macromolecular complexes in cryo electron tomography: challenges and opportunities

E. Moebel, C. Kervrann

March 10, 2022

Abstract

Background and Objectives: Cryo electron tomography visualizes native cells at nanometer resolution, but analysis is challenged by noise and artifacts. Recently, supervised deep learning methods have been applied to decipher the 3D spatial distribution of macromolecules. However, in order to discover unknown objects, unsupervised classification techniques are necessary. In this paper, we provide an overview of unsupervised deep learning techniques, discuss the challenges to analyze cryo-ET data, and provide a proof-of-concept on real data.

Methods: We propose an unsupervised sub-tomogram classification method based on transfer learning. We use a deep neural network to learn a clustering friendly representation able to capture 3D shapes in the presence of noise and artifacts. This representation is learned here from a synthetic data set.

Results: We show that when applying k-means clustering given a learning-based representation, it becomes possible to satisfyingly classify real sub-tomograms according to structural similarity. It is worth noting that no manual annotation is used for performing classification.

Conclusions: We describe the advantages and limitations of our proof-of-concept and raise several perspectives for improving classification performance.

1 Introduction

Cryo electron tomography (cryo-ET) enables to visualize sub-cellular environment at nanometer scale, allowing to identify macromolecular complexes in their native state. The field of view and the resolution are both large enough to enable a joint study of the cellular context and structures. Therefore cryo-ET acts as a link between low resolution (e.g., light microscopy) and high resolution (e.g., X-ray crystallography) imaging techniques, filling the gap of knowledge between several scales. However,

noise and artifacts in cryo-ET are such that heavy computational processing is needed to access the image content.

In recent years, a number of supervised deep learning based techniques have been developed for identifying macromolecular species in cellular cryo-ET [1, 2]. Although effective, the application of supervised methods to cryo-ET still remains limited for several reasons. First, a significant amount of annotated data is required to train deep neural networks. Unfortunately, manual and semi-automatic annotation is a time consuming and tedious task, and prone to bias. Second, the generalization is not guaranteed when an input tomogram has been acquired with different parameters and imaging conditions compared to training tomograms. While preliminary results [2] exist, transferring learning across experimental contexts is an open question. Finally, biologists are bound to identifying already known macromolecule species. Identifying new and unknown species (or conformational states of a species) requires unsupervised classification techniques as already investigated in [3] [4] [5].

In the remainder of the paper, we present in Section 2 an overview of unsupervised deep learning methods and their potential application to cryo-ET. In Section 3, we describe our proof-of-concept method, which was especially designed to cope with the limitations imposed by cryo-ET data. In Section 4, we present preliminary results showcasing the potential of our approach. Finally, in Section 5 we discuss these results, and propose several ideas to improve the performance obtained so far.

2 Unsupervised deep learning and its promises in cryo-ET

Unsupervised deep learning This research field is very active in artificial intelligence (AI) and encompasses several topics, such as representation learning, self-supervised learning, clustering, as well as transfer learning. We detail below these different concepts and explain how they relate to each other.

Deep neural networks learn from data through the optimization of an objective cost function. The optimization algorithm is exclusively of the family of stochastic gradient descent. Therefore, the necessary condition to employ deep neural networks, regardless of the setting (supervised or unsupervised), is the ability to generate a significant gradient signal. This is achieved by designing a loss function which evaluates the distance of the network output to a so-called training target. In the supervised setting this target is a set of annotations (i.e., labels). In the unsupervised setting, the target is either a data model or is obtained via self-supervised learning. In the self-supervised setting, the target is generated automatically from the data using ad-hoc heuristics. For instance, auto-encoders [6] are typical self-supervised learning methods, for which the targets are identical to the input data of the network. Here, the goal is not to classify data, but to determine low-dimensional and reversible representations. In the literature, a variety of self-supervised tasks have been proposed for image representation and

analysis, such as in-painting [7], colorization [8] or predicting transformations [9]. The fact that humans are not involved in the target generation is why deep neural networks are considered to achieve unsupervised learning.

Representation learning This process [10] allows us to learn a (non-linear) mapping between the raw data space and an alternate feature space, in which the task at hand is easier to solve. The feature space is a low dimensional representation of raw data and encodes more abstract concepts, which tends to boost the performance on downstream tasks. If the downstream task is clustering, as it is the case for unsupervised classification, then it is desirable that data points are separable with respect to the characteristics of interest. The representation learning task can be either supervised (e.g., transfer learning) or self-supervised (e.g., contrastive learning, deep clustering).

Transfer learning In this approach [11], the representation learned for condition A is used for condition B. The underlying assumption is that the same representation may be useful in both conditions. This concept is well understood in supervised image classification, and arises from the observation that the first layers of CNNs (near the input) tend to resemble Gabor filters or color blobs, regardless of the task at hand. It is well established [11] that the first layers capture low-level features such as edges and shapes, and are relevant for tasks as diverse as dog breeds and mushroom species classification.

A useful application of this feature transferability is when the available training set is small and the risk of over- or underfitting exists. In this case, one can benefit from networks trained on a huge data sets (e.g., ImageNet [12]), by *freezing* the weight-values of the first layers and therefore train only the last task-specific layers. In this way only few weights have to be trained and computation is less demanding. Another strategy is to apply *fine-tuning*; instead of freezing transferable weights, they are used to initialize a network. The consequence is that the network parameters are closer to the target configuration (compared to random initialization), therefore significantly accelerating the convergence of the training procedure.

Transfer learning is also used to acquire general purpose feature extractors (i.e., deep features). It is now common to re-use as a representation the output of intermediate layers of networks that have been trained on ImageNet (e.g. VGG or ResNet). These representations are exploited to solve various problems, including unsupervised classification. As described in [13], coupling clustering algorithms such as k-means with these representations can lead to remarkable classification accuracies. That being said, while representations derived from ImageNet are very powerful for natural images (i.e., photographs), it is not always the case when they are applied to other imaging modalities like microscopy [14], as illustrated and discussed later on. Depending on the problem, these representations

may not discriminate the desired features.

Contrastive learning These methods [15] have received interest due to success in self-supervised representation learning. The strategy relies on comparison of sample pairs, instead of individual samples. The idea being that, in the learned representation, similar samples should be mapped together, while dissimilar samples should be pushed away. The key issue is to define pairs of similar and dissimilar samples. This is resolved by applying strong image augmentation (i.e., transformations like rotation, cropping, resizing, blur etc). Similar pairs are then obtained as being distorted copies of the same sample, and dissimilar pairs as being different samples.

In the case of cryo-ET data, care must be taken as to which transformations to use. In fact, in a macromolecule classification task, the main clue is its structure (including size and chirality). Therefore transformations such as resizing, mirror operations or elastic deformations (i.e., diffeomorphism) are not appropriate. In addition, the missing wedge is responsible for anisotropic resolution in the 3D image, inducing delocalization of densities along the z direction, as well as loss of information in the xy plane along the direction perpendicular to the tilt axis. We do not believe that a network should be invariant to the orientation of this anisotropy, and hence using random 3D rotations for data augmentation should be avoided. This leaves us with few choices left, like random translations and linear transformations on voxel values (e.g., contrast changes), which limits the potential of contrastive learning methods in cryo-ET.

Joint representation learning and clustering In order to guaranty that a learned representation is clustering friendly, a family of methods named deep clustering [16, 17, 18] adapts the underlying training objective. Here, the training loss is a combination of self-supervised and clustering-specific objectives. The self-supervised objective is used to impose constraints on the learned representation, and popular choices include the reconstruction loss or the self-augmentation loss [16] (derived from contrastive learning). The clustering-specific objective ensures that the learned representation is clustering-friendly, and includes losses such as the k-means loss, cluster assignment hardening, and group sparsity loss [16]. The clustering itself is generally performed by k-means (centroid-based) or agglomerative clustering (hierarchical).

Nowadays, several unsupervised deep learning methods achieve promising accuracies on benchmark data sets such as MNIST ($> 95\%$) and ImageNet ($> 80\%$), reducing the performance gap with supervised methods. This being said, supervised methods will always perform better, as confirmed by experiments on these benchmarks.

Application to cryo-ET While these performances are encouraging and represent good opportunities in cryo-ET in the future, the aforementioned methods provided results in well-defined conditions. The conventional benchmarks constitute an ideal classification setting: the data is not corrupted by noise, the classes are equally balanced (same number of examples per class), and the data set size is huge (ImageNet consists of more than 1 million images). These properties are not satisfied with cryo-ET data. Here, the SNR level is exceptionally low, and tomograms are corrupted by imaging artifacts (e.g., missing wedge). Macromolecule populations vary significantly, and the target species may be rare. Finally, the orders of magnitude of dataset size differ: for cryo-ET, a data set composed of 50 tomograms containing 10000 instances of a macromolecule class (e.g., ribosomes) is considered as large. But more frequently, datasets are constituted of approximately 10 tomograms.

As shown in [14], the performance of deep clustering methods depends on the amount, quality, and type of the input data. The authors show that among tested data types, bioimage data was the one for which these methods perform worst, and conclude that these methods are "*not suitable for small-sized data sets*". Furthermore, it is not clear how these techniques behave when a class is rare: will clusters be detected if the number of members is too low?

When considering an application such as cryo-ET, an additional limitation of deep learning-based representation learning is interpretability. There is no direct way of identifying which attributes the classification is based on, which is a major disadvantage for gaining insights on detected clusters.

In this paper, we present a proof-of-concept of what appears to us the most sensible choice for achieving unsupervised classification in cryo-ET. We propose to use a synthetic cryo-ET data set (i.e., a data model) to learn a clustering-friendly representation. This learned representation is then used for clustering experimental data. This approach has several advantages. First, the learned representation is interpretable, because we define explicitly the training target (i.e., the labels) and therefore we control what the network encodes. Secondly, this strategy is scalable, as we can generate as much synthetic training data as needed. This makes the approach appealing to analyze both small and large data sets (i.e., real tomograms). Third, this approach is adapted for discovering rare macromolecule species, as the representation learning does not depend on manual annotation. Therefore it is not affected by the issue of under-represented classes, because training data generation is essentially free. We illustrate this concept in the setting of subtomogram classification (which can be extended to segmentation).

5

3   A proof-of-concept method for unsupervised low-dimensional representation learning

Our objective is to find a non-linear mapping $f_\theta : X \to Z$ where $X \in \mathbb{R}^D$ is the data space and $Z \in \mathbb{R}^N$ is the learned representation, where $N < D$. This representation should (i) satisfyingly characterize 3D shapes in the presence of noise and artifacts, as observed in cryo tomograms. It should also be (ii) rotation invariant, in order to avoid subtomogram alignment as in [4] and [5]. Once such a representation has been derived, one can apply classical clustering algorithms (e.g., k-means) on these feature vectors to group unknown objects according to their structural similarity (see Fig. 1 B). Consequently, applying subtomogram averaging to objects belonging to the same cluster should reveal the high-resolution structure of unknown objects. As we will see, in our case no clustering loss was necessary to yield good clustering results.

We choose f_θ to be the encoding part of the DeepFinder network from our previous work [2]. The latter is a encoder-decoder architecture designed for segmentation (i.e., Unet like). The parameters θ are learned in the frame of a subtomogram classification task. We assume that a mapping f_θ learned from such a task should meet our requirements (i) and (ii) defined above. We finally add fully connected layers as shown in Fig. 1 A to fit the encoder network to the classification task. Once the parameters θ have been learned, we obtain the desired representation (i.e., feature extractor) by pruning the output layer as illustrated in Fig. 1 A, resulting in a feature vector Z of size $N = 32$.

Our approach involves two data sets denoted X_{source} and X_{target}. The source data set X_{source} consists of simulated sub-tomograms and corresponding class labels. These labels are automatically obtained from the simulation algorithm and therefore do not require manual annotation. The data set X_{source} is used to learn the representation f_θ. On the other hand, we have the target data set X_{target} which consists of real sub-tomograms. This data set is not involved in training whatsoever. We also have labels for X_{target} which are only used to evaluate the clustering achieved on $f_\theta(X_{target})$.

4   Experimental results

In this section, we first describe the data sets X_{source} and X_{target} and our training procedure, then we define our evaluation metrics, and finally present our results.

4.1   Description of synthetic and real data sets

Source data set   To our knowledge, SHREC is the only publicly available benchmark for evaluating algorithms on localization and classification tasks in cryo-ET. This data set is based on a physics-based
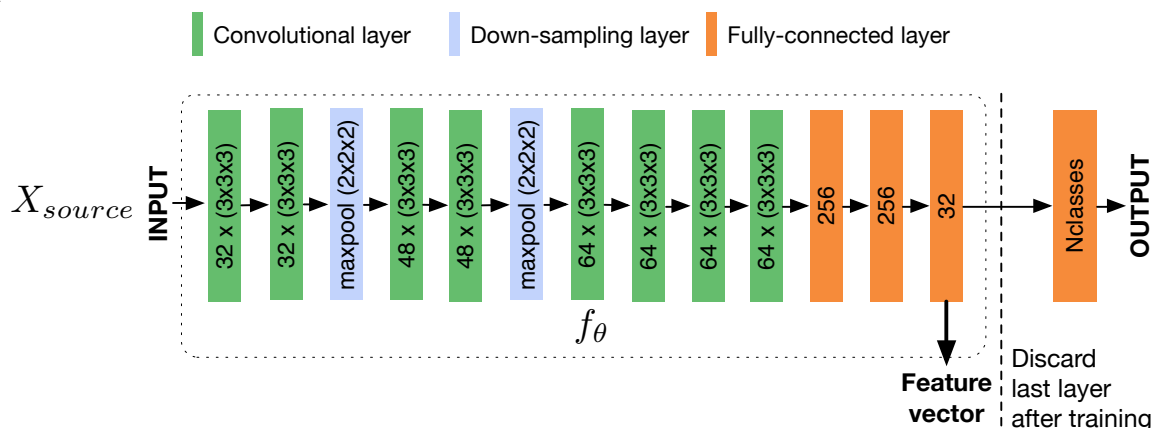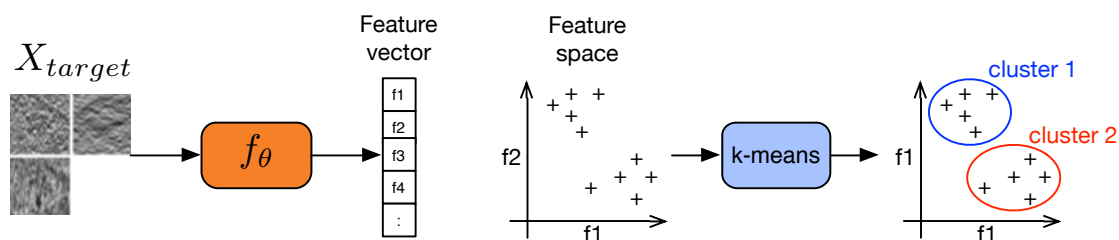
Figure 1: Unsupervised subtomogram classification. (A) Architecture of the CNN used to learn a feature space that characterizes 3D shapes. The network is trained on a synthetic data-set composed of 13 different macromolecule classes, obtained from the PDB databank. Once trained, we discard the last layer. We now have a network that takes as an input volumes and outputs feature vectors of size 32. (B) We use this network to compute the feature vectors of experimental sub-tomograms. We finally achieve unsupervised classification of these sub-tomograms by applying k-means clustering on their feature vectors. The obtained clusters group the sub-tomograms by structural similarity.

model and is constituted of 10 tomograms with 10Å/voxel resolution and a size of $512 \times 512 \times 512$ voxels. The tilt range is $\pm 60°$ with a tilt increment of $2°$. In the benchmark published in 2021 [19], each tomogram consists of up to 1500 macromolecules, as well as membranes and gold fiducials. The macromolecule population is constituted of 13 different classes (obtained from the Protein Data Bank) of varying molecular weights (from 42kDa to 3326kDa). This data set has been designed to detect the break-point of particle picking algorithm, and contains several macromolecule classes which have very low weights (5 classes are below 200kDa).

From these tomograms, we extract a set of sub-tomograms of size $40 \times 40 \times 40$ voxels, resulting in a total amount of 12905 training sub-tomograms and 1189 validation sub-tomograms.

Target data set Our target data set is obtained from a set of real tomograms depicting *Chlamydomonas Reinhardtii* cells. The tomograms have a 13.68Å/voxel resolution and a size of $928 \times 928 \times 464$ voxels. The tilt range is $\pm 60°$ with an increment of $2°$. This data set was used in several studies [20, 21, 22, 23, 2] and was carefully annotated by experts. We gathered annotations for membrane-bound ribosomes, cytoplasmic ribosomes ($\simeq 3.2$ MDa), proteasomes ($\simeq 750$ kDa) and rubisco complexes ($\simeq 560$ kDa). These annotations were exploited to extract the corresponding sub-tomograms of size $40 \times 40 \times 40$ voxels. We resized the sub-tomograms to obtain a 10Å/voxel resolution to fit the source data set. In the end, our test set is constituted of 500 sub-tomograms (100 per class).

For both data sets, we also collected sub-tomograms to constitute the negative class. In order to make sure that these do not contain macromolecules, they were sampled outside of the lamella. An illustration of both data sets is shown in Fig. 4.

4.2 Training procedure

For training we used the ADAM algorithm, chosen for its good convergence rate, by setting the learning rate to 10^{-4}, the exponential decay rate to 0.9 for the first moment estimate and to 0.999 for the second moment estimate. We used the categorical cross-entropy as training objective and set the batch size to 64.

The number of examples per class varies slightly in the SHREC data set. Therefore, we apply the same resampling strategy as in [2], so that the distribution of classes in a batch is uniform. In order to speed up the training, we initialized the network with weights obtained from the DeepFinder network, trained for the SHREC 2021 challenge [19].

4.3 Methods and metrics for visual assessment and classification evaluation

We evaluated our unsupervised proof-of-concept method on the validation and test data sets described in Section 4.1.

For a visual assesment of the learned data representation, we embedded the feature space in $\mathbb{R}^2$ using the t-SNE algorithm [24]. T-SNE is a non-linear dimensionality reduction technique, whose objective is to find a faithful low-dimensional representation of a high-dimensional space. We obtained similar results by appying UMAP [25] for visual assessment.

Next, we evaluated the k-means clustering in terms of homogeneity, completeness and V-measure. Those metrics are based on normalized conditional entropy measures of the clustering with respect to ground truth labels. The values range from 0 to 1. Values close to 1 correspond to a better classification. Homogeneity measures the ratio of samples of a single class pertaining to a single cluster. Completeness measures the ratio of members of a given class that are assigned to the same cluster. The V-measure is the harmonic mean between homogeneity and completeness.

4.4 Evaluation of the results

First, we assess the learned representation on X_{source} (see Fig. 2 A,C). In the two-dimensional space calculated by the t-SNE algorithm, the classes are organized in well separated groups, with very few data points belonging to the wrong groups. In order to test if the representation Z is clustering friendly, we apply k-means and compute metrics for different parameter values (i.e. the number of clusters). The best V-measure (0.96) is obtained when the cluster number parameter is 13, which corresponds to the number of classes.

Next, we tested the generalization potential of f_θ on X_{target}. As is illustrated in Fig. 2 B, the t-SNE plot shows that the data points are organized in three main groups: one group corresponds to the negative class, another one to the membrane-bound and cytoplasmic ribosomes, and a last one to the proteasomes and rubisco complexes. Accordingly, this feature space allows to discriminate well ribosomes and the negative class. Note that this representation has difficulties differenciating *mb-ribos* from *ct-ribos*. This can be explained by the fact that SHREC does not contain examples of membrane-bound macromolecules (altough it contains membranes). Therefore the network learned to focus the classification on the 3D structure (as expected), but without taking the neighborhood into account. Yet the context is precisely what differenciates membrane-bound from cytoplasmic ribosomes. This plot also shows that the representation does not separate well the proteasome and rubisco complex classes. That being said, we observe two sub-groups, each with an over-representation of one of these classes.

According to Fig. 2 D, the best V-measure (0.76) is obtained when the cluster number is set to 3.

9

In this case, homogeneity shows that each cluster contains only members of a single class in 63% of the cases, and completeness shows that in 96% of the cases all members of a given class are assigned to the same cluster. We note that the same completeness score as for the source data set has been obtained, but the homogeneity is significantly lower (0.63 instead of 0.96). The homogeneity measure is consistent with the visual analysis provided by t-SNE, and this value drop demonstrates the limited ability to transfer f_θ from synthetic to real data.

Nonetheless, while these results point out the weaknesses of this approach, it is also clear that f_θ, which has been obtained without manual annotations, is able to structure real data. If we examine the composition of the clusters in Fig. 3 B when the cluster number is set to 4, cluster #1 contains 100% of the negative sample class, cluster #2 contains 98.5% of ribosomes (regardless of their binding state), cluster #3 contains 73% of proteasomes and cluster #4 contains 84% of rubisco complexes. These results are still preliminary but are promising, leading us to believe that deep learning based unsupervised classification methods will play an important role in the analysis of cryo-ET data in the forthcoming years.

5   Discussion and conclusion

In this paper, we provided a brief overview of unsupervised classification methods based on deep learning. We presented the challenges, obstacles and promises to analyze cryo-ET data.

Our experiments suggest that an adapted strategy is to decouple representation learning from the clustering (as opposed to deep clustering methods). Learning the representation from a simulated source data set allows one to better control what the network encodes. This facilitates interpretability of the feature space, which is an important asset in the task of studying unknown macromolecules. As such, we have succeeded in obtaining a desired feature of this representation, namely an invariance to the macromolecule orientation, which has been achieved by design of the source data. Rotation invariance is an appealing property for reducing computation time, as it avoids the use of sub-tomogram alignment procedures, as is the case with conventional methods [3, 4, 5]. Moreover, this approach is very fast as it only took 8 seconds to classify 400 sub-tomograms of size $40 \times 40 \times 40$ on a Tesla K80 GPU. On the other hand, training took 2 hours, which is relatively short for deep learning. That being said, the advantage of the proposed approach is that the user does not need to train the network himself.

This approach can be improved in future works. First, we can improve the definition of the source data set. SHREC'21 is composed of 9 training tomograms (12905 sub-tomograms) and 13 classes, which is small compared to benchmarks like ImageNet. As a good data model is available, we can potentially generate infinite amount of data to increase the size of the training set. Also, we can increase

10

the number of macromolecule classes. In particular, we can add different conformation states of the same macromolecule species, as well as several binding states (e.g., membrane-bound). Training on such a data set would force the network to encode a more detailed representation of the macromolecule structure and its biological local context. We showed that currently, differentiating binding-states is one limitation of this approach.

Second, we would like to point out that our results have been achieved with under-sampled data, i.e. a pixel size of 10Å, while methods like [3, 4, 5] use full resolution. In fact, the results presented in [3] and [4] have been obtained with a pixel size of 6Å, and in [5] with a pixel size of 2.62Å. Consequently, reducing the voxel size of the source data set represents one axis of improvement for this strategy.

Finally, there is a generalization issue when switching from the source to the target data set. This low performance results from the fact that the source and target data do not have the same statistical distribution. The feature extractor f_θ has been fitted to the source distribution only, and hence the clustering performance on the target distribution is sub-optimal. The family of methods that aim to address this issue is known as *domain adaptation*. Here, the focus is on learning a representation that is both discriminative and invariant to the change in data distributions [26, 27]. Recently, such methods have been applied successfully to cellular electron microscopy images [28, 29]. We believe that domain adaptation constitutes a promising perspective to achieving robust unsupervised classification for cryo-ET.
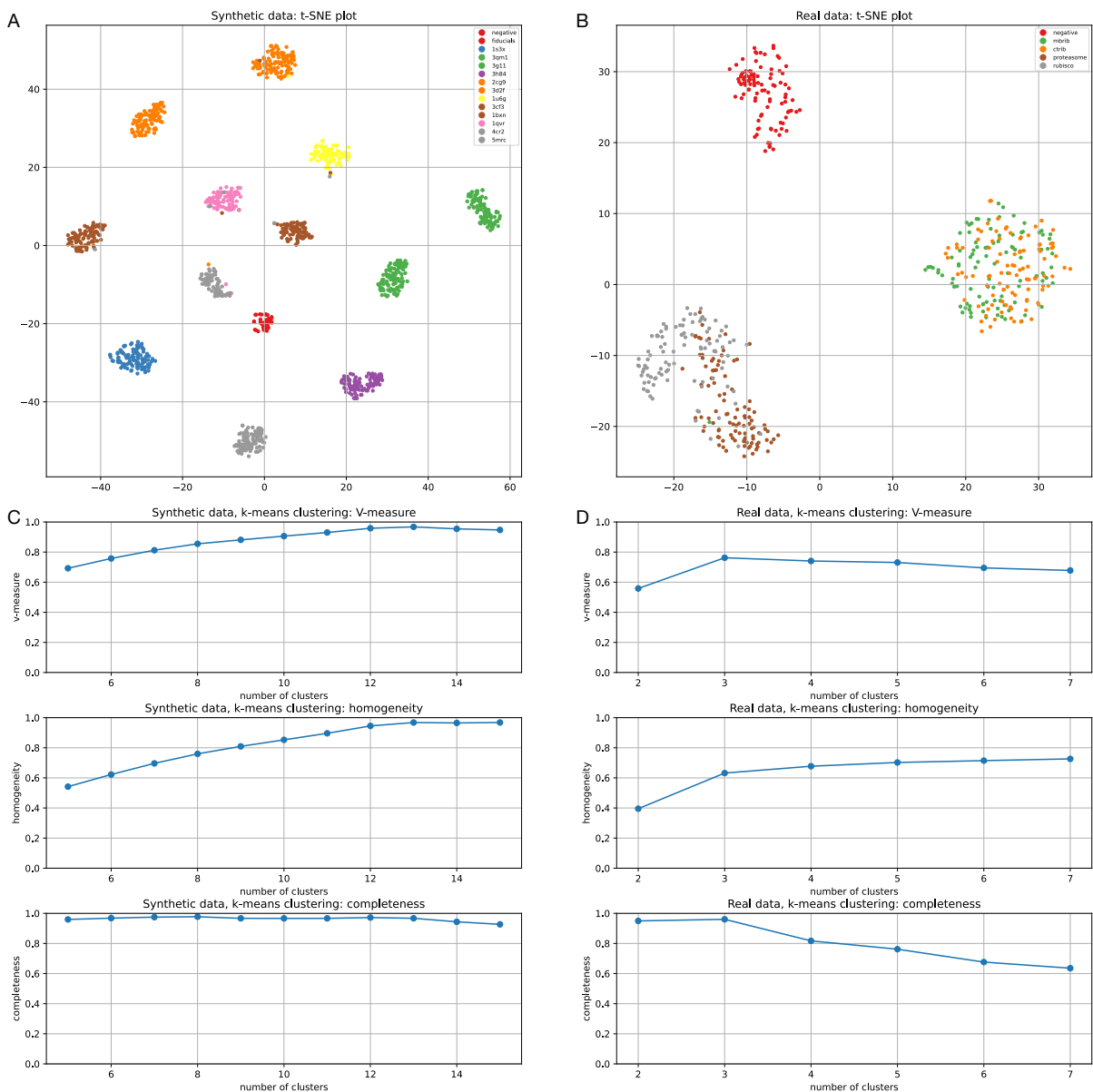
Acknowledgements

Figure 2: Evaluation of the learned representation. (A,B) 2D embedding of the 32-dimensional feature space. The embedding is obtained with the t-SNE algorithm [24]. We illustrate embeddings for (A) the source data (simulated, SHREC'21) and for (B) the target data (real, *Chlamydomonas Reinhardtii* cells). (C,D) For each data set we applied k-means clustering in conjunction with the representation, and computed the V-measure, homogeneity and completeness with respect to the number of clusters (the parameter of k-means). All results were obtained on the test data.
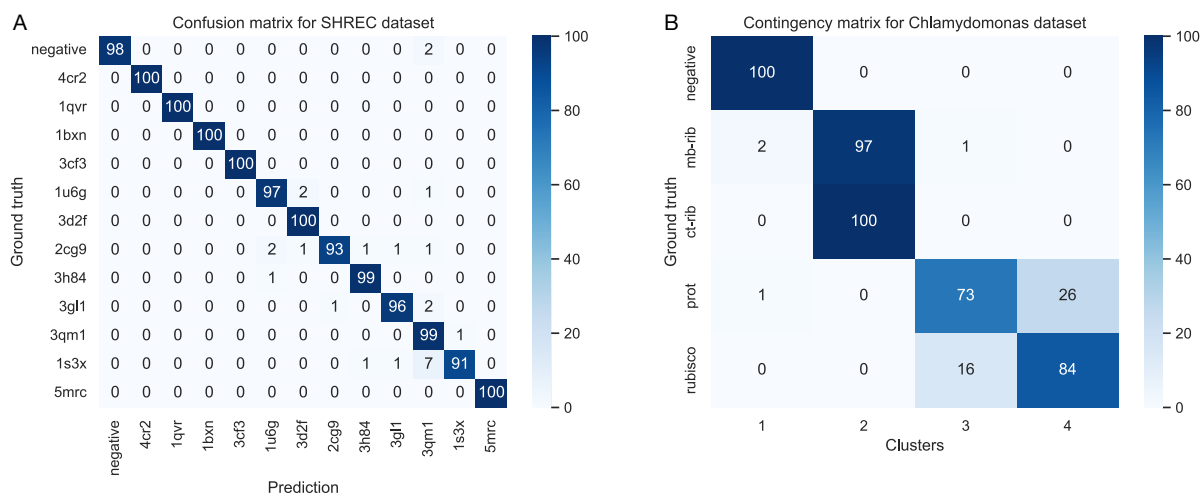
Figure 3: (A) Confusion matrix of the classification (supervised) on the SHREC dataset. (B) Contigency matrix of the k-means clustering (unsupervised) on the Chlamydomonas dataset. Displayed values are normalized and correspond to percentages.

Data availability

The synthetic dataset is available on the website of the SHREC 2021 challenge (http://www2.projects.science.uu.nl/shrec/c et/). Tomograms from the real dataset of *C. reinhardtii* cells can be found in the EMDB under accession numbers EMD-3967 and EMD-12749.

References

[1] C. Che, R. Lin, X. Zeng, K. Elmaaroufi, J. Galeotti, and M. Xu, "Improved deep learning based macromolecules structure classification from electron cryo tomograms," *Mach. Vis. Appl.*, pp. 1227–1236, 2018.

[2] E. Moebel, A. Martinez-sanchez, L. Lamm, R. D. Righetto, W. Wietrzynski, S. Albert, D. Larivière, E. Fourmentin, S. Pfeffer, J. Ortiz, W. Baumeister, T. Peng, B. D. Engel, and C. Kervrann, "Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms," *Nat. Methods*, vol. 18, pp. 1386–1394, 2021.

[3] F. Förster and R. Hegerl, "Structure determination In Situ by averaging of tomograms," in *Cell. Electron Microsc.*, vol. 79, pp. 741–767, 2007.

[4] Z. Yu and A. S. Frangakis, "Classification of electron sub-tomograms with neural networks and its application to template-matching," *J. Struct. Biol.*, vol. 174, no. 3, pp. 494–504, 2011.
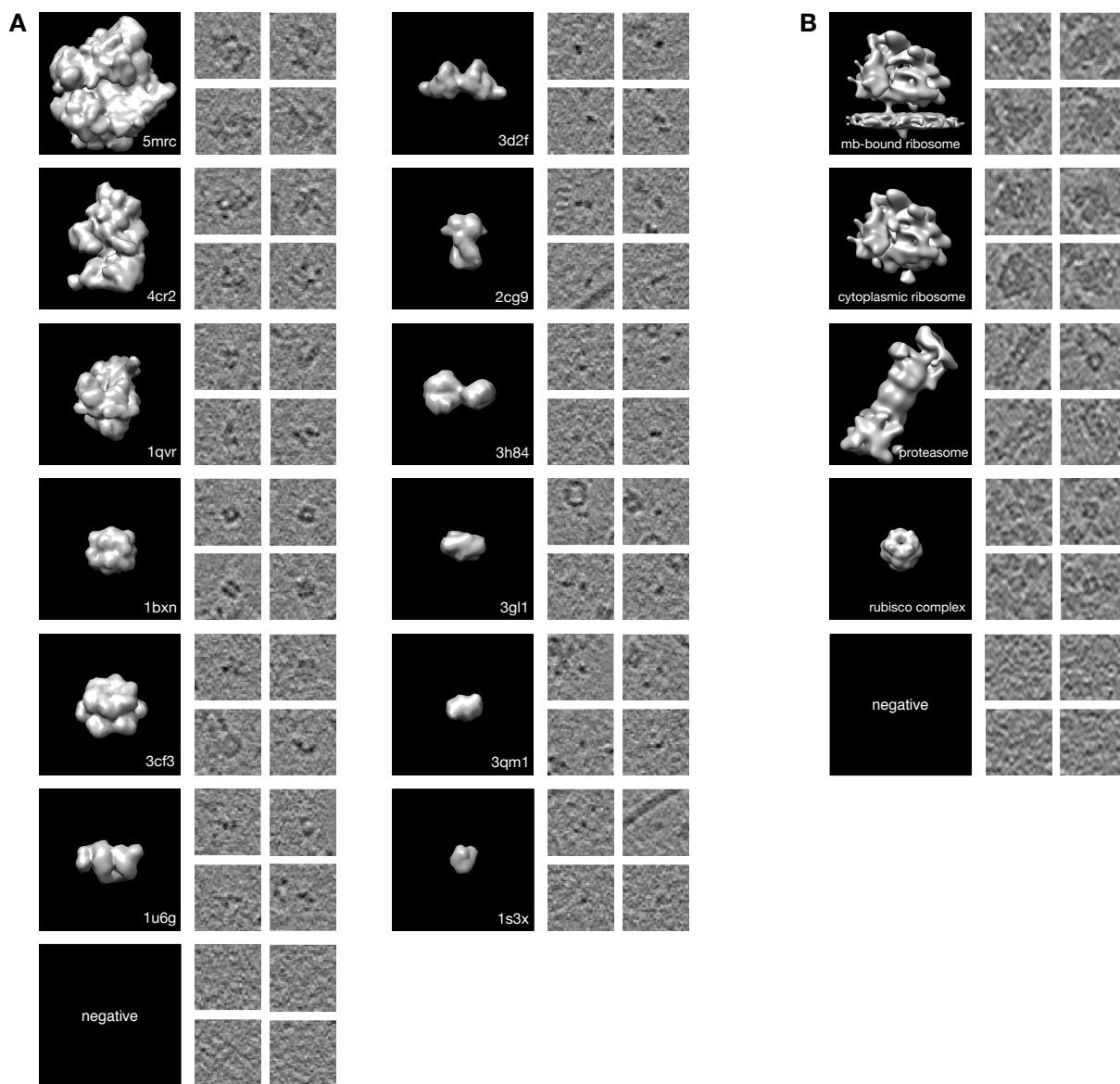
Figure 4: **Dataset illustrations** (A) Synthetic dataset SHREC'21 and (B) Real dataset (Chlamydomonas Reinhardtii). For both datasets, we display the macromolecule classes. For each class, on the left side a 3D view of the pdb structure, and on the right four instances of this class as found in the dataset. The instance images correspond to the central xy slice of the 3D volume. For enhancing the contrast, the slice has been averaged with the two neighboring slices in each direction.

[5] Y. Chen, S. Pfeffer, J. J. Fernández, C. O. S. Sorzano, and F. Förster, "Autofocused 3D Classification of Cryoelectron Subtomograms," *Structure*, pp. 1528–1537, 2014.

[6] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Work. Unsupervised Transf. Learn.*, pp. 37–50, 2012.

[7] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: feature learning by inpainting," in *Comput. Vis. Pattern Recognit.*, 2016.

[8] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *ECCV*, vol. 1, pp. 577–593, 2016.

[9] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning : a review and new perspectives," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 1798–1828, IEEE, 2013.

[11] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, IEEE, 2009.

[13] D. George, H. Shen, and E. A. Huerta, "Classification and unsupervised clustering of LIGO data with deep transfer learning," *Phys. Rev. D*, vol. 97, no. 10, 2018.

[14] R. Karim, O. Beyan, A. Zappa, I. G. Costa, D. Rebholz-schuhmann, M. Cochez, and S. Decker, "Deep learning-based clustering approaches for bioinformatics," *Brief. Bioinform.*, vol. 22, no. 1, pp. 393–415, 2021.

[15] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning : a framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[16] E. Aljalbout, V. Golkov, Y. Siddiqui, M. Strobel, and D. Cremers, "Clustering with deep learning: taxonomy and new methods," *CoRR*, vol. abs/1801.0, pp. 1–12, 2018.

[17] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *ECCV*, 2018.

[18] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," in *ICLR*, pp. 1–22, 2020.

[19] I. Gubins, M. L. Chaillet, G. V. D. Schot, M. C. Trueba, R. C. Veltkamp, F. Förster, X. Wang, D. Kihara, E. Moebel, N. P. Nguyen, T. White, F. Bunyak, G. Papoulias, S. Gerolymatos, E. I. Zacharaki, K. Moustakas, X. Zeng, S. Liu, M. Xu, Y. Wang, C. Chen, and X. Cui, "SHREC 2021 : Classification in cryo-electron tomograms," in *Eurographics Work. 3D Object Retr.*, 2021.

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

[21] S. Albert, M. Schaffer, F. Beck, S. Mosalaganti, S. Asano, H. F. Thomas, J. M. Plitzko, M. Beck, W. Baumeister, and B. D. Engel, "Proteasomes tether to two distinct sites at the nuclear pore complex," in *Proc. Natl. Acad. Sci.*, vol. 114, p. 201716305, 2017.

[22] E. S. F. Rosenzweig, B. Xu, L. K. Cuellar, B. D. Engel, L. C. M. Mackinder, M. C. Jonikas, E. S. F. Rosenzweig, B. Xu, L. K. Cuellar, A. Martinez-sanchez, M. Schaffer, N. S. Wingreen, M. Plitzko, and F. Fo, "The Eukaryotic CO2-Concentrating Organelle Is Liquid-like and Exhibits Dynamic Reorganization," *Cell*, vol. 171, pp. 148–162, 2017.

[23] S. He, H.-t. Chou, D. Matthies, T. Wunder, M. T. Meyer, N. Atkinson, A. Martinez-sanchez, P. D. Jeffrey, S. A. Port, W. Patena, G. He, V. K. Chen, F. M. Hughson, A. J. Mccormick, O. Mueller-cajar, B. D. Engel, Z. Yu, and M. C. Jonikas, "The structural basis of Rubisco phase separation in the pyrenoid," *Nat. Plants*, vol. 6, no. 12, pp. 1480–1490, 2020.

[24] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[25] L. Mcinnes, J. Healy, and J. Melville, "UMAP : Uniform manifold approximation and projection for dimension reduction," *arXiv*, vol. 1802.03426, 2020.

[26] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. L*, vol. 17, pp. 1–35, 2016.

[27] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, pp. 766–785, IEEE, 2021.

[28] R. Bermudez-Chacon, P. Marquez-Neila, and P. Fua, "A domain-adaptive two-stream U-net for electron microscopy image segmentation," in *ISBI*, pp. 400–404, 2018.

[29] J. Peng, J. Yi, and Z. Yuan, "Unsupervised mitochondria segmentation in EM images via domain adaptive multi-task learning," *J. Sel. Top. Signal Process.*, vol. 14, no. 6, pp. 1199–1209, 2020.