MEAN SHIFT CLUSTERING AS A LOSS FUNCTION FOR ACCURATE AND SEGMENTATION-AWARE LOCALIZATION OF MACROMOLECULES IN CRYO-ELECTRON TOMOGRAPHY

Lorenz Lamm[1,2,3], *Ricardo D. Righetto*[2], *Tingying Peng*[1,3]

[1] Helmholtz Munich – German Research Center for Environment and Health, Munich, Germany
[2]Biozentrum, University of Basel, Basel, Switzerland
[3]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany

ABSTRACT

Cryo-electron tomography allows us to visualize and analyze the native cellular environment on a molecular level in 3D. To reliably study structures and interactions of proteins, they need to be accurately localized. Recent detection methods train a segmentation network and use post-processing to determine protein locations, often leading to inaccurate and inconsistent locations.

We present an end-to-end learning approach for more accurate protein center identification by introducing a differentiable, scoremap-guided Mean Shift clustering implementation. To make training computationally feasible, we sample random cluster points instead of processing the entire image.

We show that our Mean Shift loss leads to more accurate cluster center positions compared to the classical Dice loss. When combining these loss functions, we can enhance 3D protein shape preservation and improve clustering with more accurate, localization-focused score maps. In addition to improved protein localization, our method provides more efficient training with sparse ground truth annotations, due to our point sampling strategy.

Index Terms— Mean Shift clustering, Cryo-electron tomography, protein localization, protein segmentation, end-to-end learning

1. INTRODUCTION

Cryo-electron tomography (Cryo-ET) is a promising imaging technique [1] that enables the study of cells in their native environment and in three dimensions. This innovative approach significantly advances our understanding of protein interactions in their native environment. A notable application is the determination of protein structures through subtomogram averaging (STA) [2], where small volumes are extracted around center positions of proteins within the tomogram, aligned, and then averaged to generate a high-resolution structure.

For STA, it is important to detect as many instances of the same protein as possible. These proteins must be precisely located to make STA efficient or even feasible. Therefore, determining initial center points as close as possible to the true protein centers is a critical step for the successful and efficient reconstruction of protein structures from native cells.

Classically, template matching [3] has been used for localizing proteins in Cryo-ET, and is still often used due to the lack of large public annotated datasets in Cryo-ET that could be used for training neural networks. The few available datasets often do not contain complete annotations and miss several true proteins [4]. Recently, template matching has been outperformed by new deep learning-based approaches in several cases [5, 6, 4, 7, 8]. Many of these methods [4, 7, 8] first train a neural network to segment protein shapes, and then use Mean Shift clustering [9] to extract protein center locations. Since the training is thus not focused on protein localization, resulting cluster centers may be inaccurate. Besides, these approaches require the often cumbersome generation of target maps depicting the protein shapes and do not directly utilize the protein center positions given by frameworks like template matching.

Mean shift clustering has been used for several deep learning tasks, including image segmentation [10, 11] and self-supervised learning [12]. However, to our knowledge, it has not been proposed as a loss function for object center location, due to its non-differentiable nature.

We propose to integrate Mean Shift clustering into our network for end-to-end optimization of protein center locations. We introduce a score-weighted, differentiable Mean Shift module and attach it to a U-Net [13] architecture, enabling training with just protein center coordinates or combined with traditional segmentation loss. We show in multiple Cryo-ET datasets that this leads to more precise protein center locations, particularly in the case of non-spherical protein shapes. Furthermore, we show that our loss function yields good results even with incomplete ground truth annotations.

The code accompanying Mean Shift loss function and generating toy data can be accessed here via GitHub.

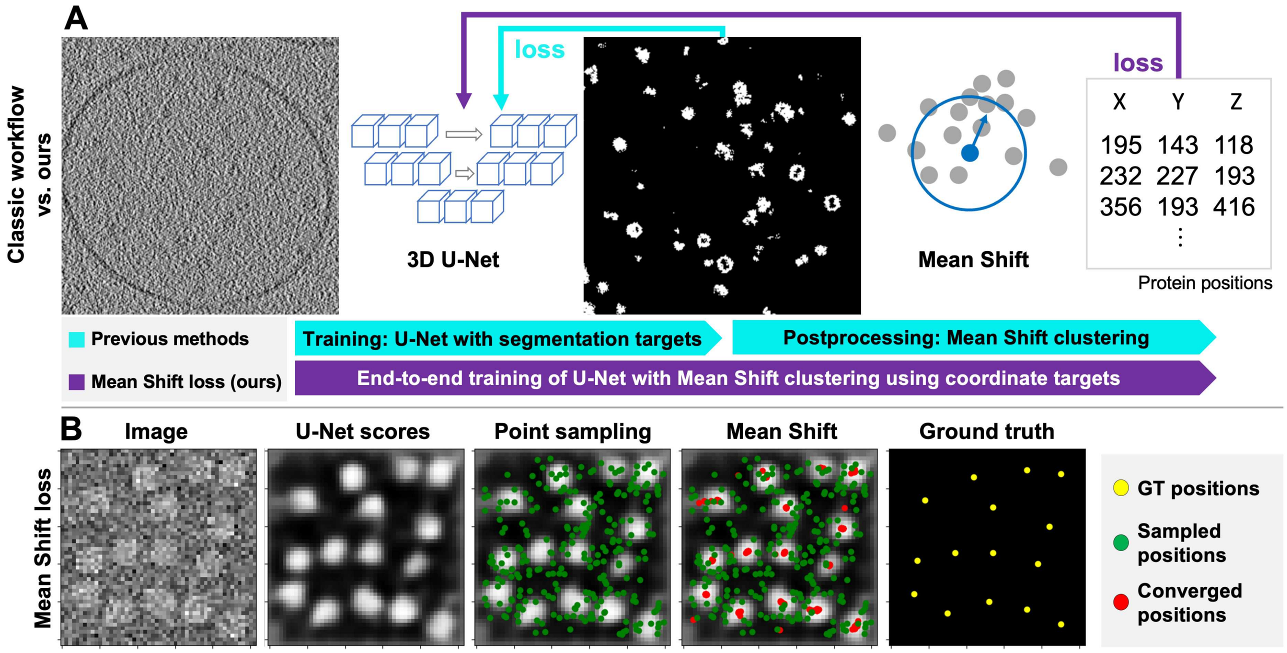

Fig. 1. Mean Shift clustering as a loss function: **A:** Existing methods (e.g., [4, 8]) train a 3D U-Net to segment protein shapes. Subsequently, Mean Shift clustering gives protein center positions. Our Mean Shift loss allows to train this process end-to-end. **B:** Implementation of our Mean Shift loss: After computing U-Net score maps, random positions (green) are sampled around GT centers. Our differentiable Mean Shift clustering gives converged positions (red) from the sampled coordinates. These can be compared to GT positions (yellow) to compute a loss value.

2. METHODS

As shown in Figure 1A, existing approaches first train a U-Net [13] to segment protein shapes, and then use Mean Shift clustering [9] as post-processing to extract protein center locations. We propose to train this workflow end-to-end by incorporating our differentiable variant of Mean Shift clustering into the architecture, enabling us to directly utilize ground truth (GT) protein center positions instead of (or in combination with) segmentation masks.

2.1. Mean Shift clustering

Mean Shift clustering [9] is a clustering technique that iteratively shifts data points towards the densest part of a dataset. It is often used when the exact number of expected cluster centers is unknown, as its only adjustable parameter is the *bandwidth* b. The clustering processes each point separately by iteratively updating the point's position by the weighted average of all points within radius b of the current point. All points thus converge to locally dense point regions.

2.2. Our differentiable Mean Shift clustering

Similar to [14], we implement Mean Shift using PyTorch on GPU in a batch-wise fashion. During inference, other methods [4, 8] perform thresholding of scoremaps to yield initial cluster coordinates. Compared to that, during training, we randomly sample points within a certain radius around ground truth (GT) locations from the voxel grid (see Figure 1B). Then, we weight the sampled positions using the network-assigned scores of the corresponding voxels. Using these weights, we perform our score-guided Mean Shift clustering by iteratively computing the weighted averages of all points within bandwidth b of a sampled point p. For further efficiency, we limit the maximum of iterations to a low number (10 in all experiments). Algorithm 1 describes our score-weighted Mean Shift clustering in more detail.

Algorithm 1 Our differentiable Mean Shift loss

```
Inputs: image, bandwidth $b$
Returns: Mean Shift loss
Predict score map $\text{scores}_{\text{U-Net}}$ for image
Sample point coordinates $p$ around ground truth positions
for $p_{\text{pred}}$ in sampled points do
    for $i = 1$ to $\text{iter}_{\max}$ do
        Find points $q$ within radius $b$: $\{q \mid \|p_{\text{pred}} - q\| < b\}$
        Compute weights: $w_q = \text{scores}_{\text{U-Net}}(q) \cdot \frac{\|p_{\text{pred}} - q\|}{b}$
        Update position $p_{\text{pred}} = \frac{1}{\sum_q w_q} \sum_q w_q \cdot q$
loss = $\text{MSE}(p_{\text{pred}}, P_{\text{GT}}) + \text{MSE}(p_{\text{GT}}, P_{\text{pred}})$
```

The introduction of network score-based weighting and random sampling of points without thresholding allows the backpropagation through the Mean Shift module and thus enables us to define our Mean Shift loss function. The advantages of sampling only a few positions (in practice, we use 256) around GT positions are twofold: First, it leads to a much more efficient clustering performance than processing all coordinates of the 3D patch. Second, this sampling allows us to focus our training on regions with available annotations: If a true protein position is not captured by the GT, we will not sample points close to this position and thus not severely distort the training process. Compared to that, classic segmentation metrics like Dice loss will be influenced strongly by to the false negative GT annotations.

After convergence, we have a set of predicted points $p_{\text{pred}} \in P_{\text{pred}}$, and we use Mean Squared Error to compare to the GT positions $p_{\text{GT}} \in P_{\text{GT}}$:

$$\text{MSE}(p_{\text{pred}}, P_{\text{GT}}) = \min_{p_{\text{GT}} \in P_{\text{GT}}} \|p_{\text{pred}} - p_{\text{GT}}\|_2^2 \qquad (1)$$

$$\text{MSE}(p_{\text{GT}}, P_{\text{pred}}) = \min_{p_{\text{pred}} \in P_{\text{pred}}} \|p_{\text{GT}} - p_{\text{pred}}\|_2^2 \qquad (2)$$

This ensures that GT positions are close to a predicted position, while prediction positions are close to a GT position.

2.3. Evaluation metrics

For evaluation, we define the average distances of predicted positions to their closest ground truth position and vice versa:

$$\text{dist}_{\text{pred}} = \frac{1}{|P|} \sum_{p \in P} \min_{g \in GT} \|p - g\|_2, \qquad (3)$$

$$\text{dist}_{\text{GT}} = \frac{1}{|GT|} \sum_{g \in GT} \min_{p \in P} \|p - g\|_2, \qquad (4)$$

where P and GT are the sets of all predicted and GT positions, respectively. We also show the F_1-score using different *hit*-radii: A GT position is counted as true positive (TP) with *hit*-radius 4 if a predicted position is within a radius of 4, and vice versa. Together with false positives (FP) and false negatives (FN), we compute precision, recall, and F_1-score:

$$\text{Prec} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP}}, \qquad \text{Rec} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}}, \qquad (5)$$

$$F_1 = \frac{2 * \text{Rec} * \text{Prec}}{\text{Rec} + \text{Prec}}. \qquad (6)$$

3. RESULTS

3.1. Datasets

We collected several datasets to benchmark our Mean Shift loss with the commonly used Dice loss. Figure 2 shows sample images of each dataset. As a proof of concept, we generated the *Sandclock* toy dataset by placing two spheres in opposite directions of a randomly drawn center point. Next, we used the synthetic *Shrec* Challenge Cryo-ET dataset [15], depicting proteins of different sizes in realistically simulated tomograms. Finally, we evaluated our approach using the *Ribo* dataset: a tomogram from an experimental dataset (EMPIAR-10045 [16]) containing 3D locations of ribosomes.

For the *Shrec* dataset, we sampled training (1558), validation (426) and test (440) patches that contained proteins from different tomograms. For the *Ribo* dataset, we generated non-overlapping patches containing at least one protein, and split them into 50 training, 11 validation, and 11 test patches. For all datasets, we used a 3D patch size of 56^3 both during training and evaluation.

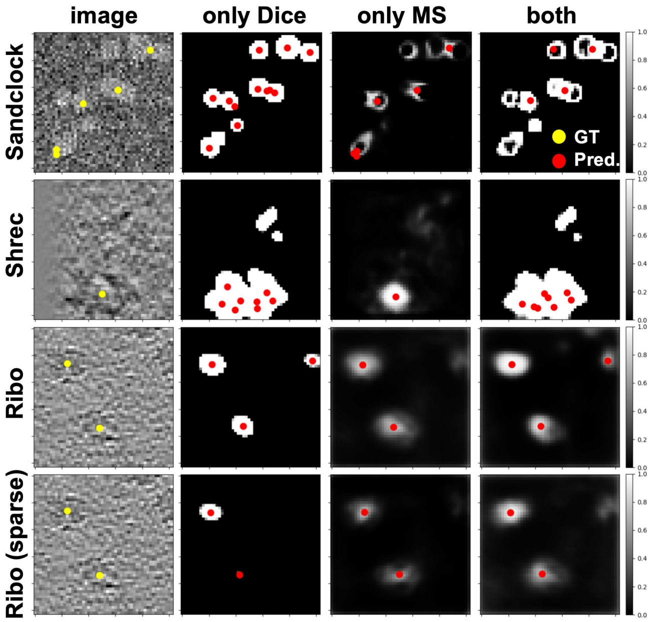

Fig. 2. Images and predictions: 2D slices of the 3D patches of the *Sandclock*, *Shrec*, *Ribo*, and *Ribo (Sparse)* experiments. Shown are the raw input together with ground truth positions (yellow), as well as score maps for the experiments using only Dice as a loss functions, only our Mean Shift loss, or a combination. Predicted cluster centers are highlighted in red.

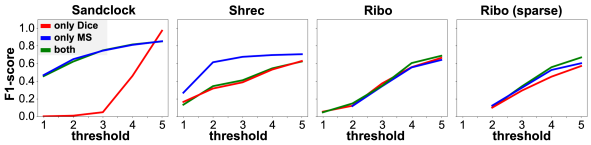

Fig. 3. F1 scores for different *hit*-thresholds for all datasets and our three training settings using only Dice loss (red), only Mean Shift loss (blue), and a combination of both (green).

3.2. Experimental evaluation

For each dataset, we performed training runs with Dice loss, Mean Shift loss, and their combination (Table 1, Figure 2),

Experiment	dist$_{\text{pred}}$	dist$_{\text{GT}}$	F1$_{\text{rad4}}$
Sandclock$_{\text{Dice}}$	4.02 ± 0.01	3.82 ± 0.01	0.46 ± 0.01
Sandclock$_{\text{MS}}$	**2.43** ± 0.17	1.64 ± 0.24	**0.81** ± 0.03
Sandclock$_{\text{both}}$	2.62 ± 0.06	**1.26** ± 0.08	**0.82** ± 0.01
Shrec$_{\text{Dice}}$	**6.21** ± 0.29	2.09 ± 0.16	0.53 ± 0.02
Shrec$_{\text{MS}}$	7.75 ± 4.70	3.99 ± 1.70	**0.70** ± 0.10
Shrec$_{\text{both}}$	7.22 ± 0.42	**1.69** ± 0.15	0.55 ± 0.01
Ribo$_{\text{Dice}}$	8.58 ± 0.35	4.03 ± 0.04	0.56 ± 0.02
Ribo$_{\text{MS}}$	8.97 ± 0.38	4.82 ± 0.85	0.55 ± 0.05
Ribo$_{\text{both}}$	8.44 ± 0.23	**3.54** ± 0.10	**0.61** ± 0.02
Ribo(sparse)$_{\text{Dice}}$	8.24 ± 0.68	8.65 ± 3.06	0.45 ± 0.10
Ribo(sparse)$_{\text{MS}}$	8.41 ± 0.44	6.46 ± 1.54	**0.53** ± 0.05
Ribo(sparse)$_{\text{both}}$	8.87 ± 0.49	**4.13** ± 0.47	**0.55** ± 0.05

Table 1. Results for different experiments. For each setting (*Sandclock*, *Shrec*, *Ribo*, *Ribo(sparse)*), we trained models using only Dice loss ($_{\text{Dice}}$), only our Mean Shift loss ($_{\text{MS}}$), and a combination of both ($_{\text{both}}$). We show the means and standard deviations (5 training runs) of each predicted position's distance to the closest ground truth position (dist$_{\text{pred}}$), and vice versa (dist$_{\text{GT}}$), as well as the F1-score with a *hit*-radius of 4.

selecting the best model from 1000 epochs based on validation loss. We used a constant learning rate of 10^{-5} without weight decay or other regularization, and a bandwidth of 4 for Mean Shift clustering, tuned on the *Shrec* validation set.

For the *Sandclock* dataset, we observe lower average distance values (dist$_{\text{pred}}$), dist$_{\text{GT}}$)) as well as better F1-scores when training with our Mean Shift loss or a combination. Figure 2 shows that while Dice loss offers more precise segmentations, it falls short in accurate cluster center identification. Conversely, our Mean Shift loss produces score maps that lead to more precise clustering and, consequently, more accurate protein center localization.

For the *Shrec* dataset, we observe mixed results: Dice loss or the combination show lower distance scores, but Mean Shift alone achieves the highest F_1-score. The score maps from Dice loss training (Figure 2) more accurately predict protein shapes, but *Shrec*'s varying protein sizes lead to ambiguous cluster centers and potential protein oversampling, as uniform bandwidth clustering struggles with size variability. Conversely, Mean Shift loss generates score maps better suited for precise cluster center prediction, reflected in higher F_1-scores. However, upon close inspection, we observed some significantly deviant outlier cluster centers, impacting distance values, as evident from the high standard deviations. The F_1-score plot in Figure 3 further supports this, with Mean Shift loss achieving higher scores already at lower *hit*-thresholds, indicating overall accuracy despite outliers.

For the experimental dataset *Ribo*, we performed two experiments: First, we generated spherical masks around all ground truth positions and trained again using Dice loss, Mean Shift loss, and their combination. Here, we observe slightly improved distance scores and F_1-scores when using the combined loss compared to only Dice loss. However, due to the roughly globular shape of the ribosomes, the advantage of using Mean Shift loss is not fully given.

Our second experiment *Ribo (sparse)* highlights our loss function's ability to deal with sparse annotations: During training, we only used a single GT position and corresponding mask per patch to optimize our network. For the test set, we considered all GT positions again. While the performance using only Dice loss decreases notably (in particular dist$_{\text{GT}}$, indicating many missed GT positions), training with the combined loss maintained similar results to full annotation training. This underscores our loss function's capability to handle the sparse annotations common in Cryo-ET, where accurately localizing all proteins is often challenging.

4. CONCLUSION

To improve the accuracy of recent protein localization programs, we introduced a Mean Shift loss function that allows end-to-end training of a segmentation task with subsequent clustering. In order to use the originally non-differentiable Mean Shift clustering for training, we introduced a network-score-based weighting to the clustering and implemented a point sampling scheme around GT positions to make the clustering computationally feasible. Using this Mean Shift loss, we can avoid tediously generating a segmentation target map and utilize ground truth locations directly.

We showed that, particularly for non-globular protein shapes, our loss function learns score maps that are more tailored towards a precise localization, compared to previous workflows with two separated steps to segment protein shapes and then perform independent clustering. Our point sampling strategy in the Mean Shift loss computation enhances robustness against sparsely annotated protein locations, a frequent issue in Cryo-ET.

In follow-up work, we would like to extend our loss function to a multi-class setting, and evaluate the benefits of the Mean Shift loss function on more experimental datasets (more diverse, non-globular protein shapes) and in more detail, e.g., by showing the effects of more accurate protein positions on downstream tasks like subtomogram averaging.

5. ACKNOWLEDGMENTS

L.L. acknowledges support from the Munich School for Data Science (MUDS) and a fellowship from the Boehringer Ingelheim Fonds. Calculations were performed at sciCORE (https://scicore.unibas.ch/) scientific computing center at the University of Basel. The authors acknowledge Ben Engel (University of Basel) for scientific discussions and access to computational resources.

6. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required.

7. REFERENCES

[1] Martin Turk and Wolfgang Baumeister, "The promise and the challenges of cryo-electron tomography," *FEBS letters*, vol. 594, no. 20, pp. 3243–3261, 2020.

[2] W Wan and John AG Briggs, "Cryo-electron tomography and subtomogram averaging," *Methods in enzymology*, vol. 579, pp. 329–367, 2016.

[3] Thomas Hrabe, Yuxiang Chen, Stefan Pfeffer, Luis Kuhn Cuellar, Ann-Victoria Mangold, and Friedrich Förster, "Pytom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis," *Journal of structural biology*, vol. 178, no. 2, pp. 177–188, 2012.

[4] Emmanuel Moebel, Antonio Martinez-Sanchez, Lorenz Lamm, Ricardo D Righetto, Wojciech Wietrzynski, Sahradha Albert, Damien Larivière, Eric Fourmentin, Stefan Pfeffer, Julio Ortiz, et al., "Deep learning improves macromolecule identification in 3d cellular cryo-electron tomograms," *Nature methods*, vol. 18, no. 11, pp. 1386–1394, 2021.

[5] Erik Genthe, Sean Miletic, Indira Tekkali, Rory Hennell James, Thomas C Marlovits, and Philipp Heuser, "Pickyolo: Fast deep learning particle detector for annotation of cryo electron tomograms," *Journal of Structural Biology*, p. 107990, 2023.

[6] Gavin Rice, Thorsten Wagner, Markus Stabrin, Oleg Sitsel, Daniel Prumbaum, and Stefan Raunser, "Tomotwin: generalized 3d localization of macromolecules in cryo-electron tomograms with structural data mining," *Nature Methods*, pp. 1–10, 2023.

[7] Lorenz Lamm, Ricardo D Righetto, Wojciech Wietrzynski, Matthias Pöge, Antonio Martinez-Sanchez, Tingying Peng, and Benjamin D Engel, "Membrain: A deep learning-aided pipeline for detection of membrane proteins in cryo-electron tomograms," *Computer methods and programs in biomedicine*, vol. 224, pp. 106990, 2022.

[8] Yu Hao, Xiaohua Wan, Rui Yan, Zhiyong Liu, Jintao Li, Shihua Zhang, Xuefeng Cui, and Fa Zhang, "Vp-detector: A 3d multi-scale dense convolutional neural network for macromolecule localization and classification in cryo-electron tomograms," *Computer Methods and Programs in Biomedicine*, vol. 221, pp. 106871, 2022.

[9] Yizong Cheng, "Mean shift, mode seeking, and clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[10] Yangxiao Lu, Yuqiao Chen, Nicholas Ruozzi, and Yu Xiang, "Mean shift mask transformer for unseen object instance segmentation," *arXiv preprint arXiv:2211.11679*, 2022.

[11] Boonnatee Sakboonyara and Pinyo Taeprasartsit, "U-net and mean-shift histogram for efficient liver segmentation from ct images," in *2019 11th International Conference on Knowledge and Smart Technology (KST)*, 2019, pp. 51–56.

[12] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash, "Mean shift for self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10326–10335.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[14] Mengyang Zhao, Aadarsh Jha, Quan Liu, Bryan A Millis, Anita Mahadevan-Jansen, Le Lu, Bennett A Landman, Matthew J Tyska, and Yuankai Huo, "Faster mean-shift: Gpu-accelerated clustering for cosine embedding-based cell segmentation and tracking," *Medical Image Analysis*, vol. 71, pp. 102048, 2021.

[15] Ilja Gubins, Marten L. Chaillet, Gijs van der Schot, M. Cristina Trueba, Remco C. Veltkamp, Friedrich Förster, Xiao Wang, Daisuke Kihara, Emmanuel Moebel, Nguyen P. Nguyen, Tommi White, Filiz Bunyak, Giorgos Papoulias, Stavros Gerolymatos, Evangelia I. Zacharaki, Konstantinos Moustakas, Xiangrui Zeng, Sinuo Liu, Min Xu, Yaoyu Wang, Cheng Chen, Xuefeng Cui, and Fa Zhang, "SHREC 2021: Classification in Cryo-electron Tomograms," in *Eurographics Workshop on 3D Object Retrieval*, Silvia Biasotti, Roberto M. Dyke, Yukun Lai, Paul L. Rosin, and Remco C. Veltkamp, Eds. 2021, The Eurographics Association.

[16] Tanmay AM Bharat and Sjors HW Scheres, "Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in relion," *Nature protocols*, vol. 11, no. 11, pp. 2054–2065, 2016.