2816
doi:10.1017/S1431927621009843
Microsc. Microanal. 27 (Suppl 1), 2021

Explore the complexity of proteins with an expanded CryoET data processing pipeline

Muyuan Chen[1], David Chmielewski[2], Wah Chiu[3] and Steven Ludtke[1]

[1]Baylor College of Medicine, United States, [2]Stanford University, United States, [3]School of Medicine, Stanford University, Stanford, California, United States

Cryo-electron tomography (CryoET) provides a direct way to image cells in 3D and visualize individual proteins within them at the native condition. From the tomograms, electron densities of proteins can be identified, and multiple copies of the same protein can be aligned and averaged to provide structural information of the protein at near-atomic resolution. With modern instruments, massive amounts of data can be produced rapidly by electron microscopes, and the development of automatic data processing methods has become the key that turns the raw microscopy data into valuable biological findings.

In 2019, we developed an integrated workflow that automates the processes from the alignment of raw tilt series to the reconstruction of high-resolution protein structures. Here, we present further development that utilizes machine learning methods and expands the capability of the data processing pipeline. The new methods address three critical problems in CryoET, and make it possible to further explore the diversity and complexity of macromolecules within their cellular environment.

First, a deep learning based algorithm is introduced to intelligently compensate for the missing wedge of raw tomograms. The program can compile a set of features from multiple tomograms of the same sample, then use these features to fill the correct information into the missing wedge of each tomogram. As an unsupervised machine learning model, the program can run on the entire dataset as a post-processing step without human intervention, and boost the interpretability of the tomograms.

Second, a graphical user interface with a built-in deep neural network model is developed for semi-automated particle selection from tomograms. Users can start by selecting only a few instances of the particle of interest, and the neural network will look for similar features in the entire dataset. Then, using the interface, users can interactively update the training set based on the results of the neural network, and the network will refine the particle selection accordingly. By iteratively refining the training set and particle selection, the tool can be used to find proteins of low abundance from the crowded cellular environment.

Finally, we present tools for solving protein structures of highly dynamic systems from CryoET datasets. This includes classical methods such as reference based classification and multi-body refinement, as well as a deep learning based algorithm that learns the conformation landscape of the macromolecular system from the subtomograms. Those tools will allow researchers to explore the complexity of proteins inside their cellular environment and study how they coordinate together to keep the cell functioning.

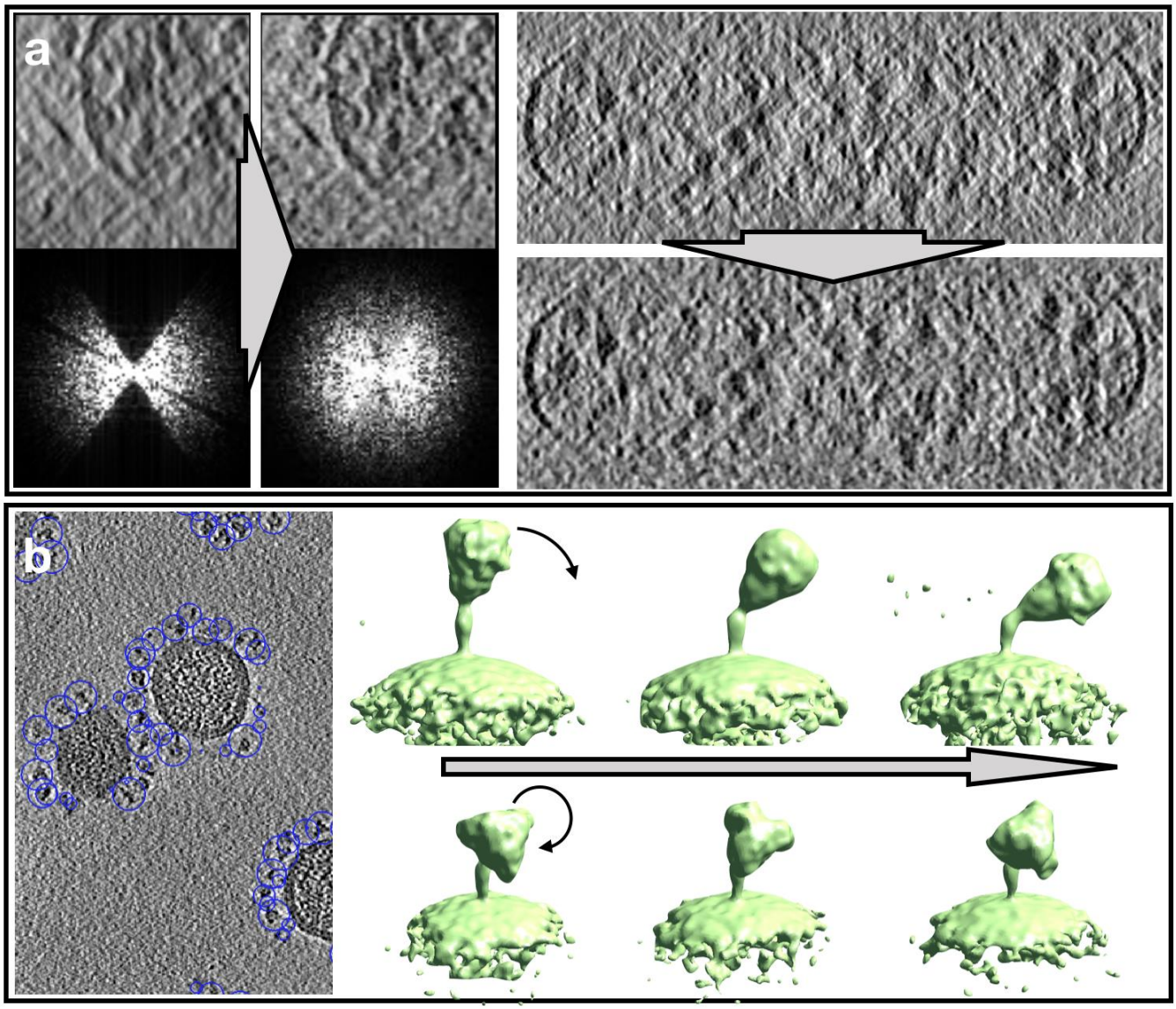

Figure 1. (a) Left: 2D training set patches and their Fourier Transform before and after the missing wedge compensation neural network is applied. Right: side view of a tomogram (EMPIAR-10499) before and after the missing wedge filling. (b) Left: automatic particle selection of coronavirus NL63 spike particles. Right: two modes of spike motion with respect to the membrane extracted from the particles.

References

[1] M. Chen, S. J. Ludtke (2021) Deep learning based mixed-dimensional GMM for characterizing variability in CryoEM. arXiv:2101.10356

[2] M. Chen, J. M. Bell, X. Shi, S. Y. Sun, Z. Wang, S. J. Ludtke (2019). A complete data processing workflow for CryoET and subtomogram averaging. *Nature Method.* 2019.

https://doi.org/10.1017/S1431927621009843 Published online by Cambridge University Press