Preparation Of Labeled Cryo-ET Datasets For Training And Evaluation Of Machine Learning Models

Aygul Ishemgulova, Alex J. Noble, Tristan Bepler, Alex De Marco
New York Structural Biology Center
aishemgulova@nysbc.org, anoble@nysbc.org, tbepler@nysbc.org, alex@nysbc.org

Abstract

We present datasets aimed at improving the efficiency of cryo-electron tomographic data analysis. While cryo-electron tomography (cryo-ET) holds immense promise as a tool for native structural biology, it faces persistent challenges in segmentation and annotation. These challenges primarily stem from the absence of diverse ground truth datasets for efficient model training, evaluation, and benchmarking. To address these challenges, we have collected and are currently annotating datasets spanning a range of complexities. Composed of carefully selected protein mixtures and organisms with small genomes, these datasets offer a broad spectrum of structures for study. The datasets are designed to provide a robust foundation for development and evaluation of machine learning models for annotation tasks, thereby enhancing the efficacy and applicability of cryo-ET in elucidating complex native biological structures and interactions. This ongoing project will soon offer the annotated datasets publicly, encouraging further innovation and research in the community.

1 Introduction

Cryo-electron tomography (cryo-ET) has emerged as a groundbreaking method in structural biology, offering a unique lens into the intricate world of proteins, lipids, sugars, nucleic acids, and organelles built of these components within their natively preserved cellular and tissue environments at sub-nanometer resolutions (1). Cryo-ET has the potential to reveal biomedically relevant insights into protein structures and interactions, such as disease- and drug-altered cellular morphologies within the cellular milieu (2,3).

In recent years, there has been major progress in cryo-ET sample preparation, namely in vitrification (4), grid micropatterning (5), cryo-focused ion beam (cryo-FIB) milling (6), waffle milling (7), and lift-out methods (8). Advancements in data collection and processing software packages, including PACE (9), Warp/M (10), AreTomo (11), and RELION (12), have likewise been significant. These advances currently allow for hundreds of cryo-ET tomograms encompassing dozens of cells and

totaling several terabytes of data to be produced per day on a high-end microscope. However, the realm of segmentation and annotation remains the Achilles' heel of the workflow. Current approaches range from manual segmentation methods with software such as Amira (Thermo Fisher Scientific), to template matching with Dynamo (13), EmClarity (14), or EMAN2 (15), and on to specialized deep learning strategies like crYOLO (16), EMAN2, and DeepPict (17). It is worth noting, however, that while these methods may excel when dealing with abundant, large proteins and cellular features, their efficacy across a broader range of cellular contents is inconsistent.

Public availability of cryo-ET tomograms is limited. Available datasets have no annotations or sparse annotations, primarily of ribosomes. The absence of universally accepted benchmark and ground truth datasets for model evaluation presents a roadblock for performance comparisons between already existing approaches and development of new approaches. To address this gap, our work introduces cryo-tomographic datasets comprised of diverse protein mixtures (< 10 proteins) with liposomes and a collection of the smallest known free-living organisms in terms of size and number of protein coding genes (from hundreds to several thousand genes). Additionally, we have obtained a collection of cellular tomograms collected on larger organisms including human, mice, algae, and protozoa (tens of thousands of genes) for internal development. A key feature of our curated datasets is that they range from low to high complexity, which will allow for stepwise machine learning (ML) development and benchmarking. The datasets to be released consist of tilt- series and reconstructed tomograms that were collected with current state-of-the-art methods in the cryo-ET field. Once fully annotated, these datasets will lay the foundation for training and assessment of ML models, advancing the frontiers of segmentation and annotation in cryo-ET.

2 Method

2.1 Preparation of samples

Purified apoferritin, PP7 virus-like particles, proteasome were mixed in phosphate buffered saline at concentrations: 3.2 mg/ml, 2.4 mg/ml, 0.25 mg/ml, respectively. Purified thyroglobulin, B-amylase, bovine serum albumin, tobacco mosaic virus, apoferritin, PP7 virus-like particles, proteasome and *E.coli* liposomes were mixed in phosphate buffered saline at concentrations of proteins: 10.4 mg/ml, 8.9 mg/ml, 6.32 mg/ml, 5.66 mg/ml, 1.21 mg/ml, 0.9 mg/ml, 0.09 mg/ml, respectively. Bovine serum albumin (A8531-1VL), B-amylase (A8781-1VL), and thyroglobulin (T9145-1VL) were ordered in Sigma. The remaining purified proteins were kindly provided to us. *Candidatus* Pelagibacter ubique HTCC1062 culture was grown at 20ºC as described elsewhere (18). *E. coli* K-12 strain WM3433 culture was grown at 37ºC. Minicells were produced as described earlier (19).

R 2/2 on 300 mesh Cu grids (Quantifoil) were glow discharged for 30 s with Hydrogen 6.4 sccm, Oxygen 27.5 sccm using Gatan Solarus II plasma cleaner. 3.8 µL of sample was applied to grids and plunge frozen in liquid ethane using a Vitrobot Mark IV (Thermo Fisher Scientific).

2.2 Data collection

Screening of the grids was performed on an TFS Glacios microscope at 200 keV. Tilt-series acquisition was performed on a TFS Titan Krios transmission electron microscope operated at 300 keV, equipped with a Gatan BioQuantum energy filter and a K3 direct electron detector using SerialEM software version 4.1.0 beta (20). Tilt-series were acquired at a magnification of 33,000x,

yielding a pixel size of 2.077 Å. Data were collected at tilt angles ranging from −45° to +45°, in 3° increments, following a dose-symmetric tilt scheme (21). Images for each tilt angle were recorded as movies, each consisting of 10 - 20 frames. The target total dose for tilt-series acquisition was 100 e−/Å^2.

2.3 Reconstruction of tomograms

Frames from movies, corresponding to images at each tilt-angle within the tilt-series, were motion-corrected, and assembled into stacks using Warp software (10). Tilts within each stack were aligned and tomograms were reconstructed using AreTomo software (11).

3 Results

3. 1 Tomograms of protein mixtures

We started with protein mixtures and prepared simpler (3 proteins) and more complex (7 proteins plus liposomes) mixtures of proteins. Protein sizes and shapes, symmetries, concentrations were considered during mixture composition.

3.1.1 Mixture of 3 proteins

The first dataset is comprised of large and symmetrical proteins: apoferritin, proteasome, and PP7 virus-like particles, which were chosen for their ease of identification (Fig. 1A). 46 tomograms were selected after manual inspection of the collected data.

3.1.2 Mixture of 7 proteins plus liposomes

To better approximate a cellular sample, we enriched the mixture with proteins with a roughly uniform size distribution. The resulting dataset incorporates a wide range of proteins: 500 kDa apoferritin, 750 kDa proteasome, 1 – 5 MDa PP7 virus-like particles, 66 kDa bovine serum albumin, 200 kDa B-amylase, 660 kDa thyroglobulin, tobacco mosaic virus which varies significantly in size, extending into the megadalton range (Fig. 1B). This diverse collection exhibits proteins of different symmetries: C1, C2, helical, T=3 icosahedral, D7, and octahedral. Liposomes extracted from *E.coli* were added to the mixture to mimic cellular membrane structures. Given that cellular protein concentrations approximately range from 100 to 300 mg/mL (22, 23) - a value substantially higher than the roughly 1 mg/mL typically utilized for single particle analysis in Cryo-EM (24), we prepared the protein mixture with a concentration of 36 mg/mL to more closely reflect cellular contexts. This approach yields a dataset that is not only representative of a diverse array of protein types but also with a protein concentration that is closer to what is found in cells. 75 tomograms were selected after manual inspection of the collected data.

3.2 Cellular Tomograms

For ease of protein annotation, we opted for datasets derived from organisms with small genomes. Small genomes imply fewer proteins in the tomograms, simplifying the annotation task. The cellular size was another crucial consideration in organism selection, with the focus being on unicellular organisms small enough to allow for tomogram collection without the need for cryo-FIB-milling -

a process that considerably extends sample preparation time. To this end, we chose cells whose minor axes are less than several hundred nanometers.

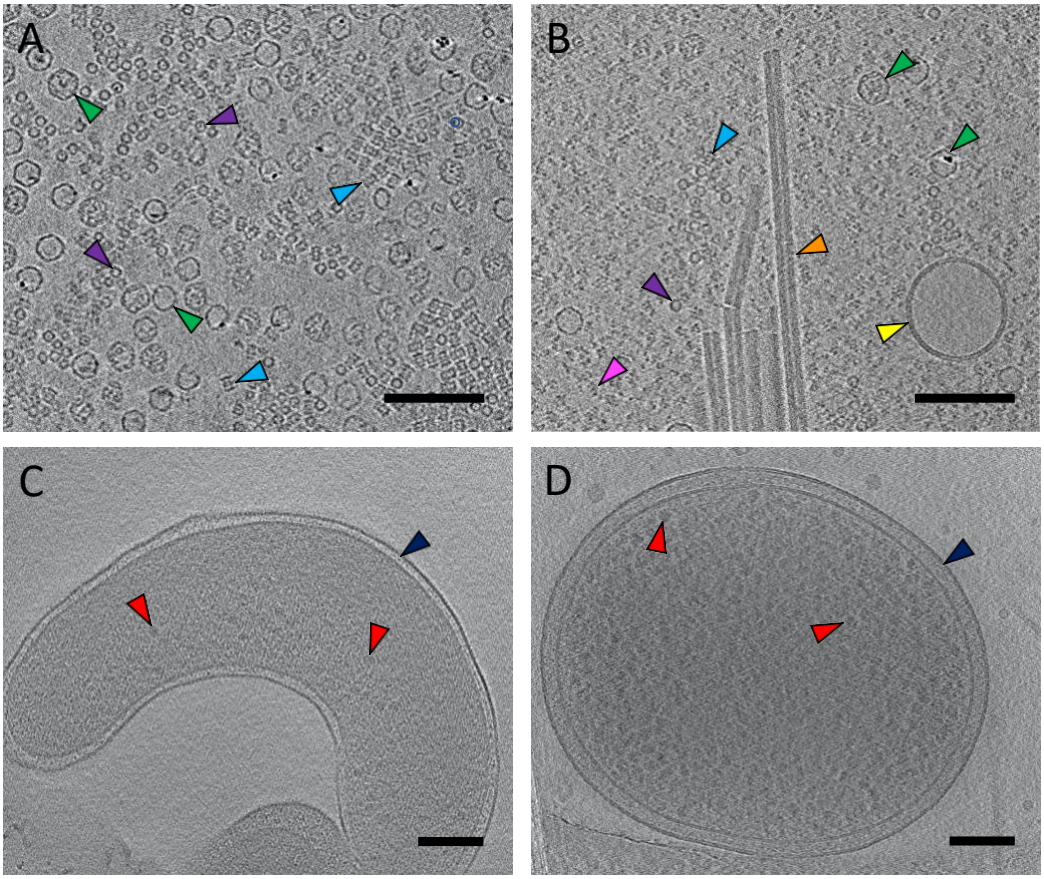

Fig.1. Cross-sections of tomograms from apoferritin, proteasome, and PP7 virus-like particles (A); from apoferritin, proteasome, PP7 virus-like particles, bovine serum albumin, B-amylase, thyroglobulin, tobacco mosaic virus, and *E. coli* liposomes (B); *Candidatus* Pelagibacter ubique HTCC1062 (C); and *E. coli* K-12 strain WM3433 minicells (D). The green arrowhead indicates a PP7 virus-like particle, purple - apoferritin, blue - proteasome, orange - tobacco mosaic virus, yellow - *E.coli* liposomes, magenta – thyroglobulin, dark blue - double membrane, red - ribosomes. The scale bars represent 100 nm.

3.2.1 *Candidatus* Pelagibacter ubique

The free-living marine bacterium, *Candidatus* Pelagibacter ubique HTCC1062 with approximately 1,400 protein-coding genes and diameters ranging from 200 to 400 nm, presented an ideal candidate. A total of 89 tomograms were selected for this organism after manual inspection of the collected data (Fig. 1C).

3.2.2 *E. coli* Minicells

While *E. coli* is a model organism, its large size and substantial genome of 4,300 to 4,400 protein-coding genes make it less ideal for cryo-ET. However, *E. coli* minicells - small cellular structures resulting from atypical cell division, devoid of chromosomal DNA and consequently chromosomal

proteins - offer a viable alternative (25). The *E. coli* K-12 strain WM3433, a mutant with increased minicells production, was chosen for our study (Fig. 1D). 65 tomograms were selected for this organism after manual inspection of the collected data.

4 Discussion and future work

In this paper, we have presented four datasets and 275 tomograms, encompassing a diverse range of protein mixtures and organisms. Additionally, one tomographic dataset of synthetically obtained JCVI-syn3A smallest living organism from Prof. John I. Glass will be added to our selection of datasets. These resources are intended to serve as benchmark datasets for the training and evaluation of ML models. Our next steps involve the identification of all proteins in cellular samples, quantification of their respective concentrations, and finding proteins they interact with through mass spectrometry. Moreover, for our internal curation and ML developments, we have obtained partially segmented cryo-ET tomograms from organisms with large genomes, including mice iBMDM cells (26), *Porphyridium purpureum* microalga (27), human SK-MEL-2 cells (28), *Toxoplasma gondii* protozoa (29), and *Cryptosporidium parvum* sporozoites (30). The models and annotated datasets resulting from this study will be crucial for enhancing the accuracy of cellular tomogram labeling. We are in the process of annotating and curating the datasets. Once this task is complete, the datasets will be made available to the public to support the development and benchmarking of ML models designed for segmentation and annotation of cryo-ET data.

Acknowledgements:

We thank Andres Cabezas for providing us with purified *E.coli* liposomes, laboratory of Prof. Stephen Giovannoni for providing us with *Candidatus* Pelagibacter ubique HTCC1062 culture, laboratory of Prof. Jun Liu for providing us with *E. coli* K-12 strain WM3433, Prof. Huilin Li for providing us with purified Mtb 20S Proteasome, Dr. Misha Kopylov for providing us with PP7 VLP, Dr. Brian Kloss for providing us with ApoF, Dr. Ruben Diaz-Avalos for providing us with tobacco mosaic virus, Prof. Yorgo Modis for providing us with iBMDM data, Prof. Sen-Fang Sui for providing us with microalga data, Prof. David Drubin for providing us with SK-MEL-2 data, Dr. Qiang Guo for providing us with protozoa data, and Prof. Yi-Wei Chang for providing us with sporozoites.

References:

1. Hong Y, Song Y, Zhang Z, Li S. Cryo-Electron Tomography: The Resolution Revolution and a Surge of In Situ Virological Discoveries. Annu Rev Biophys. 2023 May 9;52:339-360. doi: 10.1146/annurev-biophys-092022-100958. Epub 2023 Jan 31. PMID: 36719970.
2. Zhang P, Mendonca L, Howe A, Gilchrist J, Sun D, Knight M, Zanetti-Domingues L, Bateman B, Krebs AS, Chen L, Radecke J, Sheng Y, Li V, Ni T, Kounatidis I, Koronfel M, Szynkiewicz M, Harkiolaki M, Martin-Fernandez M, James W. Correlative Multi-scale Cryo-imaging Unveils SARS-CoV-2 Assembly and Egress. Res Sq [Preprint]. 2021 Jan 19:rs.3.rs-134794. doi: 10.21203/rs.3.rs-134794/v1. Update in: Nat Commun. 2021 Jul 30;12(1):4629. PMID: 33501431; PMCID: PMC7836121.
3. Madeleine A. G. Gilbert, Nayab Fatima, Joshua Jenkins, Thomas J. O'Sullivan, Andreas Schertel, Yehuda Halfon, Tjado H. J. Morrema, Mirjam Geibel, Sheena E. Radford, Jeroen J. M. Hoozemans, René A. W. Frank. In situ cryo-electron tomography of β-amyloid and

tau in post-mortem Alzheimer's disease brain. bioRxiv. 2023 July 18. doi: https://doi.org/10.1101/2023.07.17.549278.

4. Xu Y, Dang S. Recent Technical Advances in Sample Preparation for Single-Particle Cryo-EM. Front Mol Biosci. 2022 Jun 24;9:892459. doi: 10.3389/fmolb.2022.892459. PMID: 35813814; PMCID: PMC9263182.
5. Sibert BS, Kim JY, Yang JE, Wright ER. Micropatterning Transmission Electron Microscopy Grids to Direct Cell Positioning within Whole-Cell Cryo-Electron Tomography Workflows. J Vis Exp. 2021 Sep 13;(175):10.3791/62992. doi: 10.3791/62992. PMID: 34570100; PMCID: PMC8601404.
6. Wagner FR, Watanabe R, Schampers R, Singh D, Persoon H, Schaffer M, Fruhstorfer P, Plitzko J, Villa E. Preparing samples from whole cells using focused-ion-beam milling for cryo-electron tomography. Nat Protoc. 2020 Jun;15(6):2041-2070. doi: 10.1038/s41596-020-0320-x. Epub 2020 May 13. PMID: 32405053; PMCID: PMC8053421.
7. Klykov O, Bobe D, Paraan M, Johnston JD, Potter CS, Carragher B, Kopylov M, Noble AJ. In situ cryo-FIB/SEM Specimen Preparation Using the Waffle Method. Bio Protoc. 2022 Nov 5;12(21):e4544. doi: 10.21769/BioProtoc.4544. PMID: 36618877; PMCID: PMC9795037.
8. Oda H Schioetz, Christoph JO Kaiser, Sven Klumpe, Dustin R Morado, Matthias Poege, Jonathan Schneider, Florian Beck, Christopher Thompson, Juergen M Plitzko. Serial Lift-Out-Sampling the Molecular Anatomy of Whole Organisms. bioRxiv, 2023.04. 28.538734. doi.org/10.1101/2023.04.28.538734.
9. Eisenstein F, Yanagisawa H, Kashihara H, Kikkawa M, Tsukita S, Danev R. Parallel cryo electron tomography on in situ lamellae. Nat Methods. 2023 Jan;20(1):131-138. doi: 10.1038/s41592-022-01690-1. Epub 2022 Dec 1. PMID: 36456783.
10. Tegunov D, Xue L, Dienemann C, Cramer P, Mahamid J. Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells. Nat Methods. 2021 Feb;18(2):186-193. doi: 10.1038/s41592-020-01054-7. Epub 2021 Feb 4. PMID: 33542511; PMCID: PMC7611018.
11. Zheng S, Wolff G, Greenan G, Chen Z, Faas FGA, Bárcena M, Koster AJ, Cheng Y, Agard DA. AreTomo: An integrated software package for automated marker-free, motion- corrected cryo-electron tomographic alignment and reconstruction. J Struct Biol X. 2022 May 10;6:100068. doi: 10.1016/j.yjsbx.2022.100068. PMID: 35601683; PMCID: PMC9117686.
12. Zivanov J, Otón J, Ke Z, von Kügelgen A, Pyle E, Qu K, Morado D, Castaño-Díez D, Zanetti G, Bharat TAM, Briggs JAG, Scheres SHW. A Bayesian approach to single- particle electron cryo-tomography in RELION-4.0. Elife. 2022 Dec 5;11:e83724. doi: 10.7554/eLife.83724. PMID: 36468689; PMCID: PMC9815803.
13. Castaño-Díez D, Kudryashev M, Arheit M, Stahlberg H. Dynamo: a flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments. J Struct Biol. 2012 May;178(2):139-51. doi: 10.1016/j.jsb.2011.12.017. Epub 2012 Jan 8. PMID: 22245546.
14. Himes BA, Zhang P. emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging. Nat Methods. 2018 Nov;15(11):955-961. doi: 10.1038/s41592-018-0167-z. Epub 2018 Oct 22. PMID: 30349041; PMCID: PMC6281437.
15. Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I, Ludtke SJ. EMAN2: an extensible image processing suite for electron microscopy. J Struct Biol. 2007 Jan;157(1):38-46. doi: 10.1016/j.jsb.2006.05.009. Epub 2006 Jun 8. PMID: 16859925.

16. Wagner T, Merino F, Stabrin M, Moriya T, Antoni C, Apelbaum A, Hagel P, Sitsel O, Raisch T, Prumbaum D, Quentin D, Roderer D, Tacke S, Siebolds B, Schubert E, Shaikh TR, Lill P, Gatsogiannis C, Raunser S. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. Commun Biol. 2019 Jun 19;2:218. doi: 10.1038/s42003-019-0437-z. PMID: 31240256; PMCID: PMC6584505.
17. de Teresa-Trueba I, Goetz SK, Mattausch A, Stojanovska F, Zimmerli CE, Toro- Nahuelpan M, Cheng DWC, Tollervey F, Pape C, Beck M, Diz-Muñoz A, Kreshuk A, Mahamid J, Zaugg JB. Convolutional networks for supervised mining of molecular patterns within cellular context. Nat Methods. 2023 Feb;20(2):284-294. doi: 10.1038/s41592-022-01746-2. Epub 2023 Jan 23. PMID: 36690741; PMCID: PMC9911354.
18. Carini P, Steindler L, Beszteri S, Giovannoni SJ. Nutrient requirements for growth of the extreme oligotroph 'Candidatus Pelagibacter ubique' HTCC1062 on a defined medium. ISME J. 2013 Mar;7(3):592-602. doi: 10.1038/ismej.2012.122. Epub 2012 Oct 25. PMID: 23096402; PMCID: PMC3578571.
19. Hu B, Margolin W, Molineux IJ, Liu J. The bacteriophage t7 virion undergoes extensive structural remodeling during infection. Science. 2013 Feb 1;339(6119):576-9. doi: 10.1126/science.1231887. Epub 2013 Jan 10. PMID: 23306440; PMCID: PMC3873743.
20. Mastronarde DN. Automated electron microscope tomography using robust prediction of specimen movements. J Struct Biol. 2005 Oct;152(1):36-51. doi: 10.1016/j.jsb.2005.07.007. PMID: 16182563.
21. Hagen WJH, Wan W, Briggs JAG. Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. J Struct Biol. 2017 Feb;197(2):191-198. doi: 10.1016/j.jsb.2016.06.007. Epub 2016 Jun 14. PMID: 27313000; PMCID: PMC5287356.
22. Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values. Bioessays. 2013 Dec;35(12):1050-5. doi: 10.1002/bies.201300066. Epub 2013 Sep 20. PMID: 24114984; PMCID: PMC3910158.
23. Lin J, Amir A. Homeostasis of protein and mRNA concentrations in growing cells. Nat Commun. 2018 Oct 29;9(1):4496. doi: 10.1038/s41467-018-06714-z. PMID: 30374016; PMCID: PMC6206055.
24. Schmidli C, Albiez S, Rima L, Righetto R, Mohammed I, Oliva P, Kovacik L, Stahlberg H, Braun T. Microfluidic protein isolation and sample preparation for high-resolution cryo-EM. Proc Natl Acad Sci U S A. 2019 Jul 23;116(30):15007-15012. doi: 10.1073/pnas.1907214116. Epub 2019 Jul 10. PMID: 31292253; PMCID: PMC6660723.
25. Farley MM, Hu B, Margolin W, Liu J. Minicells, Back in Fashion. J Bacteriol. 2016 Mar 31;198(8):1186-95. doi: 10.1128/JB.00901-15. PMID: 26833418; PMCID: PMC4859596
26. Yangci Liu, Haoming Zhai, Helen Alemayehu, Jérôme Boulanger, Lee J. Hopkins, Alicia C. Borgeaud, Christina Heroven, Jonathan D. Howe, Kendra E. Leigh, Clare E. Bryant, Yorgo Modis. Cryo-electron tomography of NLRP3-activated ASC complexes reveals organelle co-localization. bioRxiv. 2023 April 25. https://doi.org/10.1101/2021.09.20.461078.
27. Li M, Ma J, Li X, Sui SF. In situ cryo-ET structure of phycobilisome-photosystem II supercomplex from red alga. Elife. 2021 Sep 13;10:e69635. doi: 10.7554/eLife.69635. PMID: 34515634; PMCID: PMC8437437.
28. Serwas D, Akamatsu M, Moayed A, Vegesna K, Vasan R, Hill JM, Schöneberg J, Davies KM, Rangamani P, Drubin DG. Mechanistic insights into actin force generation during vesicle formation from cryo-electron tomography. Dev Cell. 2022 May 9;57(9):1132-

1145.e5. doi: 10.1016/j.devcel.2022.04.012. Epub 2022 May 2. PMID: 35504288; PMCID: PMC9165722.

29. Li Z, Du W, Yang J, Lai DH, Lun ZR, Guo Q. Cryo-Electron Tomography of Toxoplasma gondii Indicates That the Conoid Fiber May Be Derived from Microtubules. Adv Sci (Weinh). 2023 May;10(14):e2206595. doi: 10.1002/advs.202206595. Epub 2023 Feb 25. PMID: 36840635; PMCID: PMC10190553.
30. Martinez M, Mageswaran SK, Guérin A, Chen WD, Thompson CP, Chavin S, Soldati- Favre D, Striepen B, Chang YW. Origin and arrangement of actin filaments for gliding motility in apicomplexan parasites revealed by cryo-electron tomography. Nat Commun. 2023 Aug 9;14(1):4800. doi: 10.1038/s41467-023-40520-6. PMID: 37558667; PMCID: PMC10412601.