

TomoScore: A Neural Network Approach for Quality Assessment of Cellular cryo-ET

Zhao Wang

zhaow@bcm.edu

Baylor College of Medicine <https://orcid.org/0000-0003-4897-9986>

Xuqian Tan

Baylor College of Medicine <https://orcid.org/0000-0001-6044-8876>

Ethan Boniuk

Baylor College of Medicine

Anisha Abraham

Rice University <https://orcid.org/0009-0000-0107-6400>

Xueting Zhou

Baylor College of Medicine

Zhili Yu

Baylor College of Medicine

Steve Ludtke

Baylor College of Medicine <https://orcid.org/0000-0002-1903-1574>

Article

Keywords:

Posted Date: April 28th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-5405930/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

There is **NO** Competing Interest.

Ethics Institutional Review Board for Baylor College of Medicine and Affiliated Hospitals approved this study.

TomoScore: A Neural Network Approach for Quality Assessment of Cellular cryo-ET

Xuqian Tan¹, Ethan Boniuk^{1,2}, Anisha Abraham^{1,3}, Xueting Zhou¹, Zhili Yu¹, Steven J. Ludtke^{1,4}, Zhao Wang^{1,4,5,6,7*}

¹Verna and Marrs McLean Department of Biochemistry and Molecular Pharmacology, Baylor College of Medicine, Houston, TX 77030, USA

²Bioengineering Program, Rice University, Houston, TX 77030, USA

³Chemistry and Sociology Program, Rice University, Houston, TX 77030, USA

⁴CryoEM Core (Advanced Technology Core), Baylor College of Medicine, Houston, TX 77030, USA

⁵Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA

⁶Department of Materials Science and Nanoengineering, Rice University, Houston, TX 77030, USA

⁷Department of Molecular and Cellular Oncology, Division of Basic Science, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

*Correspondence: zhaow@bcm.edu (Z.W.)

Abstract

Electron cryo-tomography (cryo-ET) is a powerful imaging tool that allows three-dimensional visualization of subcellular and molecular architecture without chemical fixation. Tomogram quality varies widely, particularly during large high-throughput data collections, and the most common strategy for initial quality assessment is empirical judgment by an expert. Tomograms may be collected for two distinct purposes: annotation of subcellular features and cellular morphology, typically performed at lower magnifications and higher defocus, and subtomogram averaging, at high magnifications, closer to focus. For the first purpose, contrast and the ability to distinguish cellular features of interest are key, whereas for subtomogram averaging, recoverable signal at high resolution is the key factor. We have developed “TomoScore” a deep-learning based tomogram screening tool targeting cellular annotation. This tool provides a single quantitative measure of the suitability of a tomogram for annotation of subcellular features, in terms of the scale of features that can be readily distinguished. We further explore the relationship between accumulated electron dose and resulting quality, suggesting an optimum dose range for cryo-ET data collection. Overall, our study streamlines data processing and reduces the need for human involvement during pre-selection for tomogram segmentation.

Background

Electron cryo-tomography (cryo-ET) is the only current technique capable of providing the 3-D structure of unlabeled/unfixed cells at nanometer resolutions^{1,2}, as well as the emerging method of subtomogram averaging, providing high-resolution macromolecular structure *in-situ*³. These two related goals, subcellular annotation and subtomogram averaging, generally require different data collection parameters. For cellular annotation, lower magnifications provide larger fields of view, and higher defocus provides improved low-resolution contrast to reduce ambiguity when averaging is not being performed. For subtomogram averaging, higher magnification is required to achieve the necessary sampling, and lower defoci are required to make accurate CTF correction possible at high resolution. For subtomogram averaging tools, such as CTFFIND⁴, can provide good estimates of whether tilt series contain enough high-resolution information to be suitable. However, for cellular annotation, CTF analysis is far less useful, as measurable Fourier contrast may be present even when tomograms lack sufficient features for annotation.

Accurate interpretation of cellular tomograms requires annotation of features at different levels of detail, ranging from individual macromolecules to organelles, depending on the purpose of the study^{5–7}. However, tomogram quality can vary quite widely with imaging

conditions and how the cells were prepared for imaging. For example, a thicker lamella provides a larger cellular volume but decreases the contrast of individual features within the cell. When a project involves hundreds to thousands of cellular tomograms, prescreening tomograms to assess suitability for annotation is critical. This process is often performed manually by a human visually screening and, in some cases, trying to annotate each tomogram in the set^{8,9}. To reduce the need for this time-consuming process, we have developed a deep-learning based quantitative quality metric that can help rank tomograms, as well as assess approximate thickness. The “quality” we assess in this context is a measure of the smallest reliably resolvable feature present in individual 2-D slices of the tomogram. The quality metric is determined for all tomogram slices, providing an overall quality metric and thickness estimate. This metric is completely independent of typical resolution measures, such as the FSC between tomograms reconstructed from even movie frames and odd movie frames, which can produce both over and underestimates of the tomogram quality for annotation purposes^{10–13}. Therefore, our metric only accesses the quality of the final cellular reconstructed tomograms.

Since the problem of scoring tomograms is mathematically ill-defined, we made use of deep learning techniques, building on ResNet¹⁴, with human-derived training data to produce our trained SliceQuality Network. The network was trained across multiple cellular species, microscopes, and instrument settings. The strategy we used for manually quantifying quality is straightforward and has minimal annotator-dependence.

Results

Human-assessed quality categories for network training

For model training, we needed to generate a training set with a wide variety of input images with corresponding human-judged categorical quality. To build a comprehensive training data set, we selected various prokaryotic and eukaryotic cells from the Cryo-ET Data Portal (Chan Zuckerberg Imaging Institute, Chan Zuckerberg Initiative)¹⁵. We also included platelet tomograms generated by our lab as a specialized type of eukaryotic cell (anucleate) due to their unique cellular structure (Supplement Table 1). Overall, 58,527 slices from 114 tomograms across 19 species were gathered and manually assessed.

Empirically, humans generally assess tomogram quality by examining 2D slices through the tomogram roughly paralleling the beam path of the specimen. While orthogonal slices can provide some additional insights, due to the missing wedge and variability between algorithms, these other views are less useful for assessing tomogram quality. Most segmentation/annotation algorithms use a similar approach of annotating 2D slices before merging the results into a 3D annotation. We followed a similar strategy for our human-

based quality assessment, presenting one 2-D slice at a time for independent assessment. While the trained network produces a continuous quality variable, we limited human assessors to 6 discrete quality values based on the size in pixels of the smallest reliably discernible biological feature in each slice (Fig. 1). In addition to the 5 discrete feature-size categories presented in Fig. 1, a sixth category is reserved for tomogram slices with no discernible biological features at all.

During the human assessment of tomogram quality, we observed a logical trend: within one tomogram, central slices tend to have the highest quality, while the top and bottom slices have much lower quality. To encourage training the network to produce a smooth gradation between discrete scores, we applied a Gaussian filter to human-judged categories to generate continuous quality estimates across each tomogram.

We employed the following two strategies during data partitioning to establish rigorous training. Firstly, we performed a tomogram-based partition for unbiased training. Specifically, 98 of 114 tomograms were randomly selected for the training set pool, while the remaining tomograms were left for validation (1/114) and testing (15/114). This ensures the model does not peek into the testing tomogram due to highly similar neighboring slices from the same tomogram. Next, we performed a sample size partition for balanced training. Among the selected 98 tomograms, 2500 slices were randomly selected from each human labeled category to be added to the training, ensuring the prediction is not affected by sample size. Subsequently, our training set contains a total of 15,000 slices evenly distributed across six categories with corresponding continuous quality estimations.

SliceQuality Network, a modified ResNet101 model, presents high accuracy for quality score prediction.

We adopted the structure of ResNet101¹⁴ and modified it to predict the quality score of a single tomogram slice, as shown in Fig. 2a. Specifically, to obtain an output quality score ranging from 0 to 1, we added a sigmoid function at the end of ResNet101. Our continuous quality estimations in the training set uniformly map to the 0-1 prediction range (0.2: 6+ pixels; 0.4: 4-5 pixels; 0.6: 3 pixels; 0.8: 2 pixels; 1.0: 1 pixel). Our model takes less than 30 seconds to produce predictions for a standard 1k*1k tomogram containing ~200 slices on a machine with an NVIDIA GeForce RTX4090 GPU. Which is much faster than manual labeling.

To test model accuracy and generalizability, we selected 15 tomograms as the test set across different cellular species, microscopes, and magnifications. To quantitatively measure model accuracy, we overlaid the label scores and predictions and then plotted them by each slice's position within one tomogram. An average Pearson correlation

coefficient of 0.845 ($p < 0.001$) is reached between the prediction and label (Fig. 2b). Notably, our model also performs well on the three species (*Sulfolobus solfataricus*, *Treponema primitia*, and *Magnetospirillum magneticum*) that were not included in the training set, demonstrating its generalizability. We also noticed that two of the tomograms (row2, col2, and row3, col3) have a wider range of higher-quality slices for human labeling. After manual inspection, we found that both tomograms had tilted samples, resulting in the clear-cellular feature proportion of top and bottom slices being smaller than usual, thus receiving a lower predicted quality score. Further, our model also demonstrated its ability to differentiate between biological features and non-biological features, such as empty holes (row1, col3) and ice layers (row1, col4).

SliceQuality Network accuracy and consistency were further validated by tomogram accumulated dose *in silico* variations.

Assuming that our model is consistent and accurate, two identical tomograms should produce the same quality score using our model. However, practically, identical tomograms do not exist. Therefore, we developed an *in-silico* method to create “identical” tomograms to test model consistency and accuracy (Fig. 3a). First, we obtained a separate dataset containing 10 platelet tomograms with a high accumulated dose. Specifically, we recorded 152 raw frames per tilt image, with the total dose per tilt series being $400 \text{ e}^-/\text{\AA}^2$. The movie frames from each angle were separated into two stacks based on their even or odd positions. These stacks are expected to contain identical low-frequency signals discernible to the human eye while exhibiting varied high-frequency noise. Consequently, we obtained two “identically resolved” tomograms (Fig. 3b) using the same reconstruction parameters. We calculated the Fourier Ring Correlation (FRC) curves between the even-frames and odd-frames reconstructed tomograms and noticed the central slice (higher-quality slice) and top slice (lower-quality slice) share a similar trend of FRC curves. Our model, on the other hand, can distinguish the quality differences clearly between the central and top slices.

We generated a series of tomograms of different accumulated doses ($200 \text{ e}^-/\text{\AA}^2$ to $13 \text{ e}^-/\text{\AA}^2$) from the same tilt series by iteratively performing the even/odd splitting. Using our SliceQuality Network, we obtained the per-slice quality scores for the tomogram series with different accumulated doses, shown in (Supplement Fig. 2a). We further confirmed consistency between tomogram score predictions generated from even and odd frames at every accumulated dose level. Notably, all even/odd quality score differences were significantly lower than 16.7% ($\frac{1}{6}$), which is the discrete step size of the human categorization. This result indicates the prediction is on par with human judgments (Supplement Fig. 2b). Thus, we confirmed that our SliceQuality Network generates robust quality predictions at different doses.

Further, we observed that the predicted quality score peaks decrease as the accumulated dose decreases, indicating that there's a decrease in overall tomogram quality along with a decrease in the simulated cumulative dose used in the reconstruction¹⁶. We performed a visual comparison at the same position in the tomogram to validate the resolvability of certain structural feature changes according to the dosage. We chose two specific positions within the tomograms that effectively encapsulate the key features considered in the initial ranking criteria. As shown in Fig. 3d, 400 e⁻/Å², 200 e⁻/Å², and 100 e⁻/Å² tomograms clearly show microtubules with linear morphology and separation between neighboring objects to allow clear segmentation. Glycogens appear as high-contrast circular structures with precise boundaries to the cell environment in 400 e⁻/Å² and 200 e⁻/Å² tomograms. At lower dosages, organelles such as mitochondria lose contrast, which compromises the ability to visually identify the organelle by distinct features such as the double membrane and cristae. These results and observations inspired us to develop the following algorithm.

A novel metric, TomoScore, was created for cryo-ET quality assessment.

As we previously defined, a predicted quality score of 0.2 maps to the threshold for the presence of biological features (6+ pixel size). We consider any slices scored above 0.2 to contain biological features. Therefore, we estimated the sample thickness by counting the number of slices scored above 0.2. We validated the estimated thickness with actual thickness measurements from YZ projection (Supplement Fig. 3a,b,c) on 137 tomograms (excluded from the training) through the RANSAC algorithm¹⁷ (Supplement Fig. 3d) and obtained a regression slope of 0.988 (p<0.001). We manually examined each outlier tomogram and concluded three reasons: 1) missing wedge between angle steps. This causes some tomograms to contain smearing structure features propagated through the z-axis after reconstruction, leading the model to make a thicker estimation; 2) tilted samples would also often make overestimation for thickness; and 3) low tomogram resolvability. The most recent method for sample thickness estimation by CTFFIND5⁴ makes estimations specifically for protein-level samples, while our estimation focuses more on cellular samples. Thus, this method was not included in our comparisons.

$$Thickness_{est} = \#(SliceQualityNetwork(slice) > 0.2)$$

To simplify inter-tomogram quality comparison, we created a metric called TomoScore. TomoScore is generated from the predicted quality scores from a single tomogram and can be interpreted as the combination of all slice quality scores normalized by the estimated thickness of the specimen. As a result, TomoScore calculates the average quality score of slices within a single cell to represent the overall quality of a tomogram. Consequently, TomoScore has an expected range between 0 and 1, where 0 indicates no slice within the tomogram reveals cellular structure and 1 would denote a theoretically perfect tomogram with all slices revealing the optimal quality of organelle structures.

$$TomoScore = \frac{\sum(SliceQualityNetwork(slice) > 0.2)}{Thickness_{est}}$$

We further validated our model's consistency on TomoScore using the same even/odd frames split tomograms. For a single set of tomograms split into different total doses, the TomoScore differences between even/odd tomograms are small across the total dose from 13 e-/Å² to 200 e-/Å² (Fig. 4a). We also have a small range of distribution for TomoScore differences between the 10 even/odd split tomograms across different total doses (Fig. 4b).

To verify TomoScore aligns with human quality assessment, we visually examined the same 15 tomograms from the test set. Fig. 4c demonstrates how higher TomoScores correspond to higher quality tomogram slices as perceived by the human eye. This trend is preserved between different cell types and magnifications, indicating the broad utility of TomoScore in predicting tomogram quality.

TomoScore identifies the optimum electron dose range.

Calculating TomoScore for one of the tomograms separated into a series of different total doses, Fig. 4a also demonstrated a trend of increasing TomoScore with increasing simulated accumulated dose, which is consistent with common expectations². Specifically, the TomoScore increases drastically in the lower total dose range (13 e-/Å² to 100 e-/Å²) and reaches a plateau in the higher total dose range (100 e-/Å² to 400 e-/Å²).

To further confirm that this is not a coincidental trend limited to a single tomogram, and to understand the correlation between TomoScore and accumulated doses, we applied similar tests to the previously mentioned 10 high-electron-dose tomograms and created 5 more accumulated doses tomograms (256 e-/Å², 128 e-/Å², 64 e-/Å², 32 e-/Å², 16 e-/Å²) by extracting different percentages of evenly spaced raw frames. We calculated and plotted the percentage TomoScore each split dose tomogram (13 e-/Å², 16 e-/Å², 24 e-/Å², 32 e-/Å², 50 e-/Å², 64 e-/Å², 100 e-/Å², 128 e-/Å², 200 e-/Å², 256 e-/Å²) can reach of its full dose tomogram (400 e-/Å²) in Fig. 5a. In each set of tomograms, we observed the same trend that TomoScore is positively correlated with the total dose. At a dose range between 80-120 e-/Å², the tomograms achieved approximately 95% of their respective maximum score. Under a 60e-80e dose range, the tomograms attained around 90% of the highest score. In other words, data collection with an accumulated dose of 60-120 e-/Å² is optimum for high-quality platelet tomograms.

Next, we converted the total dose axis into a log2 scale and replotted the average max TomoScore percentage for the 10 series of tomograms to better quantitatively characterize the trend between TomoScore and the total dose observed (Fig. 5b).

Calculating the slope and intercept of the best-fit line for each tomogram series within this logarithmic scale, we observed a nearly perfect regression line ($R^2 = 0.93$, $p\text{-value} = 2.011 \times 10^{-6}$). This result provides additional confirmation that all 10 tomograms exhibit a similar trend, where their TomoScore proportionally increases with the log2 of the total dose.

It is acknowledged that our method of simulating different accumulated doses takes less consideration of dose damage. Still, the 10 tomograms we used for dose simulation were acquired and processed using identical protocols, controlling for all other factors that could potentially influence quality disparities except for inherent cell-to-cell differences.

Discussion

The application of deep learning techniques has opened new avenues for addressing the challenges of tomogram data processing. Our TomoScore quality evaluation system is especially useful for cryo-ET, where the datasets are large, complex, and noisy. New data collection strategies, such as Parallel/montage cryo-ET^{18–21} on Thermo Scientific EPU Software or similar SerialEM methods can collect multiple locations in one single angular step, increasing the throughput of cryo-ET data collection compared to traditional collection schemes. Moreover, rapid tilt-series acquisition²² or newer continuous tilt methods can acquire a complete tilt series in ~2 minutes, which increases CryoET collection speeds even further. Batch tomogram^{23–25} and most tomogram analysis software packages like IMOD^{26,27}, EMAN2^{28,29}, Tomoauto³⁰, emClarity³¹, Aretomo³², and Warp³³ can run automated or semi-automated reconstruction of tomograms in a pipeline. These improvements have dramatically increased the throughput of cryo-ET data collection and reconstruction. Therefore, an automated quality assessment tool specific to cellular cryo-ET is a critical step in high-throughput cryo-ET analysis pipelines.

In summary, our SliceQuality Network has proven capable of per-slice quality analysis of cellular cryo-ET across different species and magnifications. The resulting tool demonstrates the ability to accurately, consistently, and, more importantly, quickly assess the resolvability and thickness of reconstructed tomograms. TomoScore thus permits a large set of tomograms resulting from a data collection to be quality-ranked in order for further analysis. TomoScore is on par with human judgment and is fully automated, saving human effort for downstream processing.

Using TomoScore, we identified the optimal dose range for platelets under a specific set of experimental conditions. Clearly, other variables play a role in this optimal dose, such as properties of the cell type, selected defocus range, lamella thickness, and other issues,

but for a given set of conditions, TomoScore does provide a method for optimizing the targeted dose^{34–37}. By using an *in-silico* approach to control the accumulated dose, the resulting resolvability differences across different doses are not considered affected by radiation damage.

Overall, our SliceQuality Network, along with TomoScore can provide an overall estimation of the quality of cellular cryo-ET, taking all possible influencers into account. We successfully demonstrated this method is robust for tomogram quality prediction with high accuracy and consistency. We are confident that our TomoScore can contribute to quality standardization in the cryo-ET community.

Methods

Platelet sample preparation and freezing.

Freshly drawn healthy human blood (Gulf Coast Regional Blood Center) was centrifuged at 200G to obtain platelet-rich plasma (PRP), and further centrifuged at 100G to remove the residual red blood cells. We applied 2–4 μ L PRP to glow-discharged Quantifoil grids and used a Vitrobot (Mark IV, FEI Corp) to quickly freeze samples to form vitreous ice. Mouse platelet samples were taken from previous research⁷.

Platelet data collection, processing and reconstruction.

Platelet data was collected using a 300 kV FEI Titan Krios microscope with a Gatan K2 Summit direct electron detector camera, through the SerialEM software. The magnification was 11,500x, with a pixel size of 13.26 Å. The tilt series were collected using a 2° angular step, unidirectionally from –50° to +50°. The defocus ranges from –10 to –15 μ m. For the normal dose tilt series, the total dose was 100 e[–]/Å². For the high-dose tile series, the total dose was 400 e[–]/Å², achieved through increasing the exposure time of each tilt image. Each 400 e[–]/Å² total dose tomogram’s micrograph was collected in 152 frames (0.052 e[–]/Å²/frame). All frames were corrected using MotionCor2³⁸ with default setting to avoid beam-induced motion along with the gain correction.

Platelet tomograms used for model training were reconstructed using the automated pipeline in EMAN2³⁹. The percentage of tilt images kept for reconstruction was 90%. The tomograms were initially output as bin2 tomograms, then another bin2 was applied by EMAN2.

For the high dose electron tomograms used for even/odd raw frames split, reconstruction was first done through EMAN2 with 100% tilt images kept on the original 400 e[–]/Å² tomograms with full raw frames, and the alignment parameters were saved for later

reduced dose reconstruction. The same alignment parameters were then imported with no alignment and 0 rounds of iteration using all tilt images for reduced-dose tomogram reconstruction to ensure consistency.

Collect data from the Cryo-ET Data Portal.

To build our dataset for training, we used API from the Cryo-ET Data Portal (Chan Zuckerberg Imaging Institute, Chan Zuckerberg Initiative)¹⁵. Specifically, searched for tomograms of sizes larger than 800*800 pixels and of magnifications smaller than 50 Å/pixel but larger than 5 Å/pixel. We randomly picked 15 tomograms from each eukaryotic species and 3 tomograms from prokaryotic species that fit the selection criteria. Tomograms containing no visible biological features were excluded from our dataset after the first round of human screening. The remaining tomograms were added to our curated training dataset, as shown in Supplement Table 1.

To generate Fig. 4c, we selectively chose both prokaryotic and eukaryotic tomograms from the CryoET Data and blood platelet tomograms generated by our lab. We limited our selection to tomograms that had an Å/pixel value of 5 or higher to match the magnification range of the tomograms used in our training set. For each magnification level, we selected three tomograms that, to the human eye, had obvious differences in quality and ran TomoScore to confirm these differences.

Human quality scoring criteria.

Human scoring was performed using the EMAN2³⁹ 2D slice display function and its included measurement tools. Each slice of a tomogram was evaluated by the eye to find the smallest recognizable feature, and then the width of that feature in pixels would be measured and recorded. Pixel measurements within a range of ~1 pixel were treated as the same measurement to reduce human error, forming five discrete categories. Features from lower-quality tomograms were more blurry and less consistent, so there was a larger variation in pixel measurements for large features. To account for the inconsistency, the ranges of categories 1 and 2 were extended to >1 pixel. A sixth category labeled “0” was included as a measurement for slices in which no features were visible.

Sample thickness measurement

When measuring the thickness of cells using a tomogram reconstruction, we took a 2D projection along the XZ or YZ planes of the tomogram and measured the bounds of the cell by hand along the vertical Z axis (Supplementary Fig. 3a). We chose to use the YZ slices for taking measurements since the missing wedge effect leaves large artifacts along the XZ slices. Using EMAN2, we averaged 513 adjacent YZ slices to generate a 2D projection along the X-axis, smoothing out the boundaries of the cell. Measurements for cell thickness were recorded as the distance in pixels between the lower and higher

cell boundaries. To make sure our measurements matched the cell boundary, we examined the Z slices we chose as the bounds for the measurement (Supplementary Fig. 3bc). Measurement bounds were reasonable if the cell was visible in the slice, but no features were visible, and unreasonable if the cell was not visible at all.

Data preprocessing and augmentation.

All tomogram slices used for this research were cropped into 960*960 in pixel size and then normalized using EMAN2 before any further analysis or training.

Due to the scarcity of slices with the smallest features of 4-5 pixels and 6+ pixels, rotation augmentation was applied to these slices such that slices of 6+ pixels were rotated 90 degrees 3 times (90°, 180°, 270°) and slices of 4-5 pixels were rotated 90 degrees 2 times (90°, 180°).

Feature Diameter	No Feature	6+ pixels	4-5 pixels	3 pixels	2 pixels	1 pixel
#Slices after augmentation	33,580	2,920	3,410	4,147	5,285	4,963

Gaussian filter ($\sigma = 3$, and z as the coordinate of the slice) was applied on human-judged ranks to generate continuous quality estimates across each tomogram.

$$g(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/(2\sigma^2)}$$

SliceQuality Network training.

To achieve better generalization to ameliorate overfitting, we adopted the pre-trained weights of ResNet101 from PyTorch⁴⁰. An adagrad variant of stochastic gradient descent was used accompanied by mean square error loss implemented in the PyTorch with a learning rate of 5×10^{-7} for 200 epochs. Each slice of tomogram put into training was of size 960*960 pixels, and the minibatch size was 8. The model was trained on a single NVIDIA GeForce RTX 4090 GPU of 24GB of VRAM, and the model took around 72 hours to train.

Acknowledgments

We thank Hongjiang Wu and Valerie Dalton for providing suggestions on the manuscript. This work is supported by Robert Welch Foundation (X-F-0006-20240723), NIH-NHLBI (R01HL162842) to Z.W., NIH (R35GM151999) to S.L. We thank Tong Huo, Yuewei Wang, and Disuke Nakada for providing mouse and human platelet data. We acknowledge the use of the Cryo-EM supported by Advanced Technology CPRIT Cryo-EM/ET Core (1RP190602) at BCM and cryo-EM core at UTHealth Center in Houston.

Contributions

S.L. and Z.W. conceived and supervised the project. X.T. and A.A. selected and downloaded tomograms from the Cryo-ET Dat Portal¹⁵. Z.Y., X.Z, and X.T. performed data collection and processing. E.B. and X.T. specified scoring criteria. E.B. and A.A. scored all the tomograms. X.T. performed model training and evaluation. X.T., S.L., and Z.W. wrote the manuscript with other authors' input.

Reference

1. Turk, M. & Baumeister, W. The promise and the challenges of cryo-electron tomography. *FEBS Lett.* **594**, 3243–3261 (2020).
2. Oikonomou, C. M. & Jensen, G. J. Cellular electron cryotomography: Toward structural biology in situ. *Annu. Rev. Biochem.* **86**, 873–896 (2017).
3. Huang, Y., Zhang, Y. & Ni, T. Towards in situ high-resolution imaging of viruses and macromolecular complexes using cryo-electron tomography. *Journal of Structural Biology* **215**, (2023).
4. Elferich, J., Kong, L., Zottig, X. & Grigorieff, N. CTFFIND5 provides improved insight into quality, tilt and thickness of TEM samples. (2024) doi:10.7554/elifelife.97227.1.
5. Fischer, T. D., Dash, P. K., Liu, J. & Waxham, M. N. Morphology of mitochondria in spatially restricted axons revealed by cryo-electron tomography. *PLoS Biol.* **16**, e2006169 (2018).
6. Liu, C. *et al.* The architecture of inactivated SARS-CoV-2 with postfusion spikes revealed by cryo-EM and cryo-ET. *Structure* **28**, 1218-1224.e4 (2020).
7. Wang, Y. *et al.* Cryo-ET distinguishes platelets during pre-acute myeloid leukemia from steady state hematopoiesis. (2021) doi:10.1101/2021.10.18.464863.
8. Gan, L. & Jensen, G. J. Electron tomography of cells. *Q. Rev. Biophys.* **45**, 27–56 (2012).
9. Mahamid, J. *et al.* Visualizing the molecular sociology at the HeLa cell nuclear periphery. *Science* **351**, 969–972 (2016).
10. Van Heel, M. Similarity measures between images. *Ultramicroscopy* **21**, 95–100 (1987).
11. Böttcher, B., Wynne, S. A. & Crowther, R. A. Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature* **386**, 88–91 (1997).
12. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).

- 473 13. van Heel, M. & Schatz, M. Fourier shell correlation threshold criteria. *J. Struct. Biol.* **151**, 250–
474 262 (2005).
- 475 14. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *2016*
476 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).
477 doi:10.1109/cvpr.2016.90.
- 478 15. Ermel, Utz and Cheng, Anchi and Ni, Jun Xi and Gadling, Jessica and Venkatakrishnan,
479 Manasa and Evans, Kira and Asuncion, Jeremy and Sweet, Andrew and Pourroy, Janeece
480 and Wang, Zun Shi and Khandwala, Kandarp and Nelson, Benjamin and McCarthy, Dannielle
481 and Wang, Eric M. and Agarwal, Richa and Carragher, Bridget. A data portal for providing
482 standardized annotations for cryo-electron tomography. *Nature Methods* (2024)
483 doi:10.1038/s41592-024-02477-2.
- 484 16. Glaeser, R. M. & Hall, R. J. Reaching the information limit in cryo-EM of biological
485 macromolecules: experimental aspects. *Biophys. J.* **100**, 2331–2337 (2011).
- 486 17. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]* (2012).
- 487 18. Eisenstein, F. *et al.* Parallel cryo electron tomography on in situ lamellae. *Nat. Methods* **20**,
488 131–138 (2023).
- 489 19. Peck, A. *et al.* Montage electron tomography of vitrified specimens. *J. Struct. Biol.* **214**,
490 107860 (2022).
- 491 20. Khavnekar, S. *et al.* Multishot tomography for high-resolution in situ subtomogram averaging.
492 *J. Struct. Biol.* **215**, 107911 (2023).
- 493 21. Bouvette, J. *et al.* Beam image-shift accelerated data acquisition for near-atomic resolution
494 single-particle cryo-electron tomography. *Nat. Commun.* **12**, 1957 (2021).
- 495 22. Chreifi, G., Chen, S., Metskas, L. A., Kaplan, M. & Jensen, G. J. Rapid tilt-series acquisition
496 for electron cryotomography. *J. Struct. Biol.* **205**, 163–169 (2019).
- 497 23. Balyschew, N. *et al.* Author Correction: Streamlined structure determination by cryo-electron
498 tomography and subtomogram averaging using TomoBEAR. *Nat. Commun.* **15**, 5594 (2024).

499 24. Liu, H.-F. *et al.* nextPYP: a comprehensive and scalable platform for characterizing protein
500 variability in situ using single-particle cryo-electron tomography. *Nat. Methods* **20**, 1909–
501 1919 (2023).

502 25. Comet, M. *et al.* Tomo Live: an on-the-fly reconstruction pipeline to judge data quality for
503 cryo-electron tomography workflows. *Acta Crystallogr. D Struct. Biol.* **80**, 247–258 (2024).

504 26. Mastronarde, D. N. Correction for non-perpendicularity of beam and tilt axis in tomographic
505 reconstructions with the IMOD package. *J. Microsc.* **230**, 212–217 (2008).

506 27. Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-
507 dimensional image data using IMOD. *J. Struct. Biol.* **116**, 71–76 (1996).

508 28. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J.*
509 *Struct. Biol.* **157**, 38–46 (2007).

510 29. Galaz-Montoya, J. G. *et al.* Alignment algorithms and per-particle CTF correction for single
511 particle cryo-electron tomography. *J. Struct. Biol.* **194**, 383–394 (2016).

512 30. Morado, D. R., Hu, B. & Liu, J. Using tomoauto: A protocol for high-throughput automated
513 cryo-electron tomography. *J. Vis. Exp.* e53608 (2016).

514 31. Himes, B. A. & Zhang, P. emClarity: software for high-resolution cryo-electron tomography
515 and subtomogram averaging. *Nat. Methods* **15**, 955–961 (2018).

516 32. Zheng, S. *et al.* AreTomo: An integrated software package for automated marker-free,
517 motion-corrected cryo-electron tomographic alignment and reconstruction. *J. Struct. Biol. X*
518 **6**, 100068 (2022).

519 33. Tegunov, D., Xue, L., Dienemann, C., Cramer, P. & Mahamid, J. Multi-particle cryo-EM
520 refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells. *Nat. Methods* **18**,
521 186–193 (2021).

522 34. Hylton, R. K. & Swulius, M. T. Challenges and triumphs in cryo-electron tomography.
523 *iScience* **24**, 102959 (2021).

- 524 35. Arnold, J. *et al.* Site-specific cryo-focused ion beam sample preparation guided by 3D
525 correlative microscopy. *Biophys. J.* **110**, 860–869 (2016).
- 526 36. Campbell, M. G. *et al.* Movies of ice-embedded particles enhance resolution in electron cryo-
527 microscopy. *Structure* **20**, 1823–1828 (2012).
- 528 37. Castaño-Díez, D. & Zanetti, G. In situ structure determination by subtomogram averaging.
529 *Curr. Opin. Struct. Biol.* **58**, 68–75 (2019).
- 530 38. Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved
531 cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
- 532 39. Chen, M. *et al.* A complete data processing workflow for cryo-ET and subtomogram
533 averaging. *Nat. Methods* **16**, 1161–1168 (2019).
- 534 40. Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. *arXiv*
535 [*cs.LG*] (2019).

536

Figures

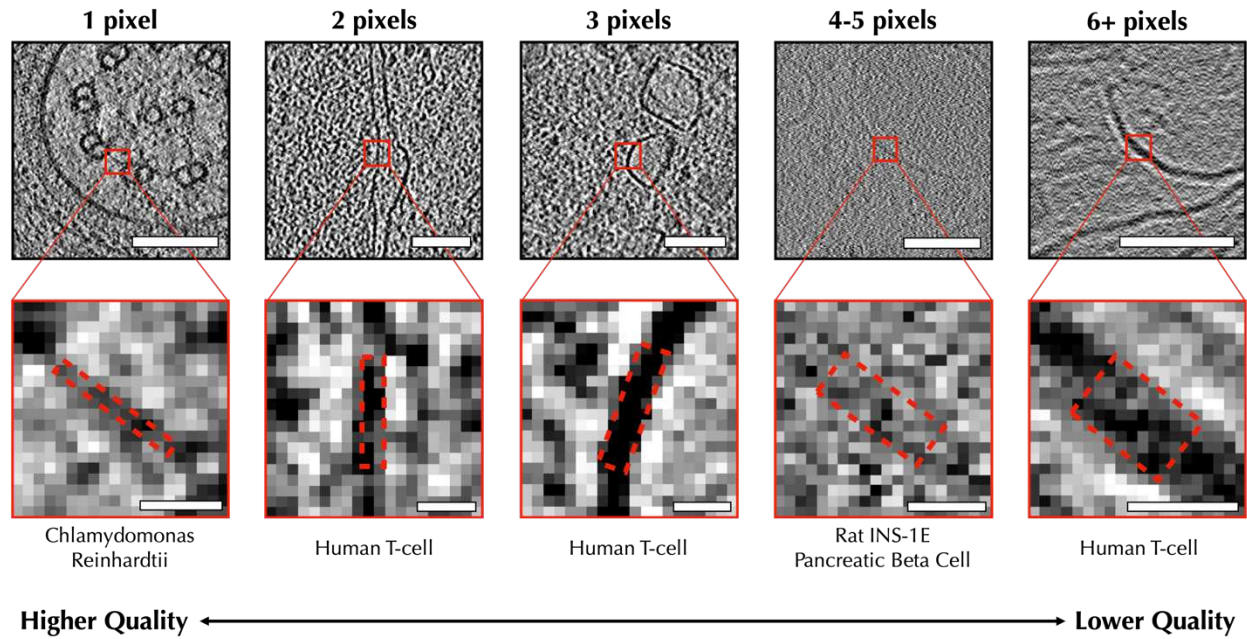


Figure 1. Demonstration of different quality tomogram slices under human-defined scoring criteria

This figure demonstrates how membrane features are scored based on their size in pixels. Example features are included from *C. reinhardtii* (left), Human T-cells (mid-left, middle, right), and Rat INS-1E pancreatic beta cells (mid-right). The dashed red rectangles indicate the width measured for each feature. Images on the top row have been cropped to 200x200 pixels with a scale bar of 100 nm. Images on the bottom row have been cropped to 20x20 pixels with a scale bar of 10 nm. From left to right, the angstrom per pixel values of each image are 13.68, 18.92, 18.92, 14.08, and 9.867 Å/pix respectively. Additional examples taken from other species, as well as measurements for particle features, can be found in Supplementary Fig. 1.

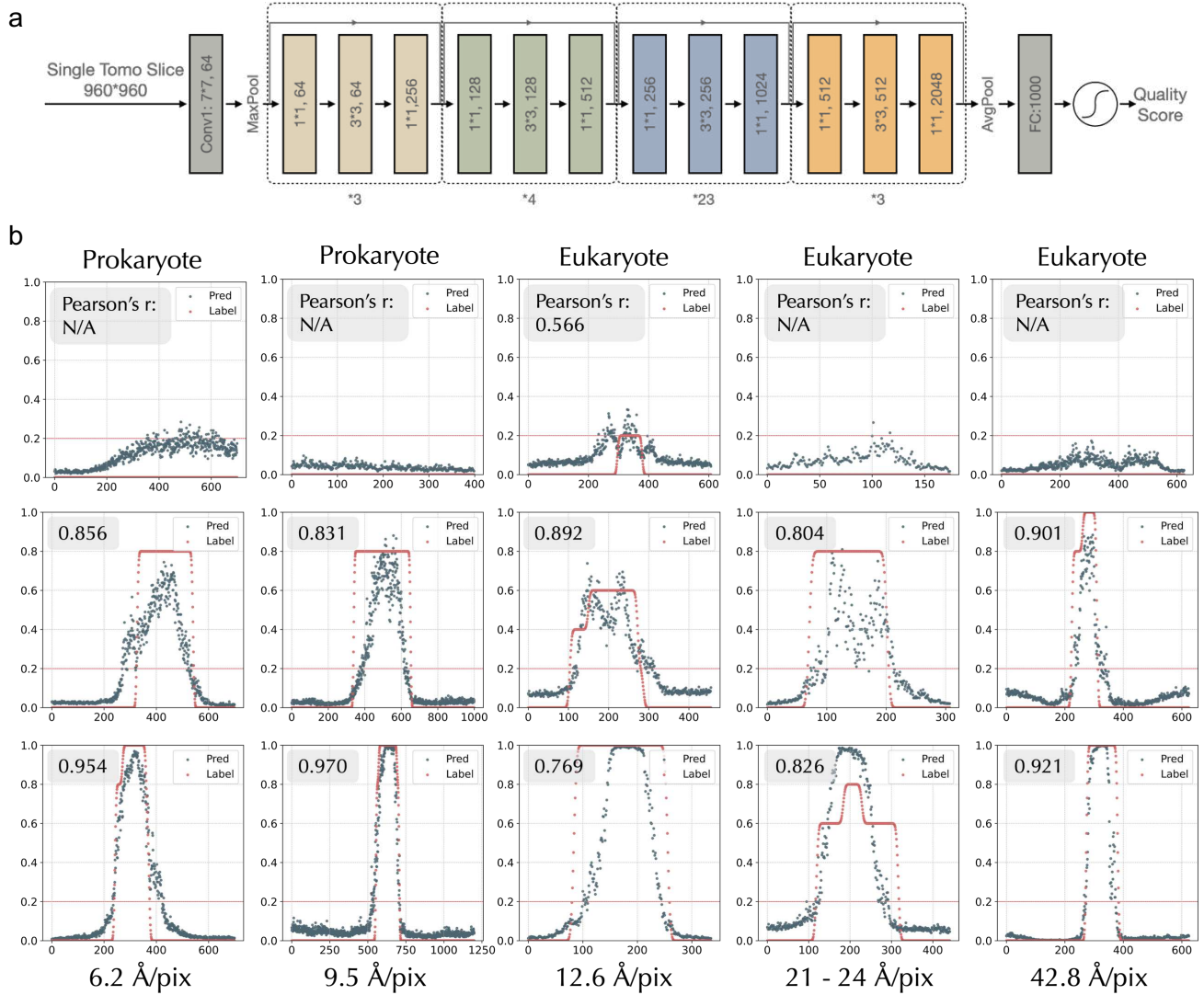
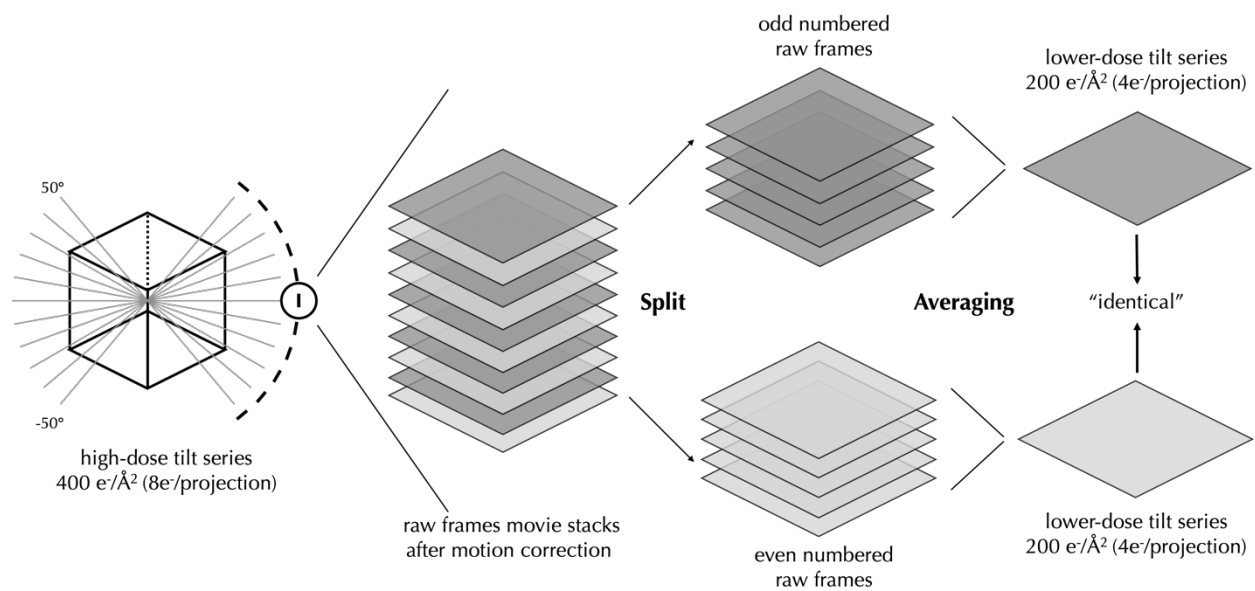


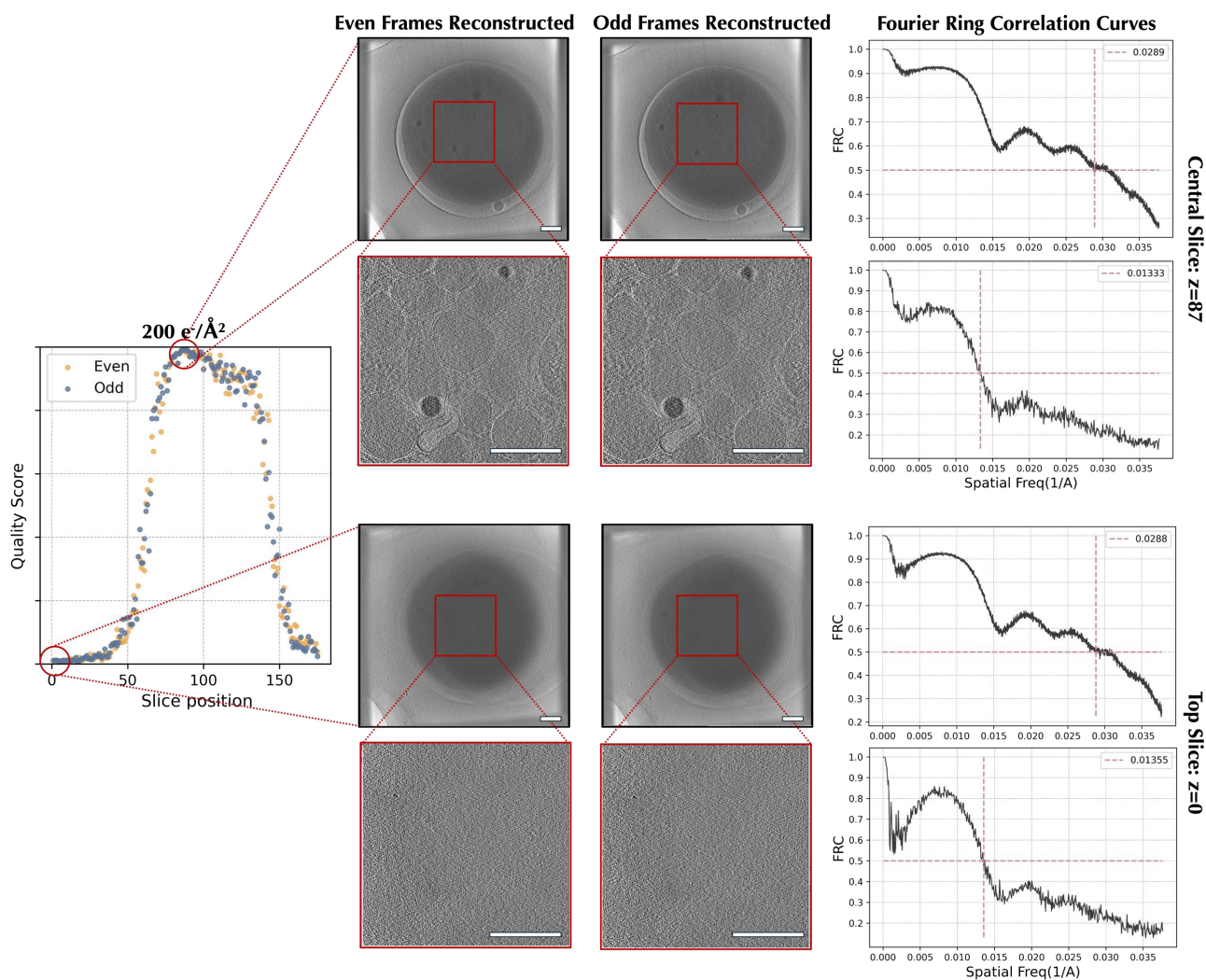
Figure 2 SliceQuality Network structure and training result

a). SliceQuality Network is modified from ResNet101 structure. A sigmoid function is appended before the output layer to give a predicted quality score within the range 0-1. b). SliceQuality Network's prediction score vs. label based on slice position curve for 15 test tomograms excluded from the training process. Pearson correlation coefficient is calculated (labeled on the top left) for each tomogram's prediction-label pair, and all of them are statistically significant ($p < 0.001$). Four of the lower-quality tomograms did not receive the Pearson correlation coefficient due to their constant labeling of 0. Each image has been cropped to 960x960 pixels for comparison. The first row captures the following, from left to right: *Sulfolobus solfataricus*, *Treponema primitia*, human umbilical vein endothelial cell (HUVEC), human blood platelet, and human blood platelet. The middle row captures the following, from left to right: *S. solfataricus*, *Vibrio cholerae*, HUVEC, HUVEC, human blood platelet. The last row captures the following, from left to right: *Magnetospirillum magneticum*, *V. cholerae*, HUVEC, HUVEC, and human blood platelet.

a



b



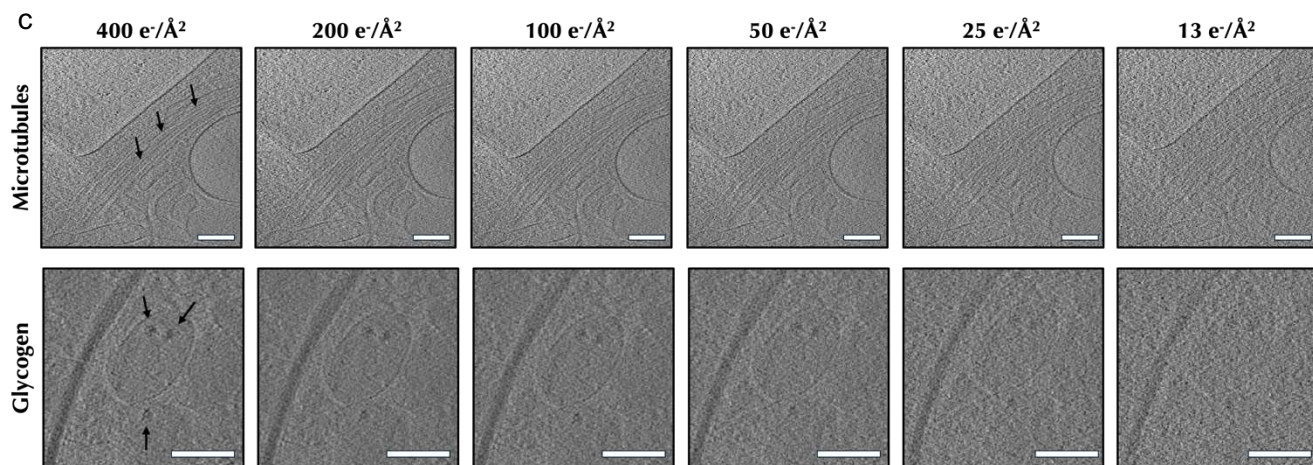


Figure 3 Model consistency tests on generated "identical" tomograms

a) A workflow of separating even/odd raw frames from a single 2D tilt projection. Separated raw frames were averaged and merged into two parallel 2D tilt projections. Repeat this process for all tilt projections in a series to generate two identical sets of tilt series such that each set is exactly half dose of the original tilt series. Each set of half-dose tilts was then reconstructed into a 3D tomogram using the same alignment parameters. b) The even/odd frames reconstructed tomogram slices are similar by visual comparison. The Fourier ring correlation (FRC) curve also shows that both central and top slices from the two tomograms share high similarities at lower resolution frequencies. On the right, quality scores predicted by our trained model indicate that central slices have better resolvability (i.e higher quality score), where the FRC curves fail to show. c) Level of radiation exposure differentially affects tomogram quality. At higher radiation doses, tomograms of the same platelet exhibit higher-quality structures of microtubules and glycogen. At lower doses, tomograms exhibit more noise and structures are less defined. Scale bars are 200 nm

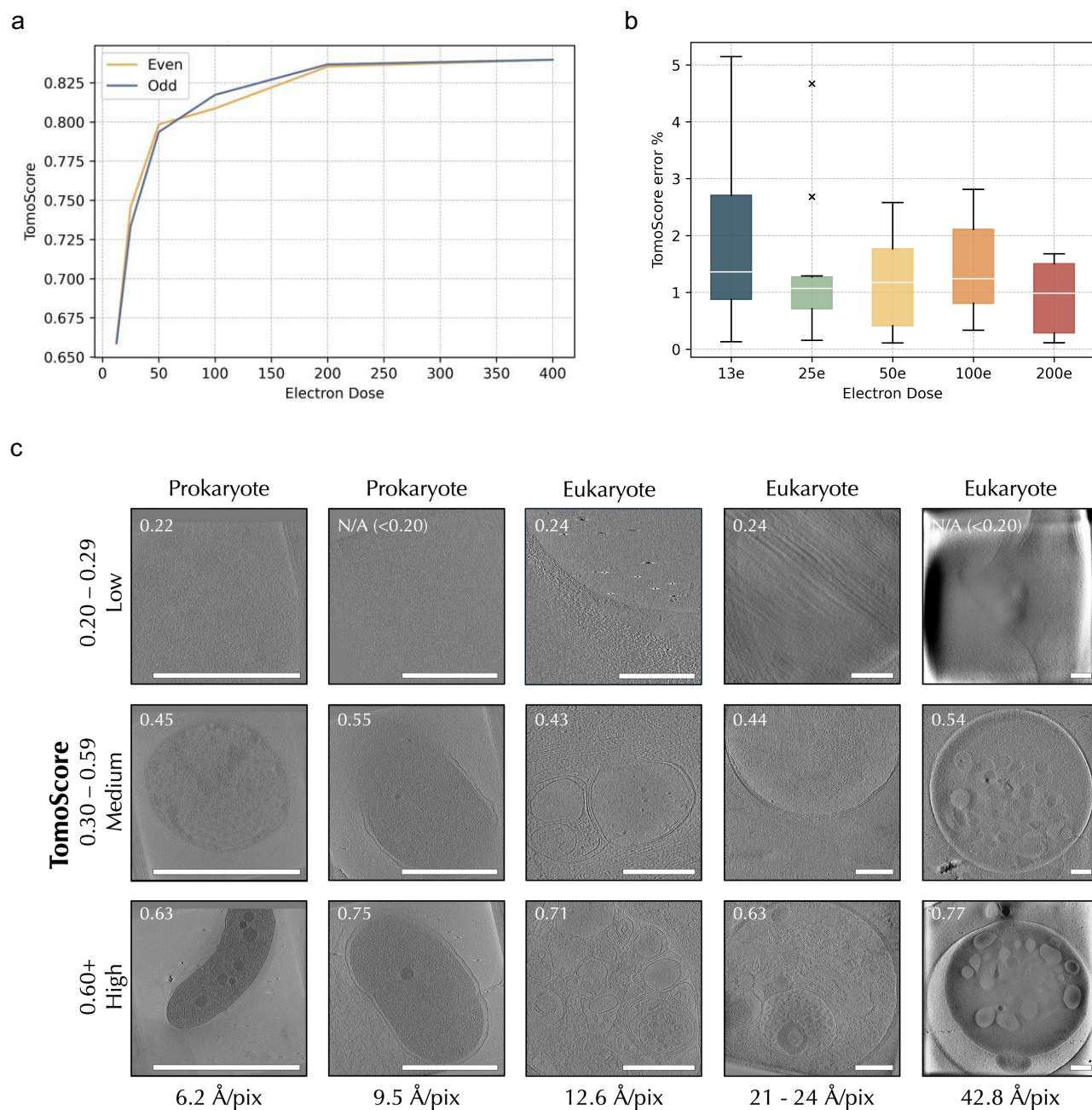


Figure 4 The trained model produces consistent TomoScore

a) Calculated TomoScores for recursive even/odd dose splits of a sample high-dose tomogram. TomoScore decreases as the total doses decrease. b) TomoScore differences distribution for all 10 high-dose tomograms across different split dosages. c) TomoScore accurately predicts tomogram quality across various cell types and magnifications. We compare the most resolvable slice, as determined by a human, of the same 15 tomograms used for Fig. 2b. As predicted, tomograms with a high TomoScore are of better quality and have more resolvable subcellular structures, and all scale bars are 500 nm.

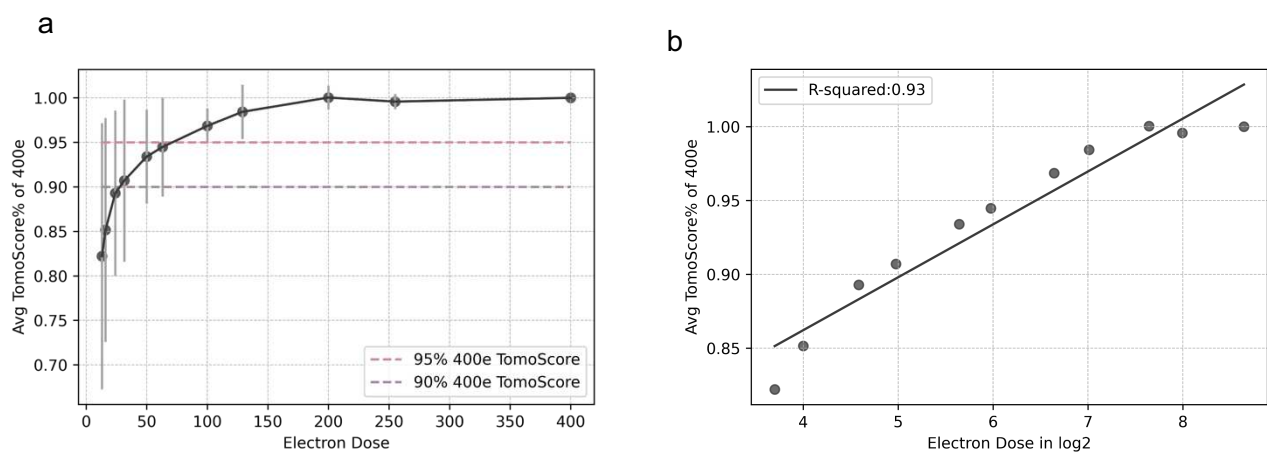


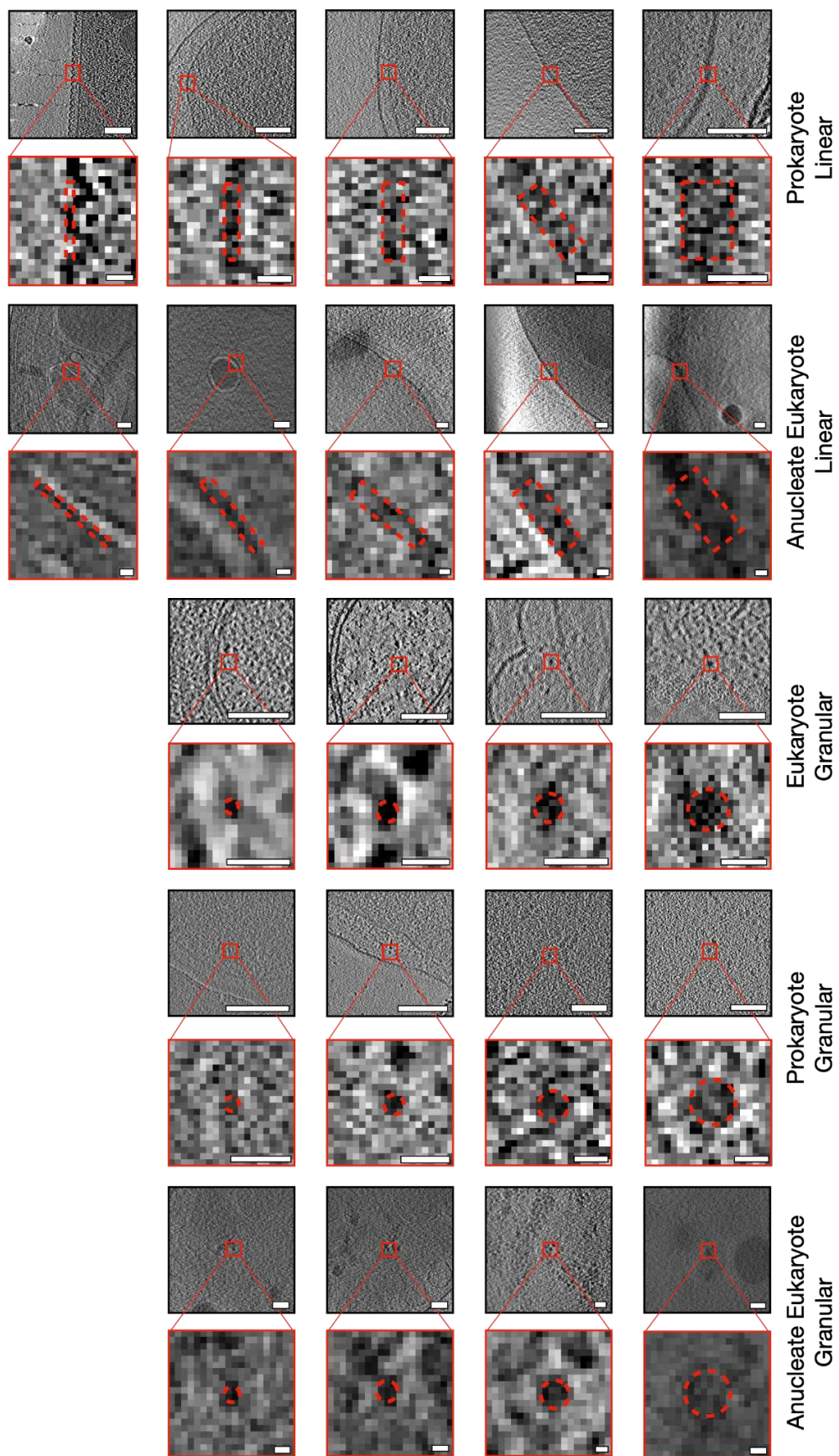
Figure 5 TomoScore determines the optimum dose range for cryo-ET data collection

For each total dose, we calculated the percentage of TomoScore each split tomogram can reach its full-dose and plot the percentage distribution of all 10 tomograms with the dose range in a) linear scale and in b) log2 scale.

Supplementary information

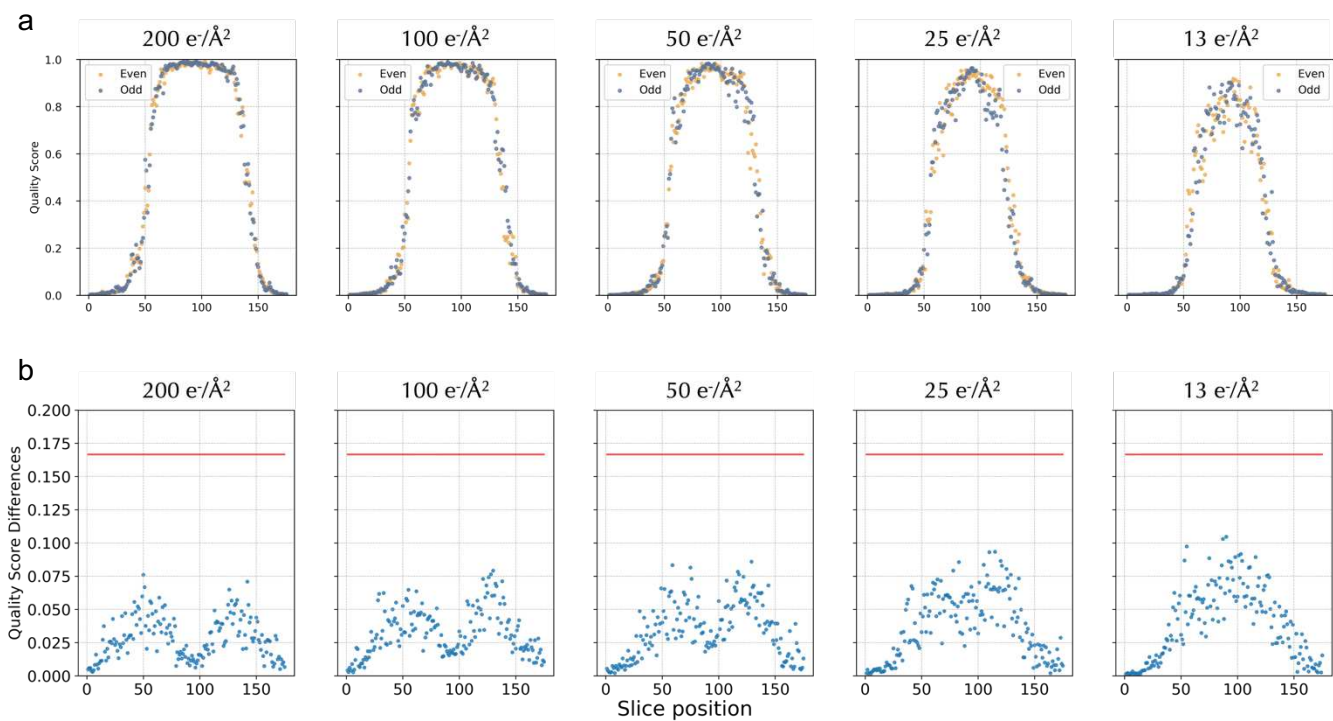
	Species	#Tomograms	#Slices
Eukaryote	<i>Chlamydomonas reinhardtii</i>	10	4616
	<i>Homo sapiens</i>	2	981
	<i>Homo sapiens</i> (T-cell)	13	3632
	<i>Homo sapiens</i> (HUVEC)	5	2150
	<i>Schizosaccharomyces pombe</i>	8	4972
	<i>Mus musculus</i>	5	1419
	<i>Rattus rattus</i> (pancreatic beta cells)	10	6078
Prokaryote	<i>Bacillus subtilis</i>	3	1126
	<i>Escherichia coli</i>	2	1791
	<i>Helicobacter pylori</i>	2	1293
	<i>Legionella pneumophila</i>	3	1489
	<i>Helicobacter hepaticus</i>	3	2388
	<i>Magnetospirillum magneticum</i>	1	700
	<i>Nitrosopumilus maritimus</i>	3	1488
	<i>Pseudomonas aeruginosa</i>	3	1789
	<i>Spiroplasma melliferum</i>	1	157
	<i>Streptococcus pneumoniae</i>	3	2092
	<i>Sulfolobus acidocaldarius</i>	3	2091
	<i>Sulfolobus solfataricus</i>	2	1400
	<i>Treponema primitia</i>	1	400
	<i>Vibrio cholerae</i>	5	4587
Anucleate Eukaryote (Platelets)	<i>Homo sapiens</i>	9	3395
	<i>Mus musculus</i>	17	8492
Total	19	114	58527

Table 1 Summary of tomogram dataset



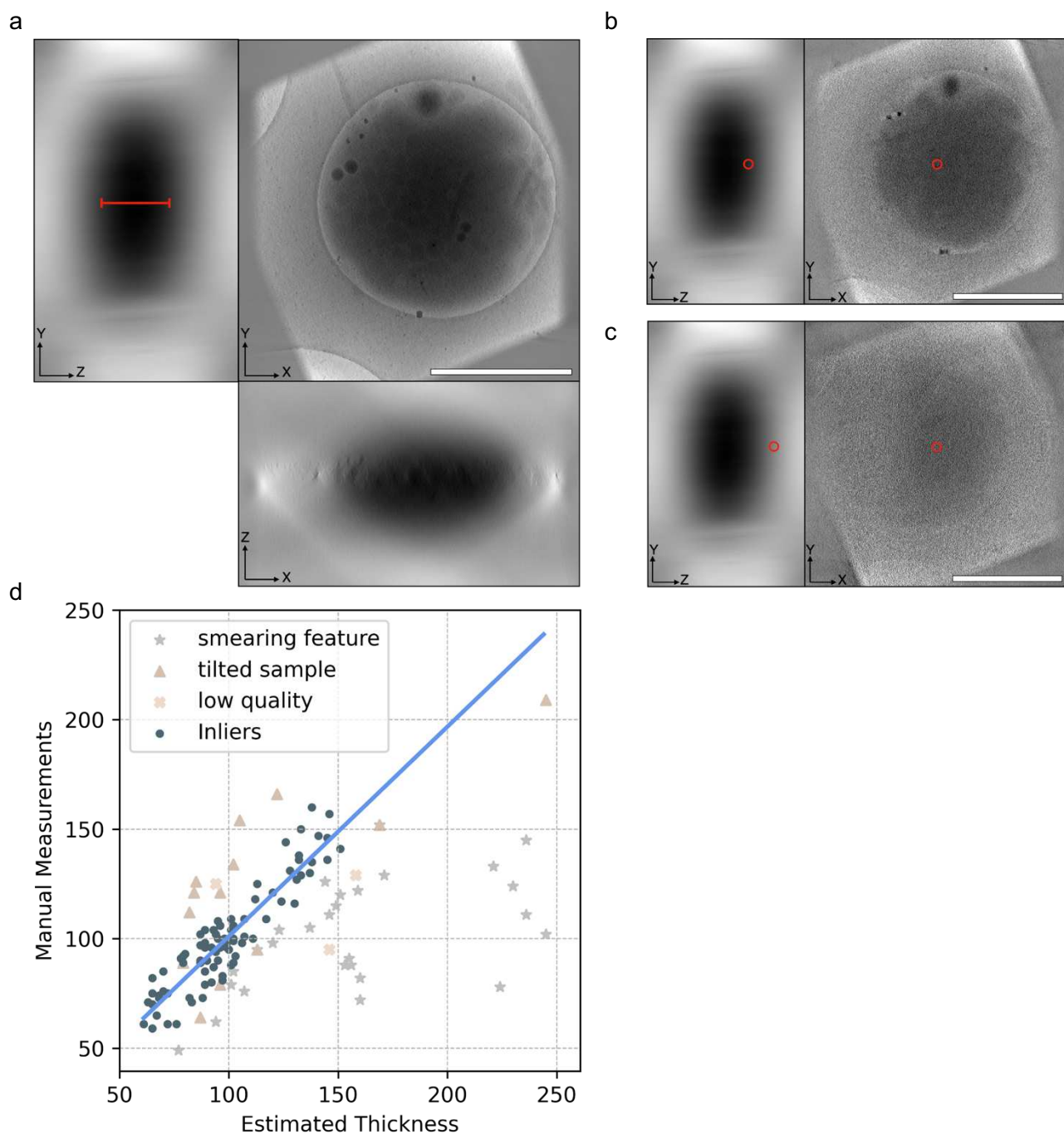
Supplementary Figure 1 Tomogram slices with quality-determining features

Additional examples of feature measurements in Eukaryotes (top rows), Prokaryotes (middle rows), and Anucleate Eukaryotes (bottom rows). Features are separated into membrane structures (left group), and particle structures (right group). Membrane features are measured by their width, while particle features are measured by their diameter. Black-bounded images are cropped to 200x200 pixels and have a scale bar of 100 nm. Red-bounded images are cropped to 20x20 pixels and have a scale bar of 10 nm.



Supplementary Figure 2 Even/Odd frames reconstructed tomograms are “identical”

a) Shows the SliceQuality Network’s predictions for the same set of tomograms, where the quality scores predicted for even/odd splits of different decreasing total dosages. b) Average quality score differences of 10 tomograms split into different dosages vary within a small range, indicating the high consistency of our trained model.



Supplementary Figure 3 Sample thickness estimation and validation

a) YZ, XY, and XZ projections of a platelet tomogram. The red bar shows which part of the image was used to measure cell thickness. b) The red circle indicates a point on the measured edge of the cell on the YZ projection. c) The red circle indicates a suggested alternative point of measurement taken at a lower contrast boundary. All scale bars are 1000 nm. d) RANSAC regression pinpointed inliers and outliers of the thickness estimations. 137 tomograms tested for thickness were all platelet cells. Reconstruction thickness varied from 175 to 600 slices in the z-direction.