emClarity: Software for High Resolution Cryo-electron Tomography and Sub-tomogram Averaging

Benjamin A. Himes[1,4] and Peijun Zhang[1,2,3*]

[1]Department of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15260, USA

[2]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK

[3]Electron Bio-Imaging Centre, Diamond Light Source, Harwell Science and Innovation Campus, Didcot OX11 0DE, UK

[4]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn VA, 20147, USA

[*]To whom correspondence may be addressed.

Email: peijun@strubi.ox.ac.uk

Abstract

Macromolecular complexes are intrinsically flexible and often challenging to purify for structure determination by single particle cryoEM. Such complexes may be studied using cryo-electron tomography combined with sub-tomogram alignment and classification, which in exceptional cases reaches sub-nanometer resolution, yielding insight into structure-function relationships. All maps currently deposited in the EMDB with resolution < 9 Å are from macromolecules that form ordered structural arrays, like viral capsids, which greatly simplifies structural determination. Extending this approach to more common specimens that exhibit conformational or compositional heterogeneity, and may be available in limited numbers, remains challenging. We developed **emClarity,** a GPU-accelerated image processing package, specifically to address fundamental hurdles to this aim, and demonstrate significant improvements in the resolution of maps compared to those generated using current state-of-the-art software. Furthermore, we devise a novel approach to sub-tomogram classification that reveals functional states not previously observed with the same data.

Introduction

Recent advances in the capabilities of cryo-electron microscopy (cryoEM) have enabled structures approaching atomic resolution. Software tools for the single particle analysis (SPA) of cryoEM data have been developed to probe macromolecular functional dynamics; for example, the maximum-likelihood approach to classification as implemented in *cis*TEM[1] or RELION[2]. For a sample to be suitable for SPA, it must yield tens to hundreds of thousands[3] of particles, purified to high compositional and conformational homogeneity[4] and subsequently imaged in many different orientations. These conditions are often difficult to realize, especially for the large and dynamic assemblies of biological complexes most relevant to cellular activities[5]. An alternative to SPA, cryo-electron tomography (cryoET), can be used to generate three-dimensional (3D) reconstructions of the specimen.

These reconstructions (tomograms) are typically limited to ~3-4 nm resolution due to the limited electron dose used to prevent excessive radiation damage to samples[6]. Additionally, the signal-to-noise ratio (SNR) is not distributed evenly in the tomogram, resulting in anisotropic resolution. This is a consequence of both increasing specimen thickness at high tilt angles and restricted angular sampling (+/- 60°) known as the "missing-wedge effect[7]". When many copies of a macromolecule are present in a tomogram, they may be extracted *in silico*, aligned to a common orientation, and averaged using procedures that share many similarities to SPA, which ameliorates the distortions due to the "missing-wedge effect". One major difference compared to SPA is that the metrics used in the orientation search are biased by the missing-wedge. This bias may be compensated by only considering the regions in Fourier space where both volumes being compared have been measured; the two most common being the constrained cross correlation[8,9] and the constrained Euclidean distance[10].

Obtaining averages of sub-tomograms at low resolutions, 15-20 Å, is now relatively routine, and is a prerequisite for further classification of the sample into multiple biological states or functional conformations[11,12]. Compared to SPA, this is arguably the greatest strength of cryoET and sub-tomogram averaging, because each particle exists as a unique, albeit distorted, 3D reconstruction. This allows for a direct analysis of the 3D variance, the value of which has been discussed extensively[13,14]. Despite substantial progress over the last decade, very few

structures have been solved at resolutions better than 8 Å using cryoET and sub-tomogram averaging, a critical threshold beyond which flexible molecular fitting approaches are more reliable[11,15].

We present here a complete set of GPU-accelerated programs called **emClarity** for **e**nhanced **m**acromolecular **cl**assification and **a**lignment for high-**r**esolut**i**on **t**omograph**y**, with the aim of routinely reaching beyond the critical sub-nanometer resolution for diverse specimens. We have focused our efforts on those areas of image processing that are likely to yield the greatest improvements, as suggested by empirical observation and theoretical calculations[16,17]: accuracy of tilt-series alignment, improved defocus determination and CTF correction, explicit treatment of anisotropic resolution, and more robust classification.

Results

emClarity workflow

A typical emClarity workflow is illustrated in figure 1, with enhancements highlighted in red text. Detailed steps are described below. A comparison of features in available sub-tomogram processing packages, including emClarity, PEET, RELION, Dynamo, Jsubtomo, pyTom and Protomo is presented in supplement table S1. Benchmarks for each major operation are in table S2, and compared among several popular Nvidia GPUs in figure S1

I) Input data

The raw data for emClarity are tilt-series, rather than tomograms, and an initial estimate of the tilt-series alignment parameters—tilt-axis angle, in-plane shifts and rotations, magnification and tilt angle—readily obtained using a separate installation of the IMOD software package[18]. For the defocus determination we use periodogram averaging as described previously[19], refined by fitting a 2D astigmatic CTF and using a low pass filtered power spectrum for background subtraction[20].

II) Tomogram WBP, template matching with 3D interactive editing

By default, emClarity limits the resolution of the initial tomogram used in template matching to 40 Å to prevent model bias. This also means CTF correction is not needed and traditional weighted back-projection is sufficient to reconstruct these initial tomograms.

Sub-tomogram positions and orientations chosen by the best scoring locally normalized cross-correlation, as described previously[21] with two main exceptions. First, our algorithm is optimized to fit in GPU memory. Second, we do not use distribution fitting to estimate the number of false-positives, rather we provide a maximum intensity projection and interface with IMOD's 3D model editing tools to enable rapid manual cleaning of the results.

III) 3D-CTF corrected WBP

To correct for the defocus gradient along the optical axis (sample thickness), emClarity uses a straightforward version of the "Defocus Gradient Corrected Back Projection" as described by Jensen and Kornberg almost twenty years ago[22]. The approach has recently been validated and re-named "3D-CTF correction[23,24]. To balance accuracy with practical compute time, emClarity determines the acceptable thickness based on the current resolution and defocus as discussed in the online methods section. For each slab of this thickness we whiten the power spectrum[25], multiply by the determined CTF and filter according to cumulative electron dose[26]. For tilted images, strips of a width corresponding to the current accepted defocus are extracted from the inverse Fourier transform of the full image multiplied by the respective CTF.

IV) Averaging sub-tomograms, CTF amplitude correction, Anisotropic SSNR weighting

The iterative alignment procedure in emClarity alternates between averaging the sub-tomograms using the current estimate of their orientations, and performing a missing-wedge constrained cross correlation grid search[27]. In addition to the data volumes, an average is also made of the "3D-sampling function" which is similar to the "weighted 3D CTF model" described in figure 1 of Bharat et al.[10] with the additional consideration of the R-weighting applied during tomogram reconstruction (Online Methods **eq 1)**. *To avoid confusion with "3D CTF correction", we do not adopt the name "3D CTF model", instead using "3D-sampling function".*

V) 3D-sampling function compensated iterative refinement of sub-tomogram alignment

The iterative refinement procedures commonly used in cryoEM are prone to erroneously fitting noise, known as "over-fitting[28]". To minimize over-fitting emClarity divides the data from the beginning into two halves which are kept separate during refinement, the so-called "gold standard" approach[29]. Additionally, the references used in the constrained search are carefully filtered as follows. In each cycle the SSNR of the average is estimated by the "gold-standard" FSC[a]. A figure-of-merit weighting[30] derived from this FSC is then combined with CTF amplitude restoration via our adaption of the post-reconstruction volume normalized single-particle Wiener Filter[31] (eq 8 in the original paper). Importantly, this adaptation involves an explicitly accounting of the directional anisotropy in the distribution of signal, as described in the Online Methods.

The iterative procedure is a local refinement, improving on the initial global alignment obtained during template matching. Importantly, we rotate the noisy particles back into the microscope reference frame for cross-correlation with the reference volume. This allows symmetry to be applied to the particle improving the SNR in the orientation search. This is not possible in SPA, where the particle is a projection and the reference must be rotated to the particle's orientation, and to our knowledge, is not implemented in any other sub-tomogram averaging package.

VI) Iterative refinement of tilt-series alignment

One of the novel features of emClarity is the iterative refinement of the tilt-series alignment by using sub-tomograms as fiducial markers; tomogram constrained-particle refinement (tomoCPR.) It is an approach similar to "particle polishing[32]" implemented for SPA in RELION with two primary differences. First, the reference projections we generate for refining the location of the sub-tomogram fiducial markers in the "raw" tilt-series, includes information from neighboring particles as well as non-particle information that is present in the tomograms as illustrated figure 2. Second, tomoCPR constrains spatially proximal particles to behave similarly within a given projection, as in SPA, while also requiring them to vary smoothly as a group from projection to projection through the tilt-

[a] These alignments are only for the FSC calculation and are never considered in the updates to the particle orientations in the iterative alignment, as necessary to maintain independence between the two half-sets.

series. The set of image transformations (shift, in-plane rotation, tilt-angle, and magnification) are fit for a grid of overlapping patches each containing a fixed number of particles, determined by the total molecular weight. The single set of image transformations that minimize the error for all fiducials in a given patch over all projections are solved using IMOD's *tiltalign*. As the patches overlap significantly (0.75), the image transformations vary smoothly over neighboring particles.

VII) 3D-sampling function compensated classification

Regions of significant variance across a data set may be visualized by overlaying a 3D "variance map" with the average structure. The "missing-wedge" produces significant artifacts that are specific to the orientation of each particle in the sample, but not necessarily its identity or conformation. Left uncorrected these artifacts obscure meaningful differences among particles, reflected in a diffuse variance across the data set which can be seen in figure 3 a-c. A previously demonstrated technique for estimating the effect of the "missing-wedge" by using a binary mask, called "wedge masked differences (WMDs)", was shown to be a good first order correction[33]; however, the accuracy of this model breaks down when higher-resolution features are considered figure 3 d-e. To allow higher-resolution information in the classification, we replace this binary wedge mask with our 3D-sampling function, resulting in a more accurate estimate of the artifacts introduced by the "missing-wedge" as shown in figure 3 g-i. It is worth noting that this does not "fill in" any missing data, rather it estimates what a given particle should look like by distorting the current sub-tomogram average by that particle's 3D-sampling function, and clusters based on the difference between this expected value and the observed particle.

VIII) Multi-scale clustering

We encode *a priori* biological information through introducing inter-voxel correlations at biologically relevant length scales, such as ~ 10 Å for alpha-helical density, 18-20 Å for RNA helices or small protein domains, and ~ 40 Å for larger protein domains. This is accomplished by selecting features of given length scale via a bandpass filter. A singular value decomposition is run at each length scale using the native MATLAB function SVD, and the singular vectors describing the greatest variance for each length scale are then concatenated into feature vectors for further clustering. While this approach is similar to existing ideas in multi-scale multi-variate

statistical analysis applied in other fields[34], because emClarity ***considers each length scale simultaneously***, the approach is capable of providing a richer description of the feature space.

emClarity improves resolution in sub-tomogram averaging

Given the inherent difficulty in working with extremely low SNR cryoEM data, and the sensitivity of the results to optimal selection of parameters[b], we have elected to test and demonstrate our software using two publicly available data sets from the Electron Microscopy Pilot Image Archive[35] (EMPIAR). We show these published/deposited maps, juxtaposed with the maps obtained with emClarity in figure 4. A total improvement in the yeast 80s ribosome from EMPIAR-10045 using RELION version 1.4 (EMD-3228[36]) from 12.9 Å to 7.8 Å is achieved using emClarity (Figure 4a). For the mammalian 80s ribosome from EMPIAR-10064 using pyTOM (EMD-3420[37]), we obtained an improvement from 11.2 Å to 8.6 Å (Figure. 4b).

To evaluate the relative impact of each of the individual features implemented in emClarity, we incrementally included them into several reconstructions of the yeast 80s ribosome. To control for errors in alignment and to have a one-to-one comparison with EMD-3228, we used precisely the same particles and orientation parameters from the star files that accompany the raw data EMPIAR-10045. We compare each map to an external reference map derived from SPA (EMD-2275[38]), via a cross-Fourier Shell Correlation (cross-FSC), starting from the RELION reconstruction as a control (Figure 4c). The results demonstrate the recovery of additional signal from the same data as each subsequent feature is incorporated.

The accuracy of our combined CTF correction approach, including a re-implementation of Grant and Grigorieff's optimal-exposure filtering and 3D-sampling function based Wiener filtering is reflected in the magenta curve in figure 4c, which shows a significant improvement over the cross-FSC of the control. The largest single improvement comes from the tomoCPR which is shown in green (and obviously includes the features in the magenta curve as well.) A more modest improvement is measured when we add in a per-tilt defocus estimation, cyan cross-FSC. When we also explicitly consider anisotropy in the SSNR in our adaptation of the single

[b] It is worth noting that the authors for the maps we use for comparison are also authors on the primary publications for their respective software packages, which helps to ensure the resolutions reported are likely optimal for the given data.

particle Wiener filter, we see another substantial improvement in the cross-FSC in the red curve. The yeast 80s sample that was used has a preferential orientation which is reflected in the FSC-cones and a plot of the angular distribution in supplemental figure S2. The final and highest resolution curve represents an alignment carried out in emClarity with all features added, illustrating the impact these advances have on the accuracy in the orientation determination.

In addition to improved resolution, there is a density outside the peptide exit tunnel of the ribosome (as noted in figure 4b, white arrow) that is present in the map derived with emClarity, but not in the map derived with pyTOM. Finally, we show the density from a peripheral region with a rigidly docked model of the yeast 80s ribosome (PDB-47VR) that underscores the difference in interpretability between the maps from the original publication and emClarity (Figure 4d).

Both ribosome samples had very few gold-fiducial markers and limited overlapping density, such that the impact of tomoCPR maybe the greatest. How does emClarity perform against more challenging cases? We tested emClarity on the HIV-1 immature Gag particle dataset of Schur et. al[39] which was recently released (EMPIAR-10164). These data yielded the highest resolution (3.4 Å) sub-tomogram average to date (EMD-3782)[23]. Using emClarity, we produce a sub-tomogram average at 3.1 Å resolution (Figure 4d, Figure S3). The quality of the density map is clearly manifested in the real-space refinement of the HIV-1 CA structure using Phenix[40] (Figure 4e).

emClarity improves classification and reveals multiple ribosome functional states

Using multi-scale clustering combined with 3D-sampling function compensated Principal Component Analysis, emClarity reveals subtle conformational differences and distinguishes minor populations from noisy and distorted images, as demonstrated with yeast 80s ribosome data from EMPIAR-10045 (Figure 5), and mammalian 80s ribosome data from EMPIAR-10064 (Figure 6). Such results were not previously obtainable using existing software[37,36].

Classification of non-translating Yeast 80s ribosomes

The ribosome is a complex molecular machine composed of RNA and protein which exists in many functional states and interacts with an array of co-factors. The eukaryotic

ribosome is composed of two major domains dubbed the large subunit (60s) and small subunit (40s). While the ribosome has a well-conserved catalytic core which mediates the peptidyl transferase reaction[41], it is increasingly subject to more complex regulation in higher organisms resulting in an expanded set of both RNA and protein components. RNA expansion segments are found primarily at the periphery of the ribosome and are typically *highly dynamic* and difficult to resolve in structural analysis. One good example is es27, an approximately 150 Å RNA helix which predominantly adopts one of two conformations separated by about 90º, shown in orange in figure 5 a-e. The first situates the end of the RNA helix just outside the peptide exit tunnel on the 60s subunit ($es27_{pet}$, figure 5a, b, d, e) and the second points toward the tRNA exit site ($es27_{L1}$, figure 5c). This dynamic domain is generally observed in cryoEM maps as a superposition of these two states, as is the case with the currently published results by ML classification in RELION[36]. A notable exception being ribosomes with accessory complexes bound at the peptide exit tunnel, e.g. Sec61, are known to bias the conformation to the $es27_{L1}$[42].

Another example of a highly dynamic ribosome domain is the L1 stalk – comprised of protein L1, and RNA helices h75, h76 and h79 from the 25s portion of the 60s subunit[43] – which moves through ~ 55Å during a translocation cycle. The motions of L1 are well correlated with several defined functional translocational states as observed using single molecule FRET and SPA[44]. Using emClarity, we can discern three of these L1 conformational states isolated from thermal (stochastic) fluctuations of the non-translating yeast 80s ribosome: $L1_{open,}$ $L1_{int,}$ and $L1_{closed}$ shown in green with variable occupancy in the five classes in figure 5. In addition to isolating dynamic states, identifying very sparsely populated classes is a particularly important and challenging task for classification of cryoEM data. We see in figure 5e the dissociated 60s subunit occupying a minor class, only ~4% of the data set or roughly ~140 sub-tomograms. In contrast, the Maximum likelihood approach implemented in RELION found three classes, one designated as a junk class and two relatively indistinguishable classes[36]. This minor population could only be isolated in the case where feature vectors built from the projection on the principal components from at least three length-scales were simultaneously clustered.

Mammalian 80s ribosome

In contrast to the non-translating yeast specimen, the mammalian ribosomes imaged in EMPIAR-10064 were prepared from clarified rabbit reticulocyte lysate using a buffer low in Mg^{2+} but lacking polyamines, such that cofactors should co-purify excepting perhaps some loss of e-Site tRNA[45]. We extracted 3,090 ribosomes from the four tilt-series deposited as the "mixed-CTEM" data set on EMPIAR, which are collected over a range of defocus values *without* a phase plate. emClarity identified five predominant classes as shown in figure 6a-e. Three of these classes show ribosomes adopting a non-rotated 40s conformation with variable tRNA eeF1A occupancy (class I-III), while two very similar classes adopted a mid-rotated (~5-6º) 40s conformation with eeF2 present (class IV-V).

A rigid body docking of the full 80s mammalian ribosome in the non-rotated POST state from PDB-4UJE[46] showed very clear agreement with the conformation of the 40s subunit, which combined with the co-factors observed suggest classes II and III are POST trans-locational ribosomes differing in retention of E site tRNA while class I is most similar to the "sampling" state. Classes IV and V both have eeF2 bound and differ in rotation of the 40s subunit of 5.9º and 5.0º, respectively (Figure 6f-g). Rigidly docking the 80s yeast structure of eeF2 from PDB-4UJO[47] into classes IV and V show overall good agreement with their eeF2•sordarin•GDP position and our density. There are differences in domain IV of eeF2 which is known to be dynamic and plays a key role in translocation[48,44]. We analyzed these differences qualitatively by comparing the Molecular Dynamics Flexible Fitting (MDFF) model (figure 6h-i) with the docked model solved with Sordarin present (figure 6j-k). The antibiotic Sordarin is highly specific for binding to fungal eeF2 and permits GTP hydrolysis, yet prevents conformational changes that result in subsequent release of eeF2 after translocation[49]. Sordarin is not present in the sample under study, yet there is a pronounced difference in electron density between domains III-V of eeF2 in class V (figure 6i black arrow) that coincides with the Sordarin binding pocket. This density is not present in class IV which also exhibits a rotation of eeF2-domain IV (figure 6j).

Discussion

We have created a set of image processing routines incorporated into the program emClarity, which have demonstrated a much greater accuracy in alignment and image restoration

compared to current state-of-the-art approaches using the same raw data sets which are publicly available. We have worked to make emClarity as easy to use as possible, limiting user specified parameters to the normal microscope and data collection information, as well as an estimate of the particle's radius and mass. The user must also select the angular search range, which is something that may be improved in the future.

In addition to a pronounced enhancement in overall resolution, we have demonstrated a powerful approach for image classification in the presence of the "missing-wedge" effect by combining the correction for wedge differences with multi-scale clustering which helps to encode biologically relevant information for the clustering algorithms. Having isolated classes IV and V of the mammalian ribosome *ex vivo* suggests both that Sordarin binding stabilizes an interaction between eeF2-domian III/V that exists on pathway in functional ribosomes, and that nearby intermediates on the energy landscape not observed in studies using the antibiotic, may be explored and observed by using this approach. In addition to isolating well-resolved class averages with minor populations, and finding nearby minima in the energy landscape, our approach also results in the production of accurate 3D-variance maps. By highlighting key regions of dynamic behavior, our approach should be useful for direct analysis and the design of complementary biophysical experiments. While these advances in classification are in the pre-processing and dimensionality reduction stage, future work to explore modern approaches in pattern recognition and machine learning, may allow another substantial improvement in the technique.

Acknowledgments:

We thank J. Frank and W. Li for very helpful discussions, D. Bevan for technical assistance with computer clusters, S. Loerch for the help with Phenix real-space refinement and COOT, and T. Brosenitsch, F. J. Alvarez, and J. P. Rickgauer for reading the manuscript. We thank F. Schur and J. Briggs for the HIV-1 immature Gag dataset, and X. Fu for testing the emClarity software. This work was supported by the National Institutes of Health (GM085043, GM082251) and the UK Wellcome Trust Investigator Award (206422/Z/17/Z).

Author contributions

This study was conceived and designed by P.Z and B.A.H. B.A.H. developed and tested the code for emClarity. B.A.H. and P.Z. analyzed the results. B.A.H. and P.Z. wrote the paper.

Competing financial interests:

The authors declare no competing financial interests.

References

1. Grant, T., Rohou, A. & Grigorieff, N. cis TEM , user-friendly software for single- particle image processing. 1–24 (2018).
2. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
3. Glaeser, R. M. & Hall, R. J. Reaching the information limit in cryo-EM of biological macromolecules: experimental aspects. *Biophys. J.* **100,** 2331–7 (2011).
4. Cheng, Y., Grigorieff, N., Penczek, P. A. & Walz, T. A primer to single-particle cryo-electron microscopy. *Cell* **161,** 439–449 (2015).
5. Oikonomou, C. M. & Jensen, G. J. Cellular Electron Cryotomography : Toward Structural Biology In Situ. *Annu Rev Biochem* 1–24 (2017). doi:10.1146/annurev-biochem-061516-044741

6. Diebolder, C. A., Koster, A. J. & Koning, R. I. Pushing the resolution limits in cryo electron tomography of biological structures. *J. Microsc.* **248,** 1–5 (2012).

7. Lučić, V., Rigort, A. & Baumeister, W. Cryo-electron tomography: The challenge of doing structural biology in situ. *J. Cell Biol.* **202,** 407–419 (2013).

8. Frangakis, A. S. *et al.* Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 14153–14158 (2002).

9. Bartesaghi, A. *et al.* Classification and 3D averaging with missing wedge correction in biological electron tomography. *J. Struct. Biol.* **162,** 436–450 (2008).

10. Bharat, T. A. M., Russo, C. J., Löwe, J., Passmore, L. A. & Scheres, S. H. W. Advances in Single-Particle Electron Cryomicroscopy Structure Determination applied to Sub-tomogram Averaging. *Structure* **23,** 1743–1753 (2015).

11. Cassidy, C. K. *et al.* CryoEM and computer simulations reveal a novel kinase conformational switch in bacterial chemotaxis signaling. *Elife* **4,** (2015).

12. Zeev-Ben-Mordehai, T. *et al.* Two distinct trimeric conformations of natively membrane-anchored full-length herpes simplex virus 1 glycoprotein B. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 4176–4181 (2016).

13. Penczek, P. A., Frank, J. & Spahn, C. M. T. A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. *J. Struct. Biol.* **154,** 184–194 (2006).

14. Liao, H. Y., Hashem, Y. & Frank, J. Efficient Estimation of Three-Dimensional Covariance and its Application in the Analysis of Heterogeneous Samples in Cryo-Electron Microscopy. *Structure* **23,** 1129–1137 (2015).

15. Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B. & Schulten, K. Molecular dynamics flexible fitting: A practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **49,** 174–180 (2009).

16. Schur, F. K. M., Hagen, W. J. H., de Marco, A. & Briggs, J. a G. Determination of protein structure at 8.5Å resolution using cryo-electron tomography and sub-tomogram averaging. *J. Struct. Biol.* **184,** 394–400 (2013).

17. Kudryashev, M., Castaño-Díez, D. & Stahlberg, H. Limiting Factors in Single Particle Cryo Electron Tomography. *Comput. Struct. Biotechnol. J.* **1,** 1–6 (2012).

18. Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116,** 71–6 (1996).

19. Fernández, J. J., Li, S. & Crowther, R. a. CTF determination and correction in electron cryotomography. *Ultramicroscopy* **106,** 587–96 (2006).

20. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192,** 216–221 (2015).

21. Hrabe, T. *et al.* PyTom : A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *J. Struct. Biol.* **178,** 177–188 (2012).

22. Jensen, G. J. & Kornberg, R. D. Defocus-gradient corrected back-projection. *Ultramicroscopy* **84,** 57–64 (2000).

23. Turoňová, B., Schur, F. K. M., Wan, W. & Briggs, J. A. G. Efficient 3D-CTF correction for cryo-electron tomography using NovaCTF improves subtomogram averaging resolution to 3.4Å. *J. Struct. Biol.* (2017). doi:10.1016/j.jsb.2017.07.007

24. Kunz, M. & Frangakis, A. S. Three-dimensional CTF correction improves the resolution of electron tomograms. *J. Struct. Biol.* **197,** 114–122 (2017).

25. Rickgauer, J. P., Grigorieff, N. & Denk, W. Single-protein detection in crowded molecular environments in cryo-EM images. *Elife* **6,** 1–22 (2017).

26. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *Elife* **4,** e06980 (2015).

27. Förster, F., Pruggnaller, S., Seybert, A. & Frangakis, A. S. Classification of cryo-electron sub-tomograms using constrained correlation. *J. Struct. Biol.* **161,** 276–286 (2008).

28. Stewart, A. & Grigorieff, N. Noise bias in the refinement of structures derived from single particles. *Ultramicroscopy* **102,** 67–84 (2004).

29. Henderson, R. *et al.* Outcome of the first electron microscopy validation task force meeting. *Structure* **20,** 205–214 (2012).

30. 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).

31. Sindelar, C. V. & Grigorieff, N. Optimal noise reduction in 3D reconstructions of single particles using a volume-normalized filter. *J. Struct. Biol.* **180,** 26–38 (2012).

32. Scheres, S. H. w. Beam-induced motion correction for sub-megadalton cryo-EM particles. *Elife* **3,** e03665 (2014).

33. Heumann, J. M., Hoenger, A. & Mastronarde, D. N. Clustering and variance maps for cryo-electron tomography using wedge-masked differences. *J. Struct. Biol.* **175,** 288–299 (2011).

34. Alsberg, B. K. Multiscale cluster analysis. *Anal. Chem.* **71,** 3092–3100 (1999).

35. Marabini, R. *et al.* The Electron Microscopy eXchange (EMX) initiative. *J. Struct. Biol.* **194,** 156–163 (2016).

36. Bharat, T. A. M. & Scheres, S. H. W. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. *Nat. Protoc.* **11,** 2054–2065 (2016).

37. Khoshouei, M., Pfeffer, S., Baumeister, W., Förster, F. & Danev, R. Subtomogram analysis using the Volta phase plate. *J. Struct. Biol.* **197,** 94–101 (2017).

38. Bai, X. C., Fernandez, I. S., McMullan, G. & Scheres, S. H. W. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife* **2013,** 2–13 (2013).

39. Schur, F. K. M. *et al.* An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation. *Science (80-.).* **353,** 506–508 (2016).

40. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66,** 213–221 (2010).

41. Gutell, R. R., Weiser, B., Woese, C. R. & Noller, H. F. Comparative anatomy of 16-S-like ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **32,** 155–216 (1985).

42. Beckmann, R. *et al.* Architecture of the Protein-Conducting Channel Associated with the Translating 80S Ribosome. *Cell* **107,** 361–372 (2001).

43. Mohan, S. & Noller, H. F. Recurring RNA structural motifs underlie the mechanics of L1 stalk movement. *Nat. Commun.* **8,** 14285 (2017).

44. Spahn, C. M. *et al.* Domain movements of elongation factor eEF2 and the eukaryotic 80S ribosome facilitate tRNA translocation. *EMBO J.* **23,** 1008–1019 (2004).

45. Wilson, D. N. & Nierhaus, K. H. The E-site story: the importance of maintaining two tRNAs on the ribosome during protein synthesis. *Cell. Mol. Life Sci.* **63,** 2725–2737 (2006).

46. Budkevich, T. V. *et al.* Regulation of the Mammalian Elongation Cycle by Subunit Rolling: A Eukaryotic-Specific Ribosome Rearrangement. *Cell* **158,** 121–131 (2014).

47. Abeyrathne, P. D., Koh, C. S., Grant, T., Grigorieff, N. & Korostelev, A. A. Ensemble cryo-EM uncovers inchworm-like translocation of a viral IRES through the ribosome. *Elife* **5,** 1–31 (2016).

48. Gomez-Lorenzo, M. G. *et al.* Three-dimensional cryo-electron microscopy localization of EF2 in the Saccharomyces cerevisiae 80S ribosome at 17.5 A resolution. *EMBO J.* **19,** 2710–8 (2000).

49. Chakraborty, B., Mukherjee, R. & Sengupta, J. Structural insights into the mechanism of translational inhibition by the fungicide sordarin. *J. Comput. Aided. Mol. Des.* **27,** 173–184 (2013).

Figure legends

Figure 1. The emClarity workflow for sub-tomogram averaging and classification. The most significant improvements introduced in emClarity are highlighted in red text, while novel additions to the process are indicated in orange boxes. Dashed lines indicate optional branches which may be included any number of times during the iterative alignment.

Figure 2. Tomo-CPR. (a) Schematic overview for reference generation in the Tomo-CPR. (b-c) Cartoons illustrating overlapping information in the projections arising from other particles,

components in the specimen, and variable defocus as a function of tilt. (d-e) Examples of non-tilted (d) and tilted (e) projections used to generate references for the yeast 80s tomo-CPR. Overlapping particles and features due to contaminants (white arrows) and the carbon edge (white chevron) are shown. Scale bars 30 nm.

Figure 3. Compensation of the missing wedge at multiple length scales. Far left, the total average filtered by Gaussian kernels of variable width to correlate voxels over the given length scales. Center, eigen-images composed of the average plus the eigenvector, sorted from the most variance explained (1) to least (21). Black and white arrows show example where es27 density is absent or present in the L1 position. Right three columns show 3D-variance maps in red over-laid with the average: without missing-wedge compensation (a-c), with binary wedge-masked-differences (WMDs) (d-f) and with 3D-sampling function estimated differences in emClarity (g-i). Positions for L1 (green arrow), es27 (orange arrow), e-Site tRNA, and the mRNA channel entrance (blue arrow) are labeled.

Figure 4. Improvement in resolution of sub-tomogram averaging using emClarity. (a-b) Comparison of the sub-tomogram average of yeast 80s ribosome (a) and of rabbit 80s ribosome (b) by RELION (EMD-3228) (a, left) or by pyTOM (EMD-3420) (b, left) and by emClarity (right). Arrow points to an additional feature only revealed in emClarity. Scale bar, 100 Å. (c) Cross-FSC between the sub-tomogram averages by emClarity and the SPA cryoEM map (EMD-2275) of yeast 80s ribosome, each accumulating the previous improvement: original orientation parameters from Relion (black), CTF estimation and correction with the optimal exposure filter (magenta), one round of tomo-CPR (green), per-tilt defocus estimation (cyan), consideration of resolution anisotropy (red), and with all features plus alignment in emClarity (blue). Right, representative views of sub-tomogram averages, with the frame color matching the plot colors, and a rigid body docking of yeast 80s atomic model (PDB-47VR). The arrow and chevron highlight the resolved alpha helices and RNA structures, respectively. These experiments were repeated at least three times with nominally identical results. (d) Comparison of the sub-tomogram average of HIV-1 immature CA-SP1 monomer (EMD-3782) (left) and by emClarity (right). (e) Enlarged views of boxed area in d overlaid with a real-space refined model.

Figure 5. Classification of yeast 80s ribosome (EMPIAR-10045) with 3D-CTF compensated missing-wedge and multi-scale PCA. (a-e) Four major classes and a minor class contributing 96.9% of sub-tomograms are shown with number and percentage of contributing units and resolution indicated below. The highly dynamic L1 protuberance and RNA expansion-segment 27 are shown in green and orange, respectively. Scale bar, 100 Å. (f-h) Enlarged views, with the small-subunit removed for clarity, of the L1 protuberance (green) in an intermediate position (f, class a), fully closed (g, class b), and fully open (h, classes c-e) respectively. The riboprotein and selected RNA helix components of the L1 protuberance (rpL1,h76,h79 from PDB 3J78) shown in magenta ribbon after rigid body docking into the respective density. (i) A cartoon illustrates the region displayed in (f-h) from the inter-subunit space with the E,P,A sites labelled and L1 (green) E-site tRNA (orange) and mRNA channel (purple ribbon.)

Figure 6. Classification of translating mammalian 80s ribosome with 3D-CTF compensated missing-wedge and multi-scale PCA. (a-c) Classes I-III represent a post translocational state with the co-factors shown in the lower row from the inter-subunit surface with the 60s subunit removed for clarity. (d-e) Classes IV-V have a mid-rotated 40s and a swiveled head corresponding to a pre-translocational intermediate. Scale bar represents 100 Å. (f-g) Class IV in dark blue overlaid with class III in gold, showing the mid-rotated 40s state. (h-i) MDFF of eeF2 (orange) with the density from class-IV (h) and class-V (i) starting from PDB-4ujo (cyan ribbon). (j-k) The rigid body docking of PDB-4ujo into the density from class IV/V, respectively. Arrows point to this density which is occupied by the antibiotic Sordarin in PDB-4ujo, but is not present in the sample used in this study.

ONLINE METHODS

Please refer to "Life Sciences Reporting Summary" for detailed information on experimental design, reagents and software.

Datasets

The datasets used in emClarity processing are from Electron Microscopy Public Image Archive (EMPIAR), including the yeast 80s ribosome (EMPIAR-10045), the mammalian 80s ribosome (EMPIAR-10064), and the HIV-1 immature Gag (EMPIAR-10164).

emClarity programs

emClarity is run from the command line and is easily scripted to run in a manner most suited to a user's particular project. A text parameter file is used to input project specific details, like microscope parameters, mask dimensions, and angular search ranges. We typically make a copy of the parameter file for each cycle of averaging and alignment, which we refer to as paramC.m, where C refers to the cycle number. The meta-data of each project is tracked in a binary database which named using the "subTomoMeta" parameter. Each tilt-series may have multiple areas slated for reconstruction, tiltN M refers to tilt-series "N" and reconstruction area "M". A brief description of the major functions (*in italic*) in emClarity is below:

emClarity init paramN.m

Read in the desired dimensions for each sub-region of each tilt-series to reconstruct, initialize the subTomoMeta (metadata binary.)

emClarity ctf estimate tiltN

estimate the ctf for a given tilt-series N

emClarity reScale mapNameIN mapNameOUT AngPixIN AngPixOUT cpu/GPU

resample a map to a new pixel size, particularly for template matching.

emClarity templateSearch tiltN M reference.mrc symmetry gpuIDX

Reconstruct tomogram M from tilt-seris N without ctf correction, run template matching on GPU # gpuIDX, randomize results at symmetry related positions.

emClarity ctf 3d paramN.m

Run 3d-CTF corrected weighted-back projection.

emClarity avg paramN.m N RawAlignment

Every cycle begins by creating a sub-tomogram average, calculating the "gold-standard" FSC, and weighting the average accordingly while compensating for amplitude attenuation by the CTF to produce references for the alignment.

emClarity alignRaw paramN.m N

Run the alignment

emClarity removeDuplicates paramN.m N

clean out any subtomograms that have drifted to the same position. Not needed every cycle.

emClarity tomoCPR paramN.m N

Run tomogram constrained particle refinement, this is generally done prior to a step down in binning, e.g. bin4 → bin3.

emClarity ctf update paramN.m N

Only needed after a run of tomoCPR, this updates the tilt-series geometry in the subTomoMeta, and also resamples the raw tilt-series applying rotation, shift and magnification scaling all in Fourier space to reduce interpolation losses of high resolution information.

Since the tilt-series alignments are update, and usually also the binning is reduced, a new round of 3D-CTF reconstructions need to be made with.

If classification is to be run, the cycle starts the same, but with the "flgClassification" parameter enabled.

emClarity avg paramN.m N

emClarity pca paramN.m N previousPCA

Run 3D-sampling function compensated PCA at each length scale specified in the "pcaScaleSpace" parameter. The command line argument previousPCA is always zero on the first run. If a random subset (25% or ~3000, whichever is larger) is to be analyzed by setting "Pca_randSubset" then a subsequent round of pca must be run with "previousPCA" set to one in order to project all of the sub-tomograms along the principal component axes.

emClarity cluster paramN.m N

Cluster the data based on selected eigenvectors from the pca step.

emClarity avg paramN.m N Cluster_cls

Notice the last argument (a string) which creates a montage of the class averages selected in the parameter file. Classes with different memberships may be selected.

At the end of processing, the half-sets may be aligned and combined by running:

emClarity avg paramN.m N FinalAlignment

Align and combine half-sets, optionally creating multiple, differently sharpened maps.

Image processing

The alignment and classification procedures are generally identical for all the samples, except for the HIV-1 Gag data, which were not classified and had C6 symmetry applied. All parameters unique to each dataset, including the angular search range and iterations used, are shown in Table S3 and S4. emClarity is only tested on Linux operating systems, and all references to command line operations are to be understood in that manner.

Project set-up and coarse tilt-series alignment:

For each specimen, we make a project directory, which we will refer to generically as "projectDir." During processing, several sub-directories will be created by emClarity in addition to the user created directories for the raw data, which we recommend calling "rawData" and a folder for the cleaned data that *must be* named "fixedStacks."

The HIV-1 Gag data consist of dose fractionated frames, which we aligned using the program "unblur" version 1.0 included in the cisTEM package. The aligned frames were summed and saved *without* any exposure-based filtering because this is handled later inside emClarity.

All tilt-series were aligned using the default parameters in IMOD version 4.10.12 using the eTomo interface, with the available gold-fiducial markers. For the ribosome datasets, all gold fiducials (~5-7/ tilt) were selected, while for the HIV-1 Gag data ~20-30 closest to the protein and distributed on both surfaces of the ice were selected. Local alignments with fixed XYZ global coordinates were run for the HIV-1 Gag data only. After generating the final aligned stack, the gold beads were located using *find beads3d*. Only the fiducial model describing the location of the beads is needed, so they were not erased. Note: for EMPIAR 10045 the pixel size in the header must be corrected to 2.17 Å prior to beginning. This may be done with the IMOD program *alterheader* from the command line.

The files describing the projection transformations, any local alignments, and fitted tilt-angles are copied to the fixedStacks directory and renamed.

```
>$ mv specimen_name_1_fid.xf        projectDir/fixedStacks/tilt1.xf

>$ mv specimen_name_1_fid.tlt       projectDir/fixedStacks/tilt1.tlt
```

>$ mv specimen_name_1_local.xf projectDir/fixedStacks/tilt1.local

>$ mv specimen_name_1_erase.fid projectDir/fixedStacks/tilt1.erase

Additionally, a file listing the tilt-angles in the order they were collected must be created for emClarity to apply an appropriate exposure filter.

projectDir/fixedStacks/tilt1.order

If outlier pixels are removed in IMOD, this "fixed" stack may be moved to projectDir/fixedStacks/tilt1.fixed, otherwise you may just link the raw data.

>$ cd projectDir/fixedStacks

>$ ln -s ../rawData/specimen_name_1.st tilt1.fixed

This is repeated for all tilts-series, of which there are 7, 4, and 41 in the yeast, mammalian, and HIV-1 Gag data sets respectively.

CTF estimation:

The mean defocus at the tilt-axis was then estimated in emClarity for each tilt-series using a 3.5 ± 2.5 μm window covering the range of expected defocus values for all three data sets.

For the HIV-1 Gag data, the per-tilt defocus was determined using "emClarity ctf refine" to produce the power-spectra, which were subsequently fit using ctffind4 with the –amplitude-spectrum input flag and default parameters.

For the yeast ribosome data which have a thin layer of carbon providing extra signal in the power spectrum, the per-tilt defocus values were refined during tomoCPR. To do so, the height of the cross-correlation peak is maximized by scanning through a small range of defocus values as applied to each reference tile[50].

Selecting sub-regions for further analysis

603 The selection of sub-regions of each tilt-series for reconstruction are defined by a text file
604 with the minimum and maximum values in x, y, z for each region. The script "recScript2.sh"
605 provided with emClarity was used to first create reconstructions of each tilt-series at a binning of
606 10 and thickness of 300 covering the full X,Y dimension of the images. Each region is then
607 defined while viewing the reconstruction in IMOD by making an IMOD model with six points
608 per region, xmin, xmax, ymin, ymax, zmin, zmax, in that order.

609 A second run of "recScript2.sh" creates a projectDir/recon directory and converts these
610 model files into the text files read in by emClarity to be used for the rest of the procedure. These
611 are called tilt1_recon.coords and list the tilt-series base name, number of regions to reconstruct,
612 and for each region the width, first and last slice in y, thickness, x-origin offset and z-origin
613 offset.

614 The ribosome data were divided on the x-axis into two regions per tilt-series. The HIV-1
615 Gag data were divided into quadrants. Additionally, the flag "fscGoldSplitOnTomos=1" is set in
616 the parameter file for the HIV-1 Gag data, so that the even/odd half sets are divided based on
617 tomogram, not randomly on sub-tomograms. This is necessary to avoid mixing neighboring
618 particles which would violate the gold-standard hypothesis.

619 *Template matching*

620 References were derived from SPA EMD-3228[39] (yeast 80s ribosome), EMD-5592[51]
621 (human 80s ribosome) and EMD-8403 (HIV-1 Gag)[52] and rescaled to the full pixel size of each
622 data set using "emClarity reScale." These references were then passed to "emClarity
623 templateSearch" binned to achieve a nominal pixel size ~ 8-12 Å depending on the size of the
624 specimen. All maps and tomograms are automatically low-pass filtered to 40 Å resolution by
625 default in emClarity. Non-CTF corrected tomograms are reconstructed by the templateSearch
626 program as needed for template matching.

627 The results for the ribosome data set were cleaned manually by comparing the maximum
628 intensity projection maps and the binned tomograms overlaid with an IMOD model showing the
629 x,y,z coordinates of each peak detected.

For the HIV-1 Gag data, emClarity removeNeighbors was used to automatically clean the results based on geometrical restraints. Only peaks that had five neighbors within 100 Å and also oriented within 20˚ were retained, resulting in 179,168 sub-tomograms to start. (This number dropped to 162,213 in the first round of averaging as particles too close to the edge to allow padding by 1.5 x particleRadius were excluded.

Particles with symmetry pose a special challenge to all missing-wedge compensation approaches as any error in the compensation will result in the particle looking different at its symmetry related orientations. To help with this, we randomize the template matching results around each symmetry related orientation, and subsequently only search an angular range small enough to not reach the neighboring positions.

Iterative alignment

Each cycle of alignment is initiated by calculating averages of the two half sets, calculating the gold-standard FSC, and then applying re-weighting each average to generate a FOM weighted reference.

We alternate searching over just the azimuthal and polar angles, and an in-plane search. For each specimen, we started at a binning of chosen to produce a pixel size of ~7-8 Å. Details may be found in table S3 for the ribosome data sets, and table S4 for the HIV-1 gag data. We then go through three rounds of averaging and alignment, followed by removing any positions that may have drifted to overlap using "emClarity removeDuplicates." We then run a round of tomo-CPR, which requires updating the aligned tilt-series and the 3dCTF corrected tomograms.

```
>$ emClarity ctf update paramX.m
>$ emClarity ctf 3d paramX.m
```

This reconstruction is generated at a binning one finer than the previous, and the same pattern was repeated until reaching full sampling.

Classification

The ribosome data for the yeast 80s were classified in a single pass, using three resolution bands 10,18, and 28Å, 36 of the top eigenvalues were saved, and five from each band were selected (parameter Pca_coefficients=[7:11;7:11;1:11])] for clustering via kmeans The class averages were then generated by running emClarity avg paramX.m X cluster_cls

The ribsome data for the mammalian 80s were classified in two passes. First, they were split into groups displaying either a rotated or un-rotated 40s small subunit. To do this, the subTomoMeta file (projectName.mat) was copied to two new files: project_smallSU.mat and project_largeSU.mat. The classes are selected for removal by viewing the class average montage in IMOD, and selecting any point in the region of a given class. These models are then used to remove their contributing members in the subTomoMeta.

>$ emClarity geometry paramX.m X RemoveClasses [X,0,0] STD.

Since both branches of the project access the same raw data, it is convenient to remain in the same project directory, and all subsequent output will be identified by the new subTomoMeta basename.

A subsequent round of classification was run using 12,22,32 Å resolutions. Unlike the yeast 80s which had some Eigen images with clear missing wedge bias, revealed as “streakiness” in the density, the mammalian displayed sufficient true variability to overpower the noise from the missing-wedge bias, and all 36 eigenvectors from each resolution band were used in clustering.

Analysis

Models PDB-3J78 for yeast were rigid body docked in using Chimera.

Models PDB-4UJO for mammalian were docked in using Chimera, in combination with the “Segger” plugin.

Models PDB-5l93 were docked in using Chimera, refined in real-space using Phenix version 1.13-2998-000, and manually edited in COOT version 0.8.9.

Algorithmic details

3D-Sampling Function

$$3D\ Sampling\ Function\ \equiv SF^{3D} = \sum_{j=1}^{S}\sum_{i=1}^{Z} T^{i,j}\left|CTF_i^{2d}\right|^2 R^{2d} ExpFilter_i^{2d} \quad eq\ 1.$$

The first term in the summation is the combined transformation of projection (i) into the tomogram, and sub-tomogram (j) into the final average. The second term is the normal expression for the CTF limited to third order Seidel aberrations[53], the third is the R-weighting, and the fourth is the optimal-exposure filter as defined[26], Z is the number of projections in each tilt-series, and S the number of sub-tomograms.

Refinement of tilt-series alignment

Tomo-CPR works by combining the strengths of the 3D-model-based and feature-tracking approaches, while also taking advantage of the robust alignment tools developed for gold-fiducial alignment available in the IMOD package, which must be installed alongside emClarity. Starting with the tomogram (as in the 3D model approach), we additionally replace the density corresponding to our particles of interest at the proper orientation, with a copy of the high SNR sub-tomogram average and then re-project that synthetic tomogram using the IMOD program *tilt* (Figure 2a). This re-projection also includes any local alignments previously determined and allows us to create a reference tilt-series along with a model for each sub-tomogram position in the 2D-projection. This model is used to cut tiles out of the data and reference projections for comparison by cross-correlation, while considering the CTF of the data projection at that point, as well as the structural noise from each particle's unique environment (supplemental figure 1b and 1c). These refined positions are then used as input to IMOD's *tiltalign* as if they were derived from gold fiducials, allowing us to take advantage of local refinement and robust fitting, as described previously[54]. The global changes to the projection geometry are applied to the tilt-series, while the local refinements are taken into consideration when the tomograms are reconstructed on the fly by emClarity. In addition to the importance of considering neighboring particles (Figure 2d), additional high-contrast features, like the edge of

the carbon foil or other particulate matter are pointed out in figure 2e, where a thin strip of one of the tomo-CPR references, prior to tiling, is shown.

Statistical optimization of the SNR in the final map

In addition to alternating phase reversals, the CTF also modulates the amplitudes of the data. Several programs for correcting the CTF phase and amplitude modulations directly in the 2D projections of tilted images are available; the two predominant being CTFPLOTTER and CTFPHASEFLIP[55] included in the IMOD package[18] for measurement and correction respectively, and TOMOCTF[19]. Both may be used to restore the amplitudes in the Fourier transforms of individual projections which inevitably amplifies noise in the process[56]. A more attractive approach is to correct the phases on the projections, via multiplication by the CTF, and then to address the amplitudes after building the 3D reconstruction (sub-tomogram average).

The amplitude modulations are compensated using a Wiener filter in both SPA and the adaptation of RELION for sub-tomogram averaging[10]. A typical Wiener filter based approach has also been described recently in the structure of the HIV-1 capsid protein[39] but is only implemented through "in-house" software. In addition to the CTF and consideration of increasing sample thickness with tilt angle, our 3D-sampling function also takes into consideration the R-weighting that is applied during tomogram reconstruction and is applied via our adaption of the "volume normalized single-particle Wiener (original eq. 8)[31] We include the original expression, with our nomenclature in eq. s1.

$$F^{SPW_t}(\boldsymbol{q}_{hkl}) = \frac{SF^{3D}}{SF^{3D} + \frac{f_{particle}}{f_{mask}}\left(\frac{1 - FSC_{mask}(\boldsymbol{q}_{hkl})}{2FSC_{mask}(\boldsymbol{q}_{hkl})}\right)\left(\frac{1}{n_q}\sum_{q \in \boldsymbol{q}_{h'k'l'}} SF^{3D}\right)} F^{LSQ}(\boldsymbol{q}_{hkl}) \qquad \boldsymbol{eq\ s1}$$

The least squares estimate, which is a Wiener filtered reconstruction with an ad-hoc Wiener constant [CITE] is defined below eq s2.

$$F^{LSQ}(\boldsymbol{q}_{hkl}) = \frac{\sum_{j=1}^{S}\sum_{i=1}^{Z} T^{i,j}\left|CTF_i^{2d}\right|^2 R^{2d} ExpFilter_i^{2d} F_i^{2d}}{\sum_{j=1}^{S}\sum_{i=1}^{Z} T^{i,j}\left|CTF_i^{2d}\right|^2 R^{2d} ExpFilter_i^{2d} F_i^{2d} \quad + \quad 1} \qquad \boldsymbol{eq.s2}$$

We have made three major changes to the filter:

1) The 3D-Sampling function is weighted for critically under-sampled regions, where the SSNR estimated by the FSC is less reliable. This is done by:
    a) choosing a minimum acceptable sampling threshold, 0.2 * median (SF^{3D} != 0)
    b) scaling SF^{3D} to replace less sampled regions by smoothly transition from this value to some new larger number, chosen by the maximum in the original SF^{3D}.
2) The FSC, normally calculated over spherical sections is replaced by an anisotropic FSC calculated over conical sections, typically 38, which has been used previously to estimate resolution anisotropy[57].
3) Finally, the average sampling over spherical shells (final term in the denominator of eq s1) that is used to scale the SSNR estimate to represent the average SSNR in a single sub-tomogram is replaced with a gaussian smoothed version of the 3D-sampling function. Again to account for anisotropy in the sampling.

$$F^{SPW_t}(\boldsymbol{q}_{hkl}) = \frac{|SF^{3D}|^2}{|SF^{3D}|^2 + \frac{f_{particle}}{f_{mask}}\left(\frac{1 - FSC_{aniso,mask}(\boldsymbol{q}_{hkl})}{2FSC_{aniso,mask}(\boldsymbol{q}_{hkl})}\right)(G \otimes |SF^{3D}|^2)} F^{LSQ}(\boldsymbol{q}_{hkl}) \qquad \boldsymbol{eq.s3}$$

3D-sampling function compensated classification

For the special case where sub-tomograms are all oriented similarly, being adsorbed to a lipid monolayer for example, they may be averaged along the direction of their missing wedge and classified in 2D[58]; however, this is obviously of limited interest for most specimen. Another popular approach involves classifying the

constrained-cross-correlation matrix[9,27], however, this can be expensive to calculate and also disregards large amounts of information – discussed in detail in the paper on wedge masked differences (WMDs)[33] on which we further expand and enhance.

The WMDs approach seeks to compensate the missing wedge by forming the difference between a given particle and its expected value. The expected value is estimated to be the global average distorted by the particle's missing wedge. These are mean centered, normalized to a variance of one, and arranged into a 2D matrix followed by singular-value decomposition (SVD). The binary wedge used in this approach is only a first approximation, and we replace it with our full 3D-sampling function. This correction allows the classification to include higher-resolution details than previously possible, which is a necessary but not sufficient condition to achieve the classification we report. We find that including the highest variance information from three to four discrete length scales at the same time is required to observe each class.

Data Availability

Cryo-EM structural data have been deposited in the EM Data Bank under accession codes EMD-8799 for the yeast 80s ribosome, EMD-8802, EMD-8803, EMD-8804, EMD-8805, and EMD-8806 for rabbit 80s ribosome classes I-V respectively, EMD-8986 for the HIV-1 Gag data.

Code availability

The software is freely available from https://www.github.com/bHimes/emClarity

Tutorial documentation and videos at https://www.github.com/bHimes/emClarity/wiki

Methods-only References

50. Meyer, R. R., Kirkland, A. I. & Saxton, W. O. A new method for the determination of the wave aberration function for high-resolution TEM. 2. Measurement of the antisymmetric aberrations. *Ultramicroscopy* **99,** 115–123 (2004).

51. Anger, A. M. *et al.* Structures of the human and Drosophila 80S ribosome. *Nature* **497,** 80–85 (2013).

52. Ning, J. *et al.* In vitro protease cleavage and computer simulations reveal the HIV-1 capsid maturation pathway. *Nat. Commun.* **7,** 13689 (2016).

53. Fernando, K. V. & Fuller, S. D. Determination of astigmatism in TEM images. *J. Struct. Biol.* **157,** 189–200 (2007).

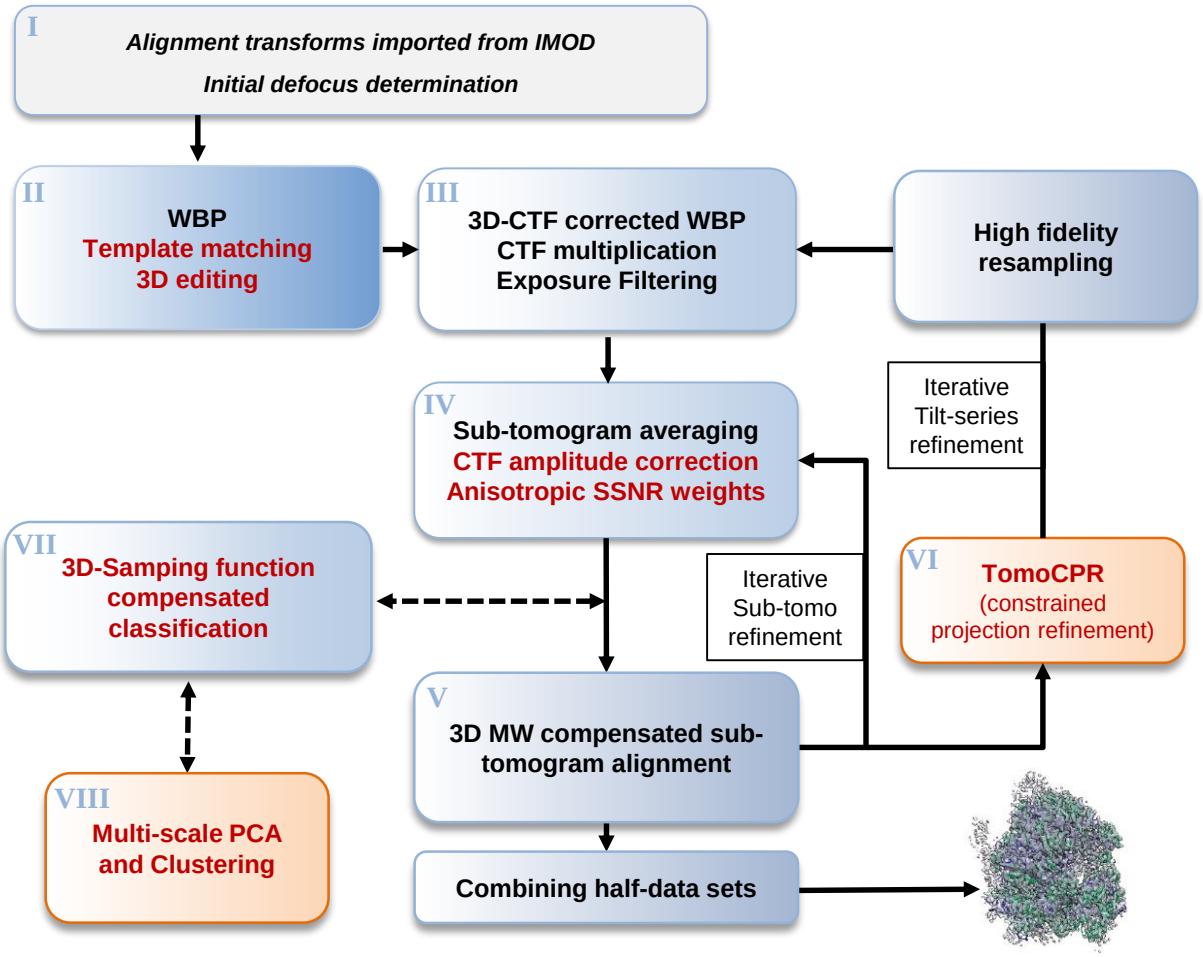
54. Mastronarde, D. N. Fiducial Marker and Hybrid Alignment Methods for Single- and Double-axis Tomography. in *Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell* (ed. Frank, J.) 163–185 (Springer New York, 2006). doi:10.1007/978-0-387-69008-7_6

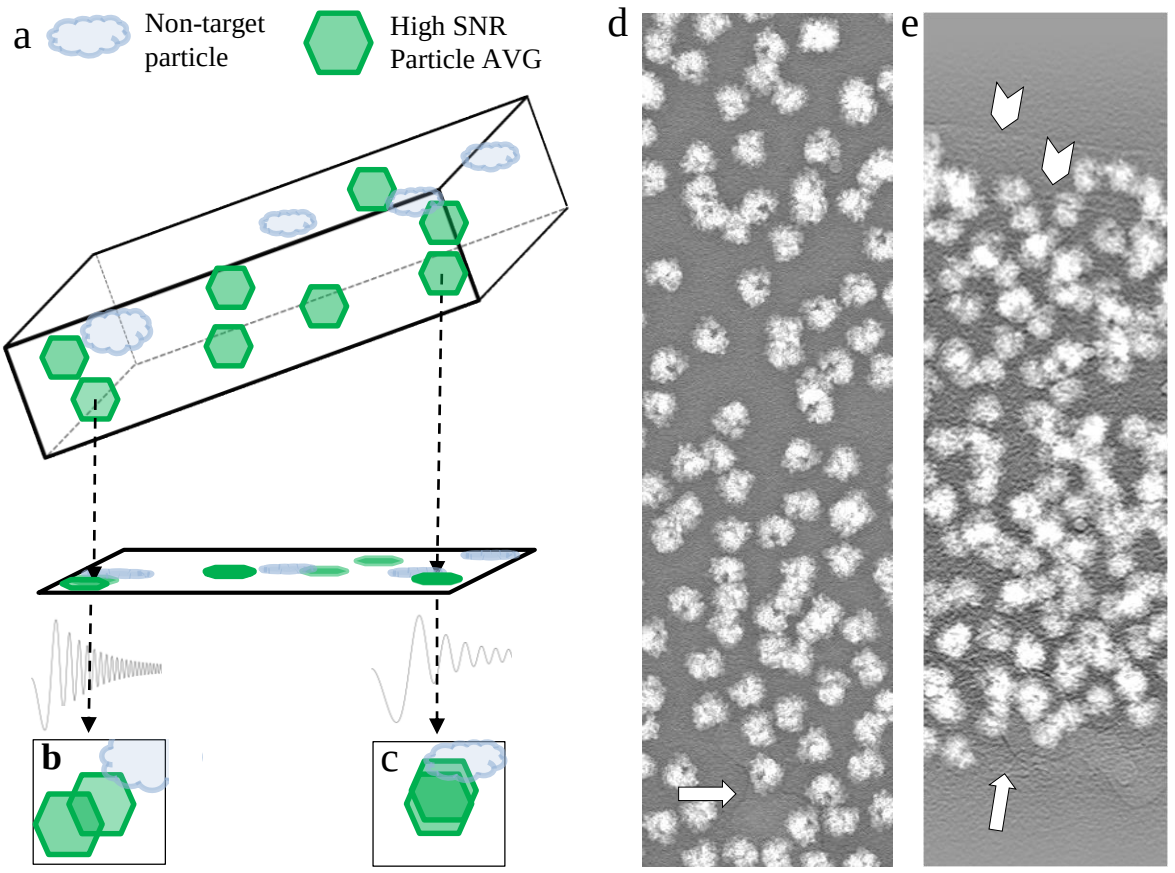
55. Xiong, Q., Morphew, M. K., Schwartz, C. L., Hoenger, A. H. & David, N. CTF Determination and Correction for Low Dose Tomographic Tilt Series. *J. Struct. Biol.* **168,** 378–387 (2010).

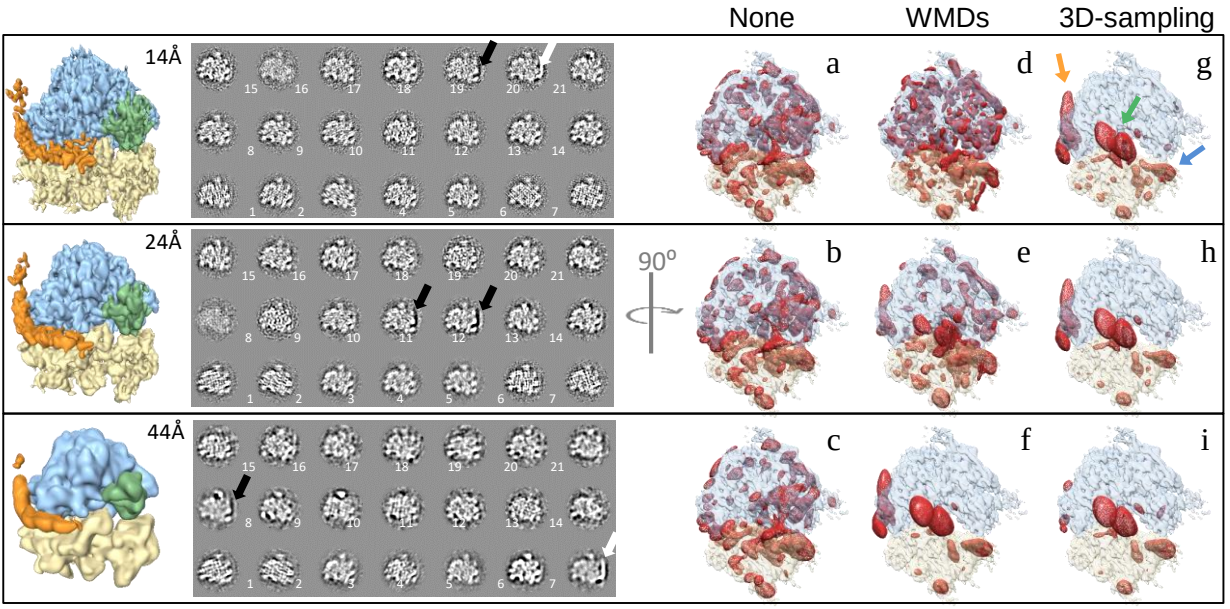
56. Frank, J. Electron Microscopy of Macromolecular Assemblies. in *Three-Dimensional Electron Microscopy of Macromolecular Assemblies* 15–69 (Oxford University Press, 2006).

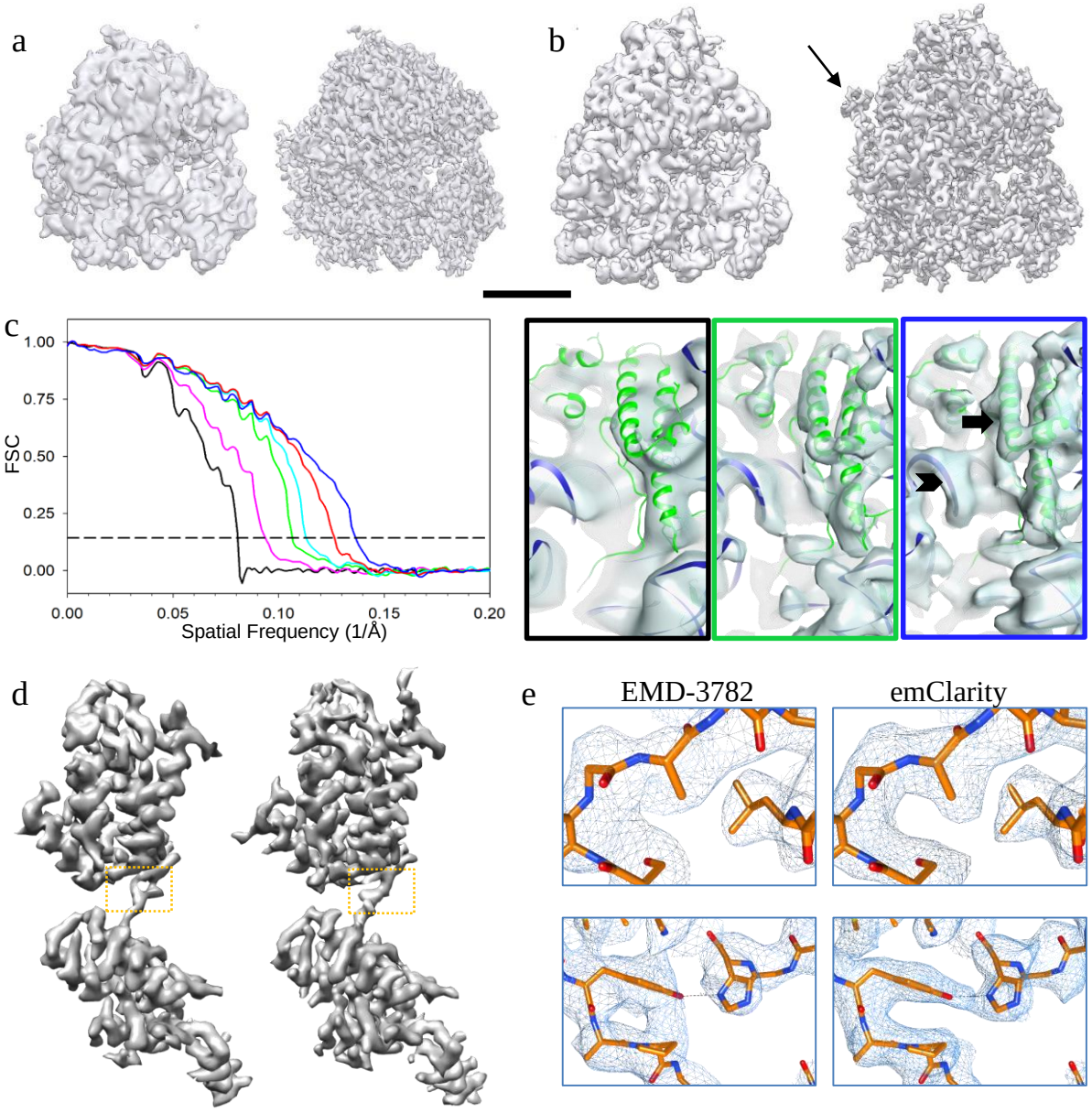
57. Diebolder, C. A., Faas, F. G. A., Koster, A. J. & Koning, R. I. Conical fourier shell correlation applied to electron tomograms. *J. Struct. Biol.* **190,** 215–223 (2015).

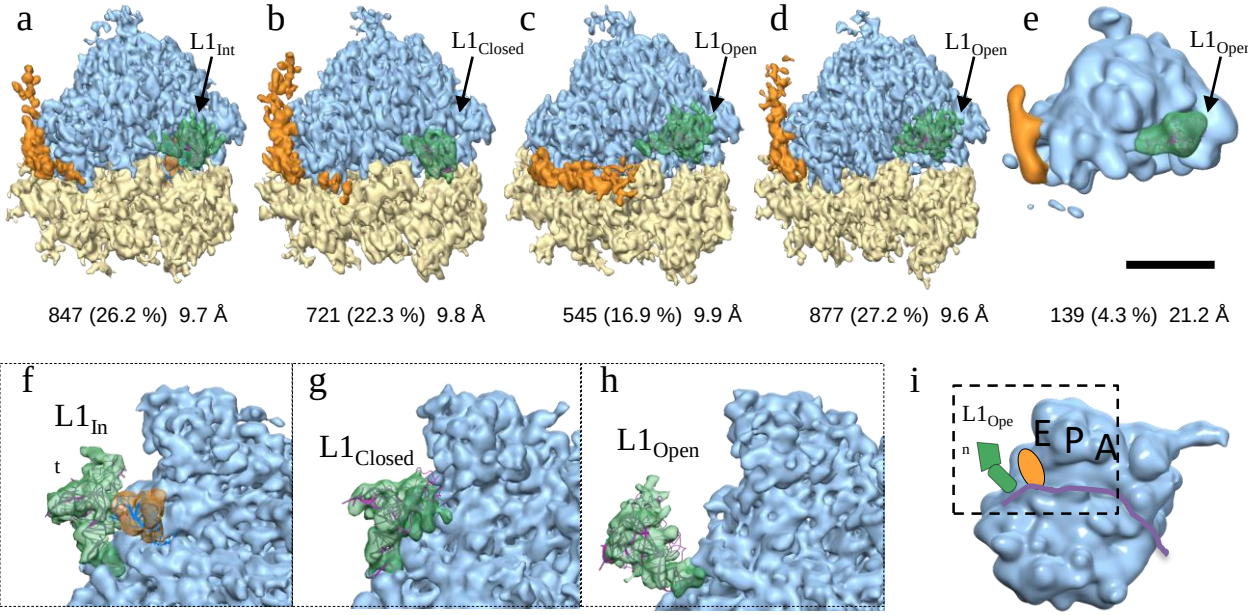
58. Winkler, H. *et al.* Tomographic subvolume alignment and subvolume classification applied to myosin V and SIV envelope spikes. *J. Struct. Biol.* **165,** 64–77 (2009).


I
Alignment transforms imported from IMOD
Initial defocus determination
II
WBP
Template matching
3D editing
III
3D-CTF corrected WBP
CTF multiplication
Exposure Filtering
High fidelity resampling
Iterative Tilt-series refinement
IV
Sub-tomogram averaging
CTF amplitude correction
Anisotropic SSNR weights
VII
3D-Samping function compensated classification
Iterative Sub-tomo refinement
VI
TomoCPR
(constrained projection refinement)
V
3D MW compensated sub-tomogram alignment
VIII
Multi-scale PCA and Clustering
Combining half-data sets


a
Non-target particle
High SNR Particle AVG
b
c
d
e

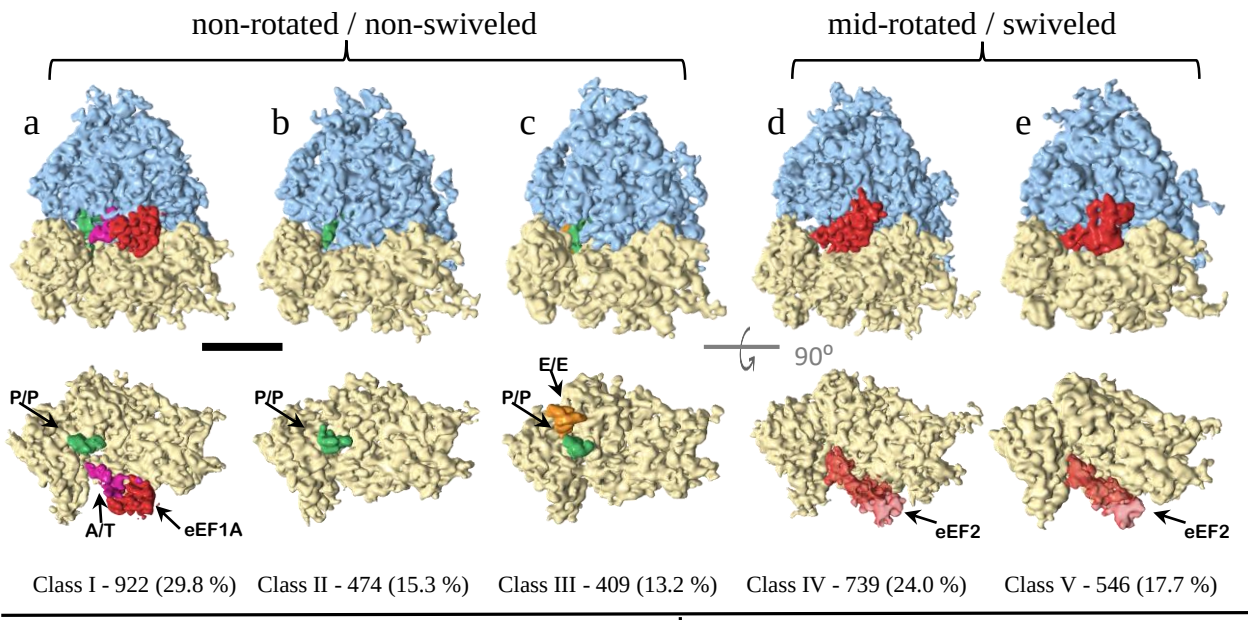

None
WMDs
3D-sampling
14Å
24Å
44Å
90°
a
b
c
d
e
f
g
h
i


a
b
c
1.00
0.75
0.50
0.25
0.00
FSC
0.00
0.05
0.10
0.15
0.20
Spatial Frequency (1/Å)
d
e
EMD-3782
emClarity

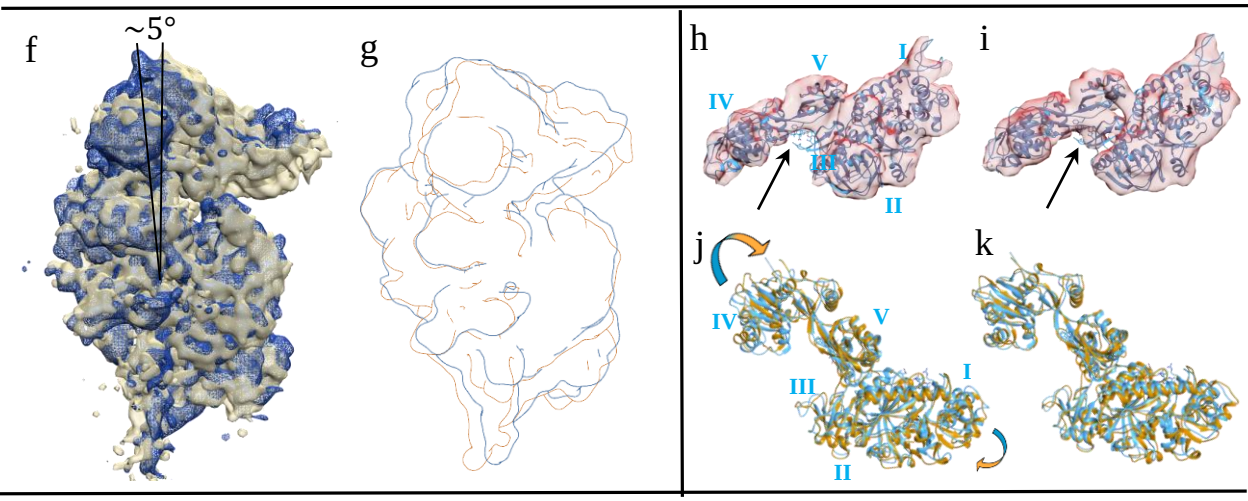

a
L1Int
b
L1Closed
c
L1Open
d
L1Open
e
L1Open
847 (26.2 %) 9.7 Å
721 (22.3 %) 9.8 Å
545 (16.9 %) 9.9 Å
877 (27.2 %) 9.6 Å
139 (4.3 %) 21.2 Å

f
L1In
t
g
L1Closed
h
L1Open

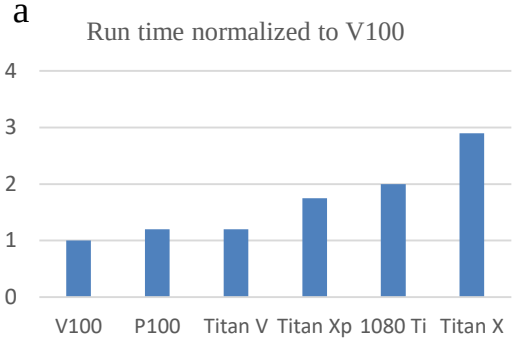
i
L1Ope
n
E
P
A



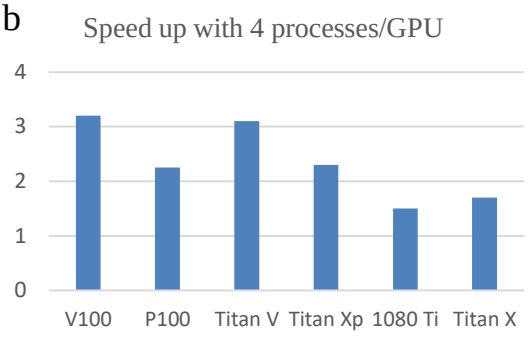
Class I - 922 (29.8 %) Class II - 474 (15.3 %) Class III - 409 (13.2 %) Class IV - 739 (24.0 %) Class V - 546 (17.7 %)

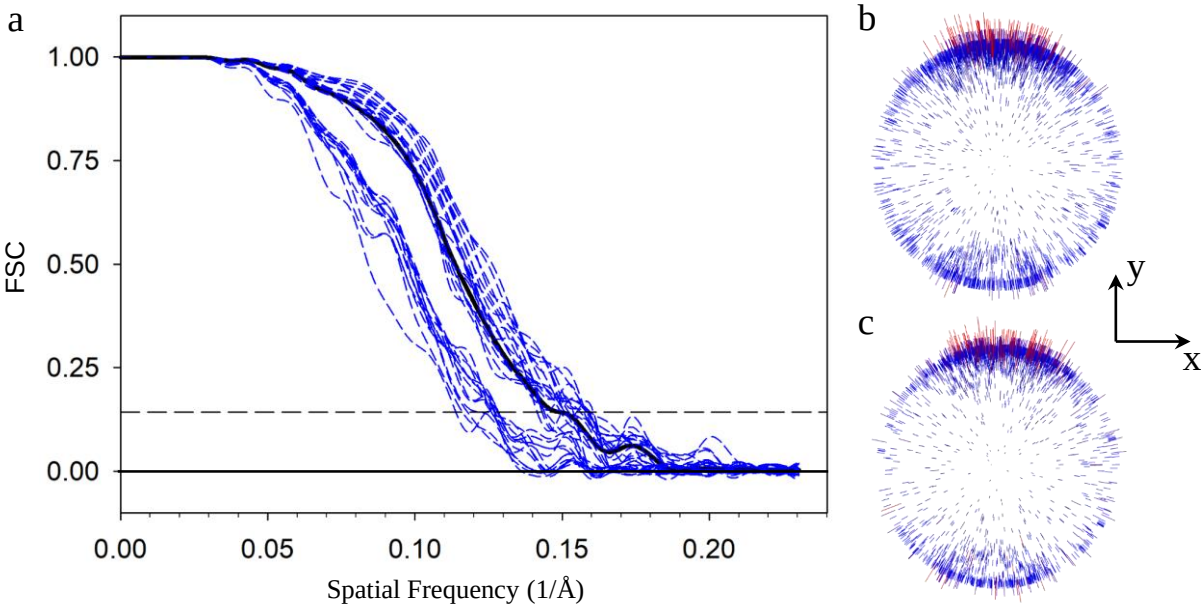


a


Run time normalized to V100
4
3
2
1
0
V100
P100
Titan V
Titan Xp
1080 Ti
Titan X

b


Speed up with 4 processes/GPU
4
3
2
1
0
V100
P100
Titan V
Titan Xp
1080 Ti
Titan X


a
b
c
1.00
0.75
0.50
0.25
0.00
FSC
0.00
0.05
0.10
0.15
0.20
Spatial Frequency (1/Å)
y
x

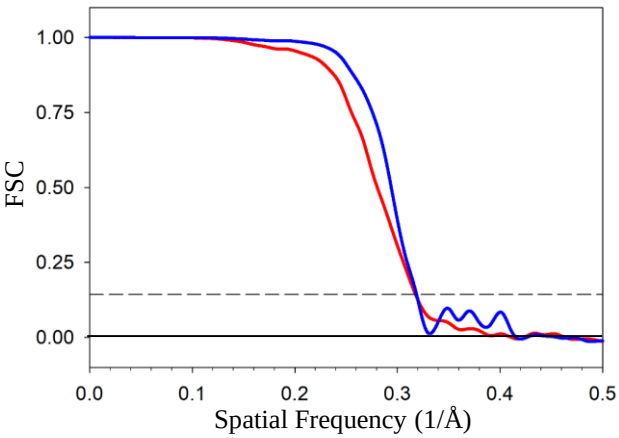

1.00
0.75
0.50
0.25
0.00
FSC
0.0
0.1
0.2
0.3
0.4
0.5
Spatial Frequency (1/Å)

Table S1. Comparison of features among major sub-tomogram averaging software packages

	emClarity‡	pyTOM[1]	Jsub-Tomo[2]	Dynam[3]	Eman2[4]	RELION[5]	PEET[6]	Protomo /i3[7]
GPU support	Yes	No	No	Yes	No	No	No	No
GUI	No	No		Yes		No	Yes	No
Template Matching	Yes	Yes	No	Manual	Manual	No	Manual	Yes
3D-CTF correction	3D-CTF WBP	No/Yes	No	No	Per-particle	Per-particle	No	No
Missing-wedge Representation	3D-Sampling Function	Binary Wedge	Binary Wedge	Binary Wedge	Fourier Intensity threshold	3D-Sampling Function	Binary Wedge	Binary Wedge (Sharp)
Tilt-series refinement using sub-volumes	Yes	No	No	No	No	No	No	No
"Gold-standard"	Yes FOM weighting	Yes FOM weighting	No	Yes	Yes Ad-Hoc weighting	Yes FOM weighting	Yes	No
Classification	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes

‡ This work

1. Hrabe, T. *et al.* PyTom : A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *J. Struct. Biol.* **178,** 177–188 (2012).
2. Maurer, U. E. *et al.* The structure of herpesvirus fusion glycoprotein B-bilayer complex reveals the protein-membrane and lateral protein-protein interaction. *Structure* **21,** 1396–1405 (2013).
3. 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.* **178,** 139–151 (2012).
4. Galaz-Montoya, J. G., Flanagan, J., Schmid, M. F. & Ludtke, S. J. Single particle tomography in EMAN2. *J. Struct. Biol.* **190,** 279–290 (2015).
5. Bharat, T. A. M. & Scheres, S. H. W. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. *Nat. Protoc.* **11,** 2054–2065 (2016).
6. Heumann, J. M., Hoenger, A. & Mastronarde, D. N. Clustering and variance maps for cryo-electron tomography using wedge-masked differences. *J. Struct. Biol.* **175,** 288–299 (2011).
7. Winkler, H. 3D reconstruction and processing of volumetric data in cryo-electron tomography. *J. Struct. Biol.* **157,** 126–137 (2007).

Table S2. Run times for the yeast 80s tutorial data[a]

Steps	Time	# of GPUs?
Average	1h 31m	2
Align	9h 45m	2
Tomo-CPR	2h 10m	2
3D-CTF	1h 16m	2
CTF update	0h 21m	2
Template Matching	0h 12m (1h 24m)	2
Classification	0h 8m	2
Total	15h 20m	

Footnote: a. Run on a Samsung 850-pro solid state scratch disk, requiring ~ 520 Gb total space. 2 x 20x Intel Xeon CPU E5-2650 v2 @ 2.3 GHz (only 12 cores used), 512 Gb of memory available, ~ 32 Gb used. 2x Nvidia Titan V GPUs.

Table S3. Alignment details for the ribosome data. The yeast 80s (EMPIAR-10045) and mammalian 80s (EMPIAR-10064) differ only in the size of the mask, and particle mass as noted in the bottom row.

Cycle	Binning	Pixel	Angular Search	Mask Type	Mask Radius (Å)	Particle Radius (Å)	Mass (MDa)
template matching	5	10.85	[180, 15, 180, 15]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
0	4	8.68	[0, 0, 16, 4]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
1	4	8.68	[15, 5, 0, 0]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
2	4	8.68	[0, 0, 18, 2]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
3	3	6.51	[12, 4, 0, 0]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
4	3	6.51	[0, 0, 9, 1.5]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
5	3	6.51	[7.5, 2.5, 0, 0]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
6	2	4.34	[7.5, 2.5, 0, 0]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
7	2	4.34	[0, 0, 6, 1]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
8	2	4.34	[4.5, 1.5, 0, 0]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
9	1	2.17	[4.5, 1.5, 0, 0]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
10	1	2.17	[0, 0, 5, 0.5]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
11	1	2.17	[3.75, 1.25, 0, 0]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
12	1	2.17	[1.5, 0.75, 0 , 0]	sphere	[180, 180, 180]	[160, 160, 160]	3.5
13	1	2.17	N/A	sphere	[180, 180, 180]	[160, 160, 160]	3.5
					mammalian + 20	mammalian + 20	mammalian + 1

Table S4. Alignment details for the HIV-1 Gag data (EMPIAR-10164)

Cycle	Binning	Pixel	Angular Search	Mask Type	Mask Radius (Å)	Particle Radius (Å)	Mass (MDa)
template matching	7	7	[180, 9, 28, 7]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
0	6	6	[0, 0, 24, 3]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
1	6	6	[16, 4, 0, 0]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
2	6	6	[0, 0, 9, 1.5]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
3	5	5	[0, 0, 12, 1.5]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
4	5	5	[15, 3, 0, 0]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
5	5	5	[0, 0, 9, 1]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
6	4	4	[0, 0, 9, 1]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
7	4	4	[10, 2, 0 , 0]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
8	4	4	[0, 0, 7.5, 0.75]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
9	3	3	[9, 1, 0, 0]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
10	3	3	[10, 2, 0, 0]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
11	3	3	[0, 0, 7.5, 0.75]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
12	2	2	[0, 0, 7.5, 0.75]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
13	2	2	[5, 1.25, 0, 0]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
14	2	2	[1.5, 0.75, 0.5, 0.5]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
15	2	2	[0, 0, 6, 0.5]	cylinder	[116, 116, 72]	[66, 66, 56]	0.5
16	1	1	[0, 0, 3, 0.3]	cylinder	[90, 90, 72, 72]	[66, 66, 56]	0.5
17	1	1	[4.5 0.75, 0, 0]	cylinder	[90, 90, 72, 72]	[66, 66, 56]	0.5
18	1	1	[1.2, 0.4, 0, 0]	cylinder	[90, 90, 72, 72]	[66, 66, 56]	0.5