nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-021-00648-5

Check for updates

High-resolution in situ structure determination by cryo-electron tomography and subtomogram averaging using emClarity

Tao Ni[1,4], Thomas Frosio[1,2,4], Luiza Mendonça[1,4], Yuewen Sheng[2], Daniel Clare[2], Benjamin A. Himes[3] and Peijun Zhang[1,2]✉

Cryo-electron tomography and subtomogram averaging (STA) has developed rapidly in recent years. It provides structures of macromolecular complexes in situ and in cellular context at or below subnanometer resolution and has led to unprecedented insights into the inner working of molecular machines in their native environment, as well as their functional relevant conformations and spatial distribution within biological cells or tissues. Given the tremendous potential of cryo-electron tomography STA in in situ structural cell biology, we previously developed emClarity, a graphics processing unit-accelerated image-processing software that offers STA and classification of macromolecular complexes at high resolution. However, the workflow remains challenging, especially for newcomers to the field. In this protocol, we describe a detailed workflow, processing and parameters associated with each step, from initial tomography tilt-series data to the final 3D density map, with several features unique to emClarity. We use four different samples, including human immunodeficiency virus type 1 Gag assemblies, ribosome and apoferritin, to illustrate the procedure and results of STA and classification. Following the processing steps described in this protocol, along with a comprehensive tutorial and guidelines for troubleshooting and parameter optimization, one can obtain density maps up to 2.8 Å resolution from six tilt series by cryo-electron tomography STA.

Introduction

Cryo-electron tomography (cryoET) has gained increasing importance in the study of molecular architectures of viruses, bacteria and cellular components in situ[1–3]. It can provide 3D reconstructions of pleomorphic objects such as organelles or cells in their close-to-native states, providing unique opportunities to capture the intermediate biological events in the cellular context. More importantly, the spatial relationship among macromolecules within a cellular tomogram can be determined[4]. In cryoET, a series of images from the same region of the specimen are recorded as the sample is tilted to various angles with respect to the incident electron beam. The images are subsequently aligned and reconstructed to generate a 3D tomogram. When there are many repeating objects, such as macromolecular complexes, in the tomogram, these objects can be aligned and averaged to improve the signal-to-noise ratio (SNR)[5], a process referred to as cryoET subtomogram averaging (STA).

Compared with cryoEM single-particle analysis (SPA), STA generally results in lower resolution. However, STA can resolve macromolecule structures in situ, unpurified and in the cellular context, as well as provide a spatial relationship between molecules, which is important for interpreting their biological functions. Nonetheless, several studies have yielded high-resolution density maps resolving secondary structural elements, including coat protein complex I (ref. [6]), nuclear pore complex[4,7], polysomes[8], chemotaxis signaling arrays[9], retroviruses assembly[10–14], bacteria surface layer[15] and ribosomes[16].

There are multiple additional challenges in STA compared with SPA[1,17,18]. First, due to the physical limits of the goniometer as well as increasing sample thickness upon tilting, tilt series are typically limited to tilt angles between −60° and 60°. The densities in a tomogram reconstructed from these tilt series therefore suffer distortions, referred to as missing-wedge effect. This distortion substantially affects the precision of subtomogram alignment and classification and must be

[1]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Diamond Light Source, Harwell Science and Innovation Campus, Didcot, UK. [3]Howard Hughes Medical Institute, RNA Therapeutics Institute, University of Massachusetts Chan Medical School, Worcester, MA, USA. [4]These authors contributed equally: Tao Ni, Thomas Frosio, Luiza Mendonça. ✉e-mail: peijun.zhang@strubi.ox.ac.uk

considered for high-resolution STA. Second, biological samples are sensitive to radiation damage, and the electron exposure applied to each tilted image is usually limited. As a result, the SNR of a tilted image is much worse compared with images in SPA. Third, specimens for cryoET are usually thick, and the effective thickness of sample increases when sample tilts. The defocus gradient due to the thickness of sample and sample tilt also needs to be considered[19]. As many biological objects adopt multiple conformations or compositions, 3D classification is required to delineate these different variances. While STA has, in principle, an advantage in 3D classification over SPA since each particle exists as a unique 3D reconstruction, thus allowing for direct analysis of the 3D variance, the low SNR and missing-wedge effect often pose significant challenges[20].

To deal with these challenges, a number of software packages have been developed for STA this far, including PEET (ref. [21]), EMAN2 (refs. [22–24]), RELION (refs. [25,26]), Dynamo (ref. [27]), Jsubtomo (ref. [28]), PyTom/AV3 (refs. [29,30]), Warp/M (ref. [16]), Protomo/i3 (ref. [31]) and emClarity (ref. [32]) (see review by Zhang[1] for a comparison). We implemented several key features in emClarity. First, an algorithm was implemented to estimate the defocus and astigmatism for each tilted image within the tilt series, to calculate the contrast transfer function (CTF). The effect of CTF modulation of images is then corrected for during tomogram reconstruction, accounting for the depth of field[32]. Second, for accurate weighting during alignment, reconstruction and classification, emClarity computes 3D sampling functions (3DSF). The 3DSF of each subtomogram, which accounts for the missing wedge information, is updated during each step of processing and used as a weight. Third, to address sample heterogeneity, emClarity implements a multiscale 3DSF-weighted, principal component analysis (PCA)-based classification method, which allows the user to emphasize specific features of different length scales. Fourth, local specimen motion and deformation place a major restriction on the quality of STA reconstructions. emClarity implemented tomogram constrained projection refinement (tomoCPR) to refine local shifts, rotations and magnification changes in the sample by using subtomograms as fiducial markers. This improves the tilt-series alignment, particularly for in situ cryoET datasets recorded from cryo-focused ion beam milled lamellae, where it would not have made sense to use gold bead fiducials because they would be removed during the milling process.

Several high-resolution cryoEM maps have been successfully obtained by various research groups using emClarity[1], including severe acute respiratory syndrome coronavirus 2 postfusion spikes[33], in situ structure of Parkinson's disease-linked leucine-rich repeat kinase 2 (ref. [34]), cellular reovirus assembly intermediates[35], Zika virus capsid protein[36], nodaviral replication protein A crown complex[37], native *Leptospira spirochete* flagellar filaments[38] and bacterial chemotaxis signaling arrays[39].

The new version of emClarity (V1.5.3.10) has some major differences from the original publication (V1.0) (ref. [32]). These include the following:

- Per-tilt CTF refinement using embedded CTFFIND4 (ref. [40])
- Handedness check during CTF estimation
- Calculation of per-particle 3DSF
- 3DSF calculation has been improved
- Switch to MATLAB 2019a
- Peak masks to limit translational search in alignment: the peak mask can be used to remove the cross-correlation peaks from a given distance of the particle origin, i.e., it defines the maximum translation allowed
- Reconstruction using the raw projection images using cisTEM

Here, we describe a detailed workflow and processing steps using the new version of emClarity. The protocol has been tested by several novice users, and the common issues that might arise during the procedure are detailed in Troubleshooting.

Overview of emClarity pipeline

emClarity streamlines all steps in the pipeline (Fig. 1). emClarity can align the raw tilt series automatically using its 'autoAlign' program. It can also import the aligned tilt series from external software packages, as long as the file formats and naming conventions follow the requirement (Step 1). It then generates aligned tilt series and estimates the CTF of each tilt series (Steps 2–4). Users define the boundary of subregion(s) in the tomogram for later reconstruction (Steps 5–7). The particles are then picked using template matching (Steps 8–12). emClarity manages the subtomogram-associated metadata in a MATLAB database and updates the metadata after each processing step throughout the pipeline (Step 13). The CTF-corrected tomograms are then generated at the requested binning (Step 14), and STA and alignment can be performed iteratively at each

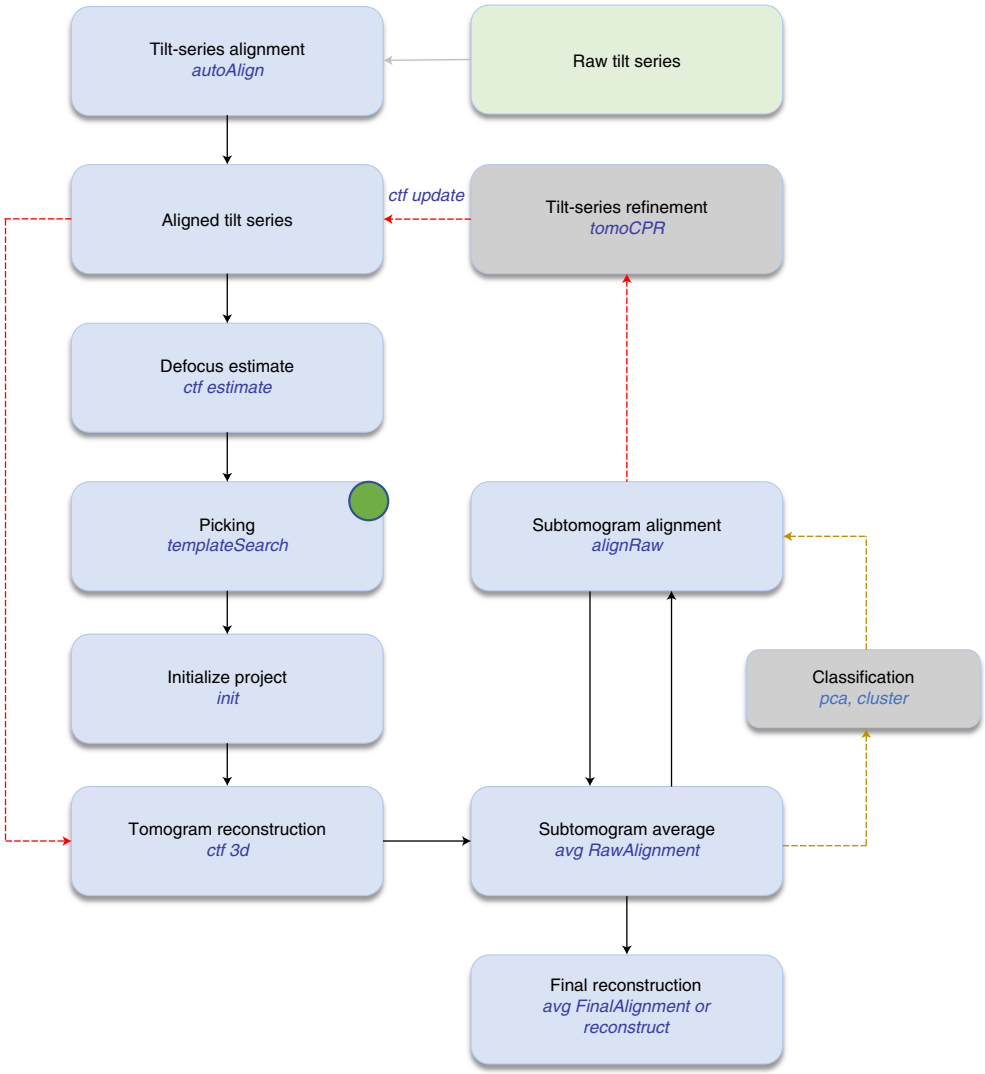

Fig. 1 | emClarity processing workflow. The green box indicates data input. emClarity processing steps are in blue boxes and optional steps are in gray boxes. Dashed red and gold lines are optional tomoCPR and classification processes, respectively. Subtomogram positions can also be imported from other software (indicated by the green circle). emClarity commands are shown in blue text.

binning (Steps 15–18). tomoCPR can be performed (Steps 19–20) to refine tilt-series alignment as well as subtomogram classification (Steps 22–30), both of which are optional steps. During the iterative alignment and averaging cycles, the data are kept in two fully separate half-sets following the 'gold-standard' refinement procedure[41]. The half-sets are used to calculate an optimal filter for weighting the reconstructions, while reducing the risk of overfitting[42]. A final map can be generated combining the two half-sets with an additional B-factor sharpening optionally applied (Step 31). A new feature is additionally implemented in emClarity, such that the raw projection images, instead of subtomograms, can also be used for the final reconstruction using cisTEM. Table 1 lists cryoET data collection and processing details. emClarity processing run time for the main steps is illustrated in Table 2, along with specific graphics processing unit (GPU) cards used for processing.

Prerequisite for using the protocol

This protocol is broadly applicable to cryoET STA projects, but is focused on providing details needed for high-resolution refinement. emClarity uses GPU accelerations and parallelization tools to cope with large datasets. Since emClarity does not have a graphic user interface, users are expected to have basic knowledge of working with the command line on Unix/Linux-based systems. It is beneficial to

Table 1 | CryoET data collection and processing details

Titan Krios[a]	GagT8I	Gag WT (EMPIAR-10164)	Ribosome (EMPIAR-10304)	Apoferritin
Voltage (kV)	300	300	300	300
Detector	Falcon 4	Gatan K2	Gatan K3 bioquantum, 20 eV	Gatan K3
Energy filter	Selectris X, 10 eV slit	Gatan		Gatan bioquantum, 20 eV
Gatan bioquantum, 20 eV				
Super-resolution mode	Yes	Yes	Yes	Yes
Pixel size (Å)	1.18	1.35	2.1	1.34
Total electron dose ($e^-/Å^2$)	122	~120	~120	102
Dose rate ($e^-/Å^2/s$)	3	3		4.2
Frame number	10	10		10
Acquisition scheme	−60°/60°, 3°	−60°/60°, 3°	−60°/60°, 3°	−60°/60°, 3°
Defocus range (μm)	−1.36 to −3.11	−1.5 to −3.96	−2.2 to −4.3	−1.5 to −3.5
Number of tilt series	5	5	12	6
Software	Etomo, emClarity	Etomo, emClarity	emClarity	Etomo, emClarity
Number of tomograms	5	5	12	6
Number of initial subtomograms	20,010	15,791	10,441	5,668
Number of subtomograms after classification	13,844	15,460	8,131	4,826
Symmetry imposed	C6	C6	C1	O
Resolution at 0.143 FSC	5.0 Å/4.5 Å[b]	3.3	7.0	2.8
Data deposited	EMPIAR-10643, EMD-13390	EMD-13354	EMD-13270	EMPIAR-10787, EMD-13271

[a]Titan Krios is a 300 kV electron microscope used for cryoEM data collection: https://www.thermofisher.com/uk/en/home/electron-microscopy/products/transmission-electron-microscopes/krios-g4-cryo-tem.html. [b]Density maps were calculated using subtomograms averaged in emClarity (5.0 Å) or using projection images (4.5 Å) reconstructed by cisTEM implemented in emClarity.

Table 2 | emClarity processing run time (five tilt series)

emClarity processing steps	Binning	GPU card	No. of GPU units	Time
CTF estimate	1	Tesla V100	1	25 min
Template search	8	Tesla V100	1	1 h
init cycle 0-2	6	Tesla V100	4	40 min
tomoCPR-1 cycle 3–5	5	Tesla V100	4	2.5 h
tomoCPR-2 cycle 6–8	4	Tesla V100	4	3 h
Classification	4	Tesla V100	4	40 min
Cycle 9–10	4	Tesla V100	4	1.5 h
tomoCPR-3 cycle 11–13	3	Tesla V100	4	2 h
tomoCPR-4 cycle 14–16	2	Tesla V100	4	3 h
tomoCPR-5 cycle 17–18	1	Tesla V100	4	10 h
avg FinalAlignment	1	Tesla V100	4	1 h
cisTEM reconstruct/refine	1	Tesla V100 (CPU)	12 CPU cores	1.5 h

Each tomogram is divided into multiple subregions (one VLP/subregion), which are processed in parallel.

have good knowledge of fiducial based alignment as implemented in Etomo[43]. Familiarity with MATLAB scripting can be helpful, but is not required. Basic knowledge of PCA and commonly used clustering method (such as *k*-means clustering) is useful when carrying out emClarity subtomogram classification. Users can also refer to the associated emClarity tutorial (Supplementary Information 1 and https://github.com/ffyr2w/emClarity-tutorial) for in-depth understanding algorithms behind each step, as well as detailed step-by-step processes using a ribosome dataset (EMPIAR-10304).

Limitations

Because emClarity uses a template-based particle picking method, it requires users to have a template for the object of interest. One should pay close attention to the template search and be cautious to template bias. We recommend using a low-pass filtered template to minimize template bias. emClarity implement template matching with either non-CTF-corrected or CTF-corrected tomograms, and comparison or combination of these two results can be informative for some challenging datasets. Small objects (<0.5 MD), such as severe acute respiratory syndrome coronavirus 2 spikes in cellular tomography dataset, can be identified through template search, albeit containing false positives. In this case, the existing prior information (such as particle position and orientation relative to membrane) can be used to exclude these false positives. The number of desired particles during template search can be either determined automatically within emClarity or set manually by user. When templates are not available, one can use other software packages, such as Dynamo[27] and PEET[21], to generate an initial template. It is also possible to import particles (coordinates and angles) picked or refined from other software into emClarity (Fig. 1, green dot). Although emClarity can refine tilt-series alignment by tomoCPR, we recommend aligning the initial tilt series to a satisfactory level using emClarity autoAlign or other packages like Etomo[43] or AreTomo (https://msg.ucsf.edu/software). In some cases, results of geometry refinement by tomoCPR might be inadequate.

Materials

Equipment and setup

A computer or a computing cluster with NVIDIA GPU cards with at least 12 GB memory, CUDA version 7.5 or greater (version 9 or newer preferred). An emClarity binary (version 1.5.3.10) and installation procedure are available and detailed in emClarity wiki (https://github.com/bHimes/emClarity/wiki).

Input data

Data: raw tilt series

Raw image movies need to be motion-corrected, but without exposure weighting, which is handled internally by emClarity. Motion-corrected images in a tilt series should be ordered in the sequence of tilt angle, from −60° to 60°, for example. Tilt series can be aligned using external software packages like Etomo and imported to emClarity. Users can also import the raw tilt series and use emClarity to align it automatically. Details of required files and formats are listed in Step 1 in the Procedure.

Data: metadata

- Microscope imaging conditions: voltage, pixel size, defocus range, amplitude contrast and Cs
- Data collection scheme (the order and exposure dose of image acquisition in a tilt series)

emClarity currently uses a parameter file to manage inputs, usually named to reflect their function and cycle, such as `param_ctf.m` for CTF estimation and param1.m for cycle 1 alignment, averaging and classification. The parameters required for individual step are listed and explained in detail in the tutorial (Supplementary Information 1). A parameter file together with run commands for the processing of human immunodeficiency virus type 1 (HIV-1) Gag dataset in this protocol is shown in Supplementary Information 2, and a template is supplied with emClarity installation.

Procedure

▲ **CRITICAL** This protocol presents a stepwise working procedure for STA and classification using emClarity. Users run all the commands through a terminal shell inside the project directory. The entire iterative alignment, averaging and classification procedure can run to the end automatically through a runscript, as long as the parameter files are set properly for each cycle. Users should modify and optimize the key parameters relevant to their projects. In the following processing steps, Steps 1–31, we provide the individual run commands with specific parameters and discuss the results, as well as troubleshoot potential issues. Novice users are recommended to follow the exact steps and check the outputs for each step and compare with the results described here. Users can refer to a more comprehensive tutorial (Supplementary Information 1) (https://github.com/ffyr2w/emClarity-tutorial), which contains a detailed explanation of all parameters and basic algorithm for each processing step in emClarity.

Preparation: arrangement of input files and directories ● Timing ~30 min when using *autoAlign*

▲ CRITICAL Tilt series can be aligned automatically inside of emClarity, or externally using software like Etomo. In this protocol, some datasets were aligned using Etomo and imported to emClarity, and some were automatically aligned using the 'emClarity *autoAlign*' program. The '*autoAlign*' function requires motion-corrected image stacks, tilt angle file and tilt axis rotation angle, and it prepares all the necessary files in `fixedStacks/`. Please refer to the tutorial in Supplementary Information 1 for the parameters. If users align the tilt series using external software like Etomo, please prepare the necessary files as indicated in Step 1.

1 Make a project directory. Within the project directory, make a new directory called `fixedStacks/`. It is essential to strictly follow the naming conventions. Copy the following files into it.
- `<prefix>.fixed`: the raw tilt series corresponding to `<prefix>.st`
- `<prefix>.xf`: the transformation file generated from tiltalign in Etomo
- `<prefix>.tlt`: the refined tilt-angle file
- (optional) `<prefix>.local`: the local alignment transformation file corresponding to `<prefix>local.xf` from tiltalign in Etomo
- (optional) `<prefix>.erase`: coordinates of the fiducial beads to erase, corresponding to <prefix>_erase.fid in Etomo
- (optional) `<prefix>.order`: refined tilt angles listed in the order of image acquisition. For example, if data collection starts from 0° and alternates between positive and negative values as follows: 3°, −3°, 6°, −6°, …, 60°, −60°, then the order file contains a single column listing these angles as 0, 3, −3, 6, −6 … 60, −60. However, we recommend generating the order file if the data acquisition scheme can not be represented by the exposure-weighting parameters (see Step 3)

If there are black images at high angle in the tilt series, we recommend removing these dark images during tilt-series alignment and making sure the corresponding `.xf` and `.tlt` are also updated. It is recommended to process the raw tilt series with IMOD CCD eraser to remove hot and dead pixels.

2 Set up appropriate working environment for emClarity (e.g., `module load emClarity/1.5.3.11`). Run emClarity using the provided command list (Supplementary Information 2). Users can run through the script entirely or run individual command separately as described below. If you have existing IMOD or UCSF Chimera in the environment, make sure there is no conflict. All the emClarity related logs are saved in `logFile/emClarity.logfile`.

Defocus estimate ● Timing ~25 min

3 Estimate the defocus of the tilt series. In this step, the raw tilt series will be transformed into aligned tilt series using the per-tilt transformation file; the gold fiducials will be removed; and the aligned tilt series will be used for per-tilt defocus and astigmatism estimation. The parameter file should contain the necessary imaging parameters. Copy a template parameter file to the project directory and rename it `param_ctf.m`.

System parameters:	
`nGPUs=4`	%% number of visible GPUs
`nCpuCores=12`	%% maximum number of processes to run in parallel
Microscope settings:	
`PIXEL_SIZE=1.179e-10`	%% pixel size of raw tilt series, in meters
`SuperResolution=0`	%% whether raw tilt-series pixel size corresponds to super-resolution image pixel size
`Cs=2.7e-3`	%% Spherical aberration of the microscope, in meters
`VOLTAGE=300e3`	%% accelerating voltage of the microscope, in volts
`AMPCONT=0.1`	%% amplitude contrast
`beadDiameter=7e-9`	%% fiducial bead diameter, in meters
Defocus range:	
`defEstimate=2.3e-6`	%% initial estimate of the defocus, in meters
`defWindow=1.5e-6`	%% defocus estimate window, in meters
Exposure-weighting parameters:	
`CUM_e_DOSE=123`	%% total exposure dose
`doseAtMinTilt=3`	%% electron dose at minimum tilt
`oneOverCosineDose=0`	%% whether Saxon scheme is used

Table continued

(continued)

```
startingAngle=0                 %% refined data collection starting angle, in degrees
startingDirection=pos           %% data collection direction
doseSymmetricIncrement=1        %% dose symmetric scheme group size
```

The last three parameters in exposure weighting are used to indicate the order of image acquisition for exposure weighting, which can also be specified by providing a `<prefix>.order` file in `fixedStacks/`. If a `<prefix>.order` is provided in the `fixedStacks/`, the exposure-weighting parameters will be ignored. For each tilt series, run the following command:

```
emClarity ctf estimate <param> <prefix>
emClarity ctf estimate param_ctf.m b2tilt20
```

A new directory `aliStacks/` will be generated in the project directory and the aligned tilt series `aliStacks/<prefix>_ali1.fixed` will be saved. For each tilt series, per-tilt defocus and astigmatism estimation results are saved as `fixedStacks/ctf/<prefix>_ali1_ctf.tlt`, which contains the tilt geometry information, accumulated exposure dose and per-tilt defocus information. Repeat CTF estimation for all tilt series:

```
#!/bin/bash
for stack in fixedStacks/*.fixed; do
prefix=${stack#fixedStacks/}
emClarity ctf estimate param_ctf.m ${prefix%.fixed}
done
```

4 Inspect the results of CTF estimation for each tilt series:
- Open the transformed tilt series in `aliStacks/<prefix>_ali1.fixed` in *3dmod* and make sure they are correctly aligned and fiducial beads are removed properly.
- emClarity also prints out the results of a tilt-series handedness check in the `logfile/emClarity.logfile`. The handedness check informs whether the expected defocus gradient matches the measured value. However, it should be noted that the handedness correctness does not necessarily indicate the biological handedness of density map is correct.
- Open `fixedStacks/ctf/<prefix>_ali1_psRadial_1.pdf` and check that the theoretical CTF estimate matches the radial average of the power spectrum of the tilt series.

? TROUBLESHOOTING

Define subregion boundaries ● Timing ~10 min

5 In many cases, the regions of interest are in some local areas (subregions) in the whole tomogram. The boundary of a subregion is defined in a binned tomogram with the entire field of view. Copy the `recScript2.sh` from emClarity installation directory to the project directory. Run the `recScript2.sh` script; a binned tomogram for each tilt series will be generated in the `bin10/` directory:

```
./recScript2.sh -1
```

6 Define the subregion boundaries in the bin10 tomogram by defining six points (x_{min}, x_{max}, y_{min}, y_{max}, z_{min} and z_{max}) to enclose the subregion. Inside the `bin10/` directory, run:

```
3dmod <prefix>_bin10.rec
```

If you have three subregions in one tomogram, you will need to define $6 \times 3 = 18$ points. Save the model (File → Save model) with the same name as the tomogram but with the .mod extension in the `bin10/` directory. One should generate one `*.mod` file per tilt series. Leave at least a few pixels from the edge of the binned reconstruction for model boundary and subregions in a tomogram should not overlap. Subregions can be as big as the whole tomogram as long as the GPU cards have enough global memory. In practice, splitting the tomogram into two subregions is supported for GPUs with

≥12 GB of memory. In this tutorial, we defined each virus-like particle as one subregion so that multiple subregions can be processed in parallel to maximize computational throughput.

7 Convert the <prefix>_bin10.mod file to an emClarity format. This generates a recon/ directory, within which <prefix>_recon.coords defines the boundary information of each subregion of every tomogram. In the project directory, run:

```
./recScript2.sh <prefix>
```

To convert all the subregions of each tomogram, run:

```
#!/bin/bash
for stack in bin10/*.mod; do
prefix=${stack#bin10/}; ./recScript2.sh ${prefix%_bin10.mod};
done
```

Pick particles ● Timing ~1.5 h

▲ **CRITICAL** emClarity uses a template-based particle picking method. A template is required (Step 8) and template search for each subregion is performed at designated binning (Steps 9 and 10). Check the template search result (Step 11).

8 Prepare the template for particle picking. The template used by emClarity needs to have the same pixel size as that of the raw tilt series (PIXEL_SIZE parameter). One may need to rescale the template from a source map to match the pixel size.

```
emClarity rescale <input> <output> <inputPixel> <outputPixel> cpu/GPU
emClarity rescale EMD-8403.mrc emd_8403rescale.mrc 3.62 1.179 cpu
```

9 Generate CTF-corrected tomograms for template search. This step generates the binned tilt series and CTF-corrected (i.e., CTF multiplied) tomograms for each subregions and saves them as cache/<prefix>_<sub-region>_binX.rec.

Parameters:

`Tmp_samplingRate=8`	%% binning factor for tomogram for template search

```
emClarity ctf 3d param_ts.m templateSearch
```

10 Run a template search for each subregion from each tomogram. One needs to decide the binning of tomogram for template search. Depending on the subtomogram size, we typically recommend running template search with tomograms at a final pixel size ~8–10 Å/pixel. Ali_mRadius is the alignment mask radii. Test different Ali_mRadius and particleRadius to optimize particle picking, especially for subtomograms arranged in a lattice-like assembly. For the HIV Gag assembly, we set Ali_mRadius with the size of seven Gag hexamers and particleRadius with size of one hexamer, so that the cross-correlation is calculated with a large molecular mass, while the individual hexamers positions can be picked. For the ribosome or apoferritin dataset, Ali_mRadius and particleRadius can be very close. Tmp_angleSearch defines the range and step of out-plane and in-plane angular search as $[\theta_{out}, \Delta_{out}, \theta_{in}, \Delta_{in}]$ in degrees. For example, [180,9,35,7] specifies a ±180° out of plane search, with 9° each step, and ±35° in plane search with a 7° step. For subtomogram with cyclic symmetry, the in-plane search range can be limited to ±180/<symmetry>. Copy a template parameter file, rename it param_ts.m and update the following parameters. The microscope parameters should remain constant as in ctf estimate.

Parameters:

`Tmp_samplingRate=8`	%% binning factor for tomogram for template search
`particleRadius=[66,66,56]`	%% X,Y,Z particle radius in Å. Cross-correlation peak radius to remove from consideration after a particle in the current peak is selected
`Ali_mRadius=[116,116,72]`	%% radius of alignment mask in Å

Table continued

(continued)

```
Tmp_angleSearch= [180,9,35,7]        %% in degrees
Tmp_threshold=1000                   %% estimate number of particles
symmetry=C6                          %% particle symmetry
```

In the project directory, run:

```
emClarity templateSearch <param> <prefix> <sub-region> <template>
<symmetry> <GPU_id>
emClarity templateSearch param_ts.m b2tilt20 1 emd_8403rescale.mrc C6 1
```

A new directory called `convmap_wedge_Type2_binX/` contains the cross-correlation (CC) convolution map `<prefix>_<region>_binX_convmap.mrc` and model `<prefix>_<region>_binX.mod`, corresponding to the coordinates of picked particles. The resulting `<prefix>_<region>_binX.csv` file contains the unbinned coordinate and orientation information on all picked particles. Please refer to emClarity wiki for the convention and format of this file. A representative tomogram (bin8) and convolution map is shown in Fig. 2.

11 Clean the false-positive points using *3dmod*. In the `convmap_wedge_Type2_binX/` directory, run:

```
3dmod <prefix>_<sub-region>_binX_convmap.mrc <prefix>_<sub-region>_
binX.mod
```

It is also useful to overlay the raw tomograms with convmap and model:

```
3dmod ../cache/<prefix>_<sub-region>_binX.rec <prefix>_<sub-region>_
binX.mod
```

Check the `<prefix>_<sub-region>_binX_convmap.mrc` about the summed CC peaks to see whether they correspond to the desired subtomogram positions. Remove the false positive points, which are common in regions with strong features such as ice contamination, carbon edges and gold bead residues. Save the remaining points using the same model file name. Before averaging and alignment, one should ensure that the picked particles were mostly correct. It might not be necessary to clean all the false positive points as 3D classification usually can remove them.

12 Rename the `convmap_wedge_Type2_binX/` to `convmap/`, as emClarity will look into the `convmap/` directory for subtomogram information in the next step.

Initialize the project ● Timing ~1 min

▲ **CRITICAL** As mentioned above, emClarity stores all the project information in a MATLAB database. The database records information on the tilt series and subtomograms including: subregion boundary (`recon/<prefix>.coords`), per-tilt CTF estimate (`fixedStacks/ctf/<prefix>_ali1.tlt`) and information on each subtomogram (`convmap/`). These metadata will be used and updated throughout the emClarity data processing pipeline. Backup metadata will be saved as `cycleXXX_<project>_backup.mat` before a new cycle starts. Users can open the database in MATLAB to check the database structure.

13 Generate an emClarity database `<project>.mat`. Copy `param_ctf.m` to `param0.m` and update the following parameters:

Parameters:

```
subTomoMeta=gag                      %% project name
Tmp_samplingRate=8                   %% binning of the tomograms for template matching binning
fscGoldSplitOnTomos=1                %% whether or not the particles from the same subregions should
                                     be kept in the same half-set or distributed randomly
```

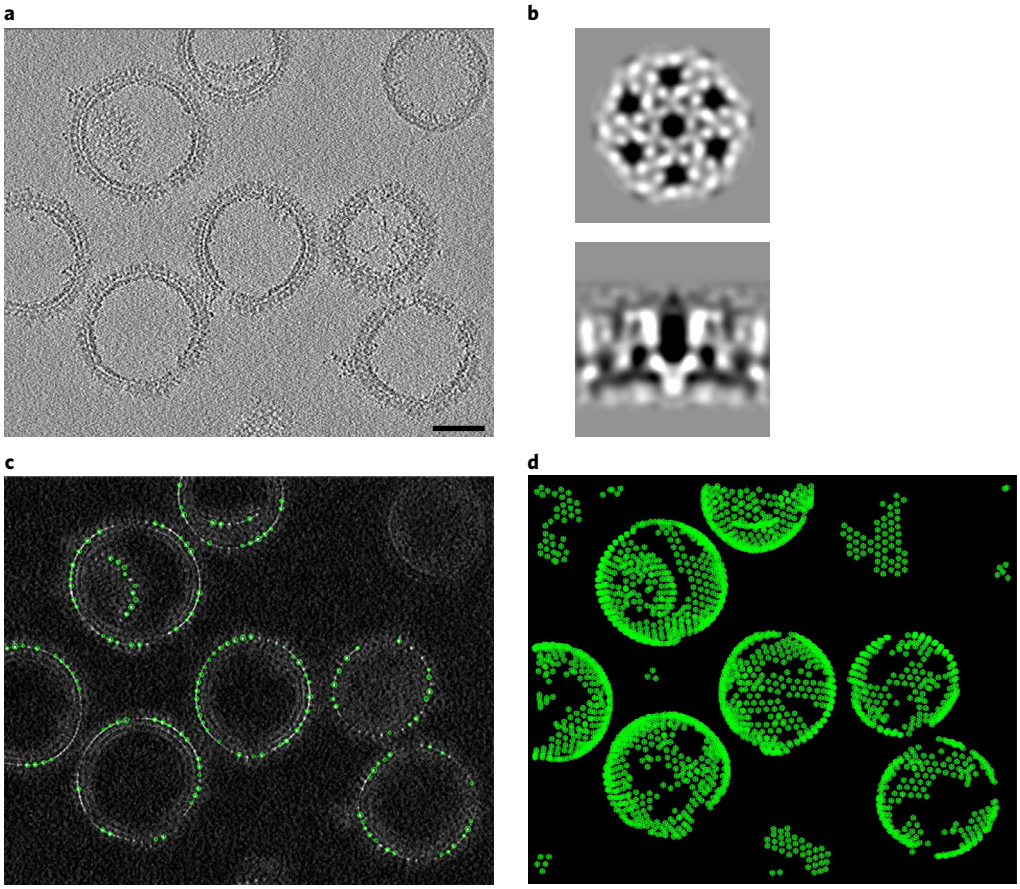

Fig. 2 | Template matching. a, A typical tomographic slice (6 nm thick) depicting HIV-1 Gag T8I assemblies from the raw data. **b**, The template used for particle picking, top and side views of HIV-1 Gag map (EMD-8403) low-pass filtered to 25 Å. **c**, A tomographic slice of resulting convolution map overlaid with template matched model points of top cross-correlation peaks. **d**, A projection view of model points through the tomogram volume. Scale bar, 50 nm.

Run the command as follows, which generates a metadata as `gag.mat`

```
emClarity init <param>
emClarity init param0.m
```

Note: `fscGoldSplitOnTomos` is typically set to 0 (randomly splitting subtomograms from each subregion into ODD and EVEN datasets). However, if the particles within the alignment mask overlap substantially with their neighbor particles, such as in the Gag lattice, we used '1' to split subregions instead of subtomograms for ODD and EVEN datasets to avoid floating the Fourier shell correlation (FSC). For a small dataset with a limited number of tilt series, we recommend defining more than two subregions for each tilt series.

Reconstruct the tomograms for alignment and averaging ● Timing ~5 min

14 Reconstruct the subregions for all the tilt series. This step generates the binned tilt series and CTF-corrected (actually CTF multiplied) subregions tomograms, which are saved in the `cache/` directory and are then used for the subtomograms extraction, averaging and alignment.

Parameters:

```
subTomoMeta=gag
PIXEL_SIZE=1.179e-10
Ali_samplingRate=6          %% binning of the tomograms for alignment
```

To generate a tomogram at a binning factor of 6, run:

```
emClarity ctf 3d <param>
emClarity ctf 3d param0.m
```

CTF-corrected tomograms `cache/<prefix>_<sub-region>_binX.rec` will be generated and one can check the tomogram with *3dmod* in IMOD.

STA and alignment ● Timing variable, depending on subtomogram number, size and binning

▲ **CRITICAL** STA and alignment are performed iteratively using tomograms at a progressively reduced bin (e.g., from bin6 to bin1). The binned tomograms can enhance the SNR and help subtomogram alignment, at the cost of losing high-resolution information. emClarity does not update alignment parameters automatically and allows users to set the tomogram binning factor (`Ali_samplingRate`), angular search range and step (`Raw_angleSearch`) for each cycle and judge whether the refinement has converged. Each cycle starts by generating an average for each half map (Step 15), which is then used as reference for alignment (Step 16). For each binning, it is generally recommended to run several cycles (Step 17). Similar to a template search, for samples with lattice-like structure, it is generally helpful to include several repetitive units (such as Gag hexamers) during the averaging and alignment.

15 emClarity does not extract the subtomograms onto disk by default; instead, the subtomograms will be extracted on the fly when needed, which can save large amounts of disk space for crowded samples.

Parameters:

```
subTomoMeta=gag
PIXEL_SIZE=1.179e-10                %% pixel size in meters
Ali_mRadius=[116,116,72]            %% in Å, enclosing seven hexamers
Ali_mCenter=[0,0,0]                 %% in Å
particleMass= 1                     %% in Megadalton
Ali_mType=sphere                    %% alignment mask type: sphere, cylinder, rectangle
particleRadius=[66,66,56]           %% corresponding to central hexamer size
Raw_className=0                     %% class 0
FSC_bfactor=10                      %% b-factor applied to half maps
Ali_samplingRate=6                  %% binning factor
symmetry=C6                         %% symmetry
```

Run the following command:

```
emClarity avg <param.m> <cycle_nb> RawAlignment
emClarity avg param0.m 0 RawAlignment
```

This generates two half maps in the project directory: `cycleXXX_<project>_class0_REF_EVE/ODD.mrc`. The dimensions of maps are calculated based on `Ali_mRadius` with additional padding. Open these two maps in UCSF Chimera or *3dmod*, or any software of your choice able to read MRC files, to check whether the maps match expectation. The corresponding (conical) FSC is available in `FSC/cycleXXX_<project>_Raw-1-fsc_GLD.pdf`, in which the dashed lines are conical FSC and the solid line is the overall FSC. The total sampling functions for both half maps `cycleXXX_<project>_class0_REF_EVE/ODD_Wgt.mrc` should be isotropic, if particles do not have preferred orientations in tomograms. Note that a molecular mask (`FSC/cycleXXX_<project>_Raw-1-shapeMask_*mrc`) is applied during FSC calculation. The overall sampling function and conical FSCs will indicate whether the subtomograms adopts preferred orientation. One can open the sampling function in *3dmod* and look through the *x*–*z* plane to see whether the amplitude weight is isotropic.

16 After the reference is generated with `avg`, emClarity can use this reference to align the particles. Similar to `Tmp_angleSearch` in template search, `Raw_angleSearch` in alignment step is also

defined as [θ_{out}, Δ_{out}, θ_{in}, Δ_{in}]. Since most of the particles are picked correctly for the Gag dataset (Step 9), the angular search ranges and step sizes for alignment are quite small.

Parameters (other parameters are identical as avg)
`Raw_angleSearch=[0,0,20,5];` %% angular search, in degrees.

```
emClarity alignRaw <param> <cycle_nb>
emClarity alignRaw param0.m 0
```

The changes of rotation and translation for every subtomogram in each subregion are saved in `alignResume/cycleXXX_<project>/<prefix>_<sub-region>.txt`. The number of lines in each file corresponds to the number of particles aligned in the current cycle. After all the subtomograms are processed, the metadata `<project>.mat` will be updated.

17 Copy `param0.m` to `param1.m` and `param2.m`, update `Raw_angleSearch` in these parameter files and repeat STA and alignment for a few cycles (Steps 14 and 15). For the speed of alignment, we usually alternate the in-plane and out-plane angular searches and perform a few cycles at each binning until the changes of rotation and shifts drop to around zero. In the same binning, one can repeat the same angular searches or gradually confine to finer angular searches. For the Gag dataset, two more cycles (cycle 1, 2) were run at bin6. Refer to Supplementary Information 2 for the list of commands and parameters at each cycle.

Parameters:

```
Raw_angleSearch=[16,4,0,0];          %% in param1.m
Raw_angleSearch=[0,0,9,3];           %% in param2.m
```

```
emClarity avg param1.m 1 RawAlignment
emClarity alignRaw param1.m 1
emClarity avg param2.m 2 RawAlignment
emClarity alignRaw param2.m 2
```

18 Remove duplicated particles after alignment.

```
emClarity removeDuplicates param2.m 2
```

After these averaging and alignment cycles, one can run a tilt-series refinement by tomoCPR (Steps 19 and 20, optional) and/or generate new tomograms and continue averaging and alignment (Step 21).

(Optional) Tilt-series refinement by tomoCPR ● Timing variable, depending on subtomogram number, size and binning

▲ **CRITICAL** Tilt series can be optionally refined by tomoCPR. STA provides accurate estimates of both particle positions and high SNR reconstructions, making them excellent fiducial markers. It is thus possible to leverage this information for improving the alignment of a tilt series. In this protocol, we run tomoCPR for each binning.

19 When using tomoCPR to refine the tilt-series geometry, the subtomograms are mapped back into raw tomograms to generate a synthetic tomogram containing an estimate of the background noise, plus the higher SNR particle, and projected into each view. A tile is cut out around each projected particle, convoluted with local CTF, and aligned to the corresponding particle in the raw data, to give rise to the particle position in the tilt series. These new positions of particles after local refinement will be used as new fiducial markers in tiltalign to refine the tilt-series alignment. Run the following command:

```
emClarity tomoCPR <param> <cycle_nb>
emClarity tomoCPR param2.m 2
```

A temporary directory `mapBack<n>/` is generated in `cache/` and will be moved to project directory only after all the tilt series are successfully processed. `<n>` indicates the current tomoCPR

number. The overall and local transformation files will be written as `mapBack<n>/<prefix>_ali<n>_ctf.tltxf` and `mapBack<n>/<prefix>_ali<n>_ctf.local` for each tilt series. The `mapBack<n>/` directory should not be deleted since the local transformation file `mapBack<n>/<prefix>_ali<n>_ctf.local` will be used to generate new tomograms, although any of the image files can be deleted to save disk space. The metadata `<project>.mat` will be updated to record the current round of tomoCPR.

20 Update the aligned tilt series and geometry file. Copy `param2.m` to `param3.m`.

Parameters:

```
Ali_samplingRate=5;                    %% tomogram binning
```

```
emClarity ctf update <param>
emClarity ctf update param3.m
```

A new geometry file `fixedStacks/ctf/<prefix>_ali<n+1>_ctf.tlt` and newly aligned tilt series `aliStacks/<prefix>_ali<n+1>.fixed` will be created, which will be used to generate new tomograms. One can check whether the newly transformed tilt series look well aligned and do not deviate substantially from original aligned stacks.

21 Generate the new tomogram at next binning (bin5). Run the following command:

```
emClarity ctf 3d <param>
emClarity ctf 3d param3.m
```

This is essentially repeating Step 14 at a new binning, followed by the STA and alignment cycle (Step 15 and 16), subtomogram duplicates removal (Step 18) and tomoCPR (Steps 19 and 20). The cycle then continues as the binning reduces.

For the Gag dataset, we run three cycles of averaging and alignment using 6×, 5× and 4× binned subtomograms before 3D classification. Update the `Ali_samplingRate` and `Raw_angleSearch` in the parameter files at each cycle. Refer to the command list in Supplementary Information 2.

(Optional) Subtomogram classification ● Timing ~40 min, depending on subtomogram number, size and binning

▲ **CRITICAL** Subtomogram classification (Steps 22–29) is optional in emClarity pipeline. In this protocol, we perform one cycle of 3D classification with bin4 subtomograms after two rounds of tomoCPR and six cycles of STA and alignment (Steps 14–21). emClarity uses a PCA-based classification method, with subtomograms band-pass filtered at various resolutions defined by users. It first computes an average map from all the subtomograms (Step 22). emClarity will then analyze the heterogeneity of the dataset by comparing individual subtomograms with the current average map (the reference). Briefly, difference maps are calculated between each particle and the references, for each resolution band that the user defines. These maps are then analyzed by PCA, using singular value decomposition. This results in a decomposition revealing the major directions of variance (eigenimages) (Step 23). Users will then select eigenimages corresponding to major direction of variance (Step 24), and emClarity will project the whole dataset along each of these eigenvectors. The projected data, which are now denoised and much smaller in size, are then clustered (by default with *k*-means clustering algorithm, Step 25). Then, the class averages will be generated for each cluster as a montage (Step 26), and particles from the undesired classes can be optionally removed from further analysis (these could be subtomograms that are 'noise' or conformations that are not of interest to the user) (Steps 27 and 28).

In principle, one can do classification at any binning and at any cycle. In practice, it is beneficial to have several rounds of alignment before classification and use an intermediate binning factor for a better SNR in tomograms (such as bin4, bin3). It is generally not recommended to conduct classification at bin1 if it was already done at higher binning.

22 Generate an average map for classification. Copy `param7.m` to `param8.m` and update `flgClassify=1` to turn on classification flag in the parameter file. Besides the parameters inherited from previous alignment cycles, other parameters specific to classification include:

Parameters:

```
Ali_mRadius=[116,116,72]                %% in Å, enclosing seven hexamers
Ali_mCenter=[0,0,0]                     %% in Å
Ali_mType=sphere
Ali_samplingRate=4                      %% binning factor for averaging
Raw_classes_odd=[0;1.*ones(2,1)]        %% C1 symmetry for half map 1
Raw_classes_eve=[0;1.*ones(2,1)]        %% C1 symmetry for half map 2
Cls_mRadius=[92,92,76]                  %% classification mask radius
Cls_mCenter=[0,0,0]
Cls_mType=sphere                        %% classification mask type
Cls_samplingRate=4                      %% binning factor for classification
flgClassify=1                           %% classification flag
```

```
emClarity avg param8.m 8 RawAlignment
```

This will generate two half maps: `cycleXXX_<project>_class0_Raw_EVE.mrc` and `cycleXXX_<project>_class0_Raw_ODD.mrc`.

23 Compute the difference map for each particle, with different band-pass filters. We set three band-pass filters at 10, 20 and 40 Å. The band-pass filters are selected according to the object one wishes to classify and typically below the maximum resolution of the current iteration. Most of variance is explained within the first 20–30 eigenimages, and `Pca_maxEigs` is used to limit the number of eigenimages to save.

Parameters:

```
pcaScaleSpace=[10,20,40]    %% one can select as many band-pass filters as possible, though three is typically sufficient
Pca_maxEigs=25              %% maximum number of eigenimages to save
```

Run the following command:

```
emClarity pca <param> <cycle_nb> <subset>
emClarity pca param8.m 8 0
```

It generates variance maps for each resolution band as `cycleXXX_<project>_varianceMap25-STD-*.mrc` and principal eigenimages as `cycleXXX_<project>_eigenImage25-STD-*.mrc`. To aid analysis, it is usually easier to look at `cycleXXX_<project>_eigenImage25-SUM-STD-mont_*.mrc`, which add a common reference to the eigenimages.

24 Select the main eigenimages by looking into each `cycleXXX_<project>_ eigenImage25-SUM-STD-mont_*.mrc` in *3dmod* and save the eigenimages numbering into `Pca_coeffs`. The eigenimages are numbered from 1 to `<Pca_maxEigs>`, counting from bottom left to top right by rows. For Gag dataset, eigenimages with hexagonal lattice feature can be selected and eigenimages that display missing-wedge effect are usually abandoned. Each resolution band requires the same number of eigenimages to be selected, which can be filled with zeros if there are not enough eigenimages in some resolution bands. Fill `Pca_coeffs=[zeros(1,12);7:18;7:18]` in `param8.m`.

25 Cluster the PCA results according to the selected eigenimages; this step groups the subtomograms into different number of classes (`Pca_clusters`). Multiple classes can be generated.

Parameters:

```
Pca_clusters=[9 12 16]          %% different number of clusters
```

```
emClarity cluster <param> <cycle_nb>
emClarity cluster param8.m 8
```

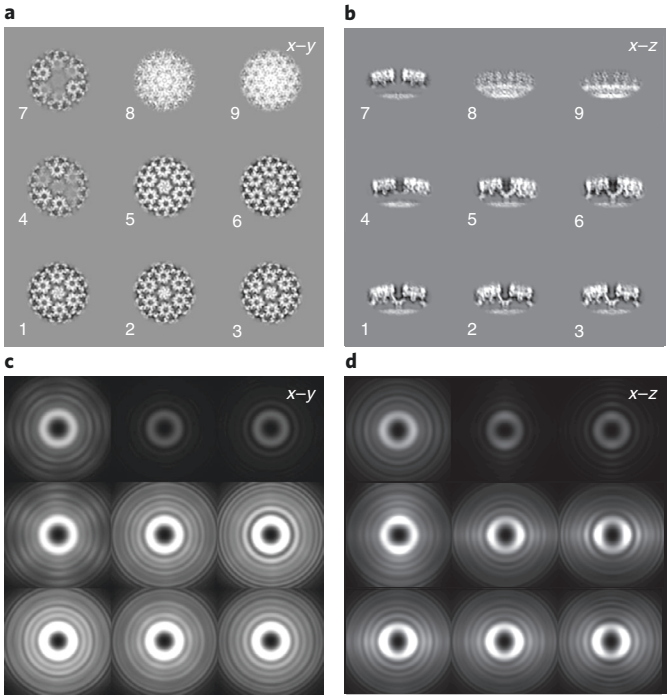

Fig. 3 | 3D classification and sampling function. a,**b**, A montage of nine 3D classes in *x*–*y* (**a**) and *x*–*z* slices (**b**). **c**,**d**, 3DSFs of the corresponding classes in *x*–*y* (**c**) and *x*–*z* slices (**d**). Note 3DSF confirms that the classification is not biased by particle orientations, as different classes have similar sampling functions and are nearly isotropic in different orientations.

This will use the `Pca_coeffs` and perform *k*-means clustering with 9, 12 and 16 target classes. The metadata will be updated and a text file `<project>_cycleXXX_ClassIDX.txt` listing the number of particles in each class will be generated.

26 Generate the class averages as a 3D montage. For the Gag dataset, we generated nine classes; the class average is numbered from 1 to `<Cls_className>`, counting from bottom left to top right by rows (Fig. 3). Set `Cls_classes_odd=[1:9;1.*ones(1,9)]`, the first row specifying the class ID and the second row specifying the cyclic symmetry.

Parameters:

```
Cls_className=9                          %% name of classes
Cls_classes_odd=[1:9;1.*ones(1,9)]       %% C1 symmetry for half map 1
Cls_classes_eve=[1:9;1.*ones(1,9)]       %% C1 symmetry for half map 2
symmetry=C1
```

```
emClarity avg <param> <cycle_nb> Cluster_cls
emClarity avg param8.m 8 Cluster_cls
```

? TROUBLESHOOTING

27 Inspect the class averages in 3dmod or UCSF Chimera.

```
3dmod cycle008_gag_class9_Cls_EVE.mrc
```

We classified the particles into nine classes (Fig. 3a,b). Seven of nine classes show clear hexagonal Gag lattice (classes 1–7) and were merged for further processing. It is generally informative to look at the sampling functions `cycle008_gag_class9_Cls_EVE/ODD.Wgt` to check whether the resulting classes have isotropic sampling function and proper coverage of defocus range (Fig. 3). Depending on the selection of eigenimages, the missing-wedge effect may dominate the classification, resulting in stretched structures. Create a new model point for each class to remove and save the model file such as `cycle008_remove.mod`.

28 Remove particles from the selected classes. `STD` refers to both the even and odd dataset.

```
emClarity geometry <param> <cycle_nb> RemoveClasses <remove.mod> STD
emClarity geometry param8.m 8 Cluster_cls RemoveClasses cycle008_
remove.mod STD
```

Subtomograms in these selected classes will be ignored for further analysis. The `cycle008_ClassMods_STD.txt` records the classes and number of subtomograms that have been removed. This should correspond exactly to the class populations from the clustering (Step 27) listed in file `<project>_cycleXXX_ClassIDX.txt`. If it does not, stop and make sure you followed the instructions from Step 24.

29 Skip the alignment for the current cycle, which prepares the metadata for the next cycle.

```
emClarity skip <param> <cycle_nb>
emClarity skip param8.m 8
```

30 Continue alignment and averaging cycles and tompCPR (optional) as in Steps 15–21. Turn off the classification flag in these parameter files by setting `flgClassify=0` and update the `Ali_samplingRate` and `Raw_angleSearch` for each cycle. For the Gag project, we ran several cycles of alignment with each binned tomogram and ran tomoCPR in the end of alignment at each binning factor (bin3, bin2 and bin1). Refer to the command list (Supplementary Information 2) for a summary of all the cycles for the Gag project.

Final reconstruction ● Timing ~2.5 h

31 For the final reconstruction, the two half datasets are combined. The updated versions of emClarity now offer two possibilities using either 3D subtomograms or their corresponding original 2D projections. To reconstruct through subtomograms, two half maps are reconstructed using `avg` as Step 15 and the conical FSCs are calculated, as well as the transformation between the two maps. The subtomograms from the second group are re-extracted and aligned to the first group using the aforementioned transformation. A final combined map is then generated averaging all aligned subtomograms from both halfsets and filtered using the FSC calculated, which is further sharpened with various b-factors.

Parameters:
`Fsc_bfactor=[10,25,75,100,250]`

```
emClarity avg param19.m 19 RawAlignment
emClarity avg param19.m 19 FinalAlignment
```

This generates the final reconstruction map `cycleXXX_<project>_class0_final_<b-factor>.mrc`. If one wants to use external software (e.g., RELION[44], cisTEM[45], Bsoft[46]) to apply different b-factors, masks or FSC weighting, one can take the raw half maps in the final cycle without FSC weighting `FSC/cycleXXX_<project>_Raw-*Ali.mrc`.

Alternatively, the final reconstruction can also be calculated from the 2D particles using cisTEM, as implemented in the updated version of emClarity. In this case, emClarity reprojects the 3D coordinates of the particles. A cisTEM STAR file is created, containing parameters such as, for each particle and for each view of the tilt series, its *x* and *y* position, rotation, defocus, and pre- and post-exposure. cisTEM will then calculate an initial reconstruction using its reconstruct3d program, then refine it using *refine3d* (note that the angles are not refined) and then finally calculates the final reconstruction with reconstruct3d using this refinement. For this protocol, we set maximum exposure to 60 electrons to include only the images within this exposure and generated the final map as `gag60e_refFilt_refined.mrc`. The `particleRadius` is set to be equivalent to `Ali_mRadius` to reconstruct the final density map with the same area as alignment.

```
emClarity reconstruct <param> <cycle_nb> <prefix> <symmetry> <max_exposure>
emClarity reconstruct param18recon.m 18 gag60e C6 60
```

Troubleshooting

Troubleshooting advice can be found in Table 3.

Table 3 | Troubleshooting table

Steps	Problem	Possible reason	Solution
3–4	Gold fiducial beads are not removed correctly	The gold fiducial file fixedStacks/<prefix>.erase is not present or incorrect	Check Etomo alignment and re-create the <prefix>.erase and redo CTF estimate
	The handedness is wrong	Tilt-series rotation angle is not correct (incorrect by 180°) during Etomo alignment	Redo Etomo alignment with correct tilt-series rotation angle (plus or minus 180°)
		Some detectors or software may save the raw image frames with additional rotation or flipping	Flip or rotate the tilt series to match the anticipated angle
	The estimated defocus is wrong	The anticipated defocus range defined in parameter `defEstimate` ± `defWindow` does not cover the real range	Check whether the theoretical CTF estimate matches the radial average of the power spectrum of the tilt series in `fixedStacks/ctf/<prefix>_ali1_psRadial_1.pdf`. Adjust the defocus range parameter `defEstimate` ± `defWindow` and rerun CTF estimate for the current tilt series. The correct defocus peak can be found in `fixedStacks/ctf/*_ccFIT.pdf`
9–11	Convmap does not show clear local CC peaks	Template should be at full pixel sampling, as it is binned internally; the template pixel size is not calibrated The tomogram is too noisy. For example, the tomogram is too thick	The template matching is very sensitive to the correct pixel size. When starting with an external reference, it is probably best to process 10% or so of the data to generate a new reference for the full run Use external software to improve the template Optimize the `Ali_mRadius` and `Ali_mType`, especially for particles in lattice assembly The subregion is filtered at the spatial frequency of the first CTF at zero of each tilt series by default. Overwrite it by including a different low-pass filter (parameter `lowResCut=40`) in the parameter file
9–12	Multiple points are picked on the same particle, or points are too close to one another	Ali_mRadius is too small	Increase `particleRadius` during template search, since it defines a region around a cross-correlation peak to remove from consideration after a particle is selected
13	Init fails to generate database file	convmap directory is not available	Remember to rename `convmap_wedgeType2_binX/` to convmap/ before running init
		Each <prefix>_<sub-region>_binX.mod should contain at least one point, i.e., one subtomogram	Make sure the number of `convmap/<prefix>_<sub-region>_binX.mod` matches the number of `recon/<prefix>_recon.coords` files in the recon/ folder
15	The average contrast is inverted	The non-CTF-corrected tomograms are used for template search. The first cycle of average was performed at the same binning as template search, which uses tomograms generated from template search instead of ctf 3d	Remove the previous `cache/<prefix>_<sub-region>_binX.rec` files generated during templateSearch before running ctf 3d if one wants to do the initial alignment at the same binning factor of template search. We generally start averaging and alignment at a different binning from the one used for template search
15–16	Failure in average or alignment step 'PEET error'	The subregion tomograms (*rec) or sampling function (*wgt) in cache/ directory are corrupted or not generated, which can happen when system disk is full during an emClarity step	Check the integrity of the `cache/<prefix>_<sub-region>_binX.rec` and cache/<prefix>_binX.wgt using *header* or open with *3dmod* command from IMOD. Remove the corrupted files and rerun ctf 3d
	'Reference to non-existent field cyclexxx"	The previous step (averaging, alignment, classification, etc.) did not finish successfully	Rerun the previous step (averaging, alignment, classification, etc.) to update the `<project>.mat`. Check the `logFile/emClarity.logfile` to make sure it finishes properly
	Out of memory	emClarity exits when the required GPU memory is not available	Reduce the number of parallel processing (set `nCpuCores=2`, for example). The requirement of GPU memory is related to the box size in average and alignment. In later stages of refinement with small binning factor, the required GPU memory for

Table continued

Table 3 (continued)

Steps	Problem	Possible reason	Solution
			each process is substantially higher. Check usage of GPU memory (type `nvidia-smi` in the command line)
19	tomoCPR fails to run	The average and alignment have not finished successfully	Rerun the average and alignment for the current cycle
		There is no particle in a subregion	Make sure there is at least one particle in a subregion. One can check the metadata `<project>`.mat or check the `alignResume/cycleXXX_/*.txt`. Each text file should contain at least one line. Particle in a subregion can be removed automatically if it drifts to the edge of subregion
	'Error using BH_synthetic_mapBack (line 978) mapBackRePrjSize = 4 is still too much for the Error in emClarity (line 366)'	The amount of GPU memory needed for reconstruction depends on the size of the local shifts. Generally, this error occurs only if the local shifts are unrealistically large	Remove this tilt series from your analysis
	tomoCPR results in misaligned tilt series	There are too few particles in the subregion All the particles are in one corner of field of view	To test whether tomoCPR helps, run average and alignment at one binning for several cycles until rotation and shifts are close to zero, then run tomoCPR and generate new tomogram in the same binning, and redo averaging and alignment to see whether density map or FSC improves. tomoCPR may not improve tilt-series alignment equally for all tilt series
22–30	Classification does not result in different classes	Suboptimal selection of eigenimages	Try a few different sets of Pca_coeffs and Pca_clusters and rerun cluster and average
25	Cluster does not run	The `Pca_coeffs file` is not formatted correctly	Make sure that `Pca_coeffs` contains the same number for each `pcaScaleSpace`. The rows of `Pca_coeffs` should be equivalent to number of `pcaScaleSpace`
31	Out-of-memory in cisTEM reconstruct	CPU memory is not sufficient	Reduce the `<max_exposure>` to include fewer images

Timing

The run time for each emClarity processing is listed in Table 2. Please note that the data processing times are for the Gag T8I dataset. The data processing time varies depending on the size of dataset, particle size, number of cycles, GPU models and other factors.

Steps 1–2, arrangement of input files and directories: ~30 min when using `autoAlign`

Steps 3–4, defocus estimate: ~25 min

Steps 5–7, define subregion boundaries: ~10 min

Steps 8–12, pick particles: ~1.5 h

Step 13, initialize the project: ~1 min

Step 14, reconstruct the tomograms for alignment and averaging: ~5 min, depending on the tomogram binning

Steps 15–18, STA and alignment: variable, depending on dataset size, particle size and binning

Steps 19–21, tilt-series refinement by tomoCPR: variable, depending on dataset size, particle size and binning and other factors

Step 22–30, subtomogram classification: ~40 min, depending on dataset size, particle size, binning and other factors

Step 31, final reconstruction: ~2.5 h, depending on dataset size, particle size and binning

Anticipated results

We illustrate the protocol using four datasets: a wild-type Gag dataset (a subset of 5 tilt series) and a ribosome dataset (a subset of 12 tilt series) from EMPIAR (EMPIAR-10164 and EMPIAR-10304),

a GagT8I assembly dataset (5 tilt series) from a previous study[47] and a new apoferritin dataset (6 tilt series) collected in-house (Table 1).

HIV-1 Gag T8I spherical assemblies

A challenging non-single-particle dataset of HIV-1 Gag T8I immature spherical assemblies with overlapping densities, but no icosahedral symmetry, is illustrated in detail in this protocol. These assemblies were produced in *Escherichia coli* as part of a study aiming to resolve the extended six-helix bundle of HIV-1 Gag hexamer.

The per-tilt CTF estimation of the tilt series is consistent with expected values from experimental setting. After the template search, the convolution map reveals local peaks corresponding to each Gag hexamer. Most of the hexamers in the lattice are picked for further analysis; a small number of particles were found to be false positives (Fig. 2). Subtomograms from each subregion were assigned to the same half datasets to avoid mixing halfsets that had overlapping peripheral density (`fscGoldSplitOnTomos=1`). STA and alignment was conducted using subtomograms binned at different factors (from 6× binned tomograms to 1× binned tomograms). After alignment was completed with each binned tomogram (except bin1), a tomoCPR tilt-series refinement was performed.

Since tomoCPR is an optional step and requires tuning of some parameters, we recommend users work on a new STA project to run through iterative STA and alignment without tomoCPR for the first instance.

A 3D classification was performed using bin4, which gave nine classes of images (Fig. 3). The classes display different features as shown in *x*–*y* and *x*–*z* slices (Fig. 3a,b), along with their corresponding overall 3DSFs in *x*–*y* and *x*–*z* slices (Fig. 3c,d). Classes 8 and 9 showed no clear Gag lattice (Fig. 3a,b); therefore, objects in these classes were removed from further processing. The sampling functions of the remaining classes reveal no preferential orientation, indicating that the 3D classification is not biased by the particle orientations in the raw tomogram.

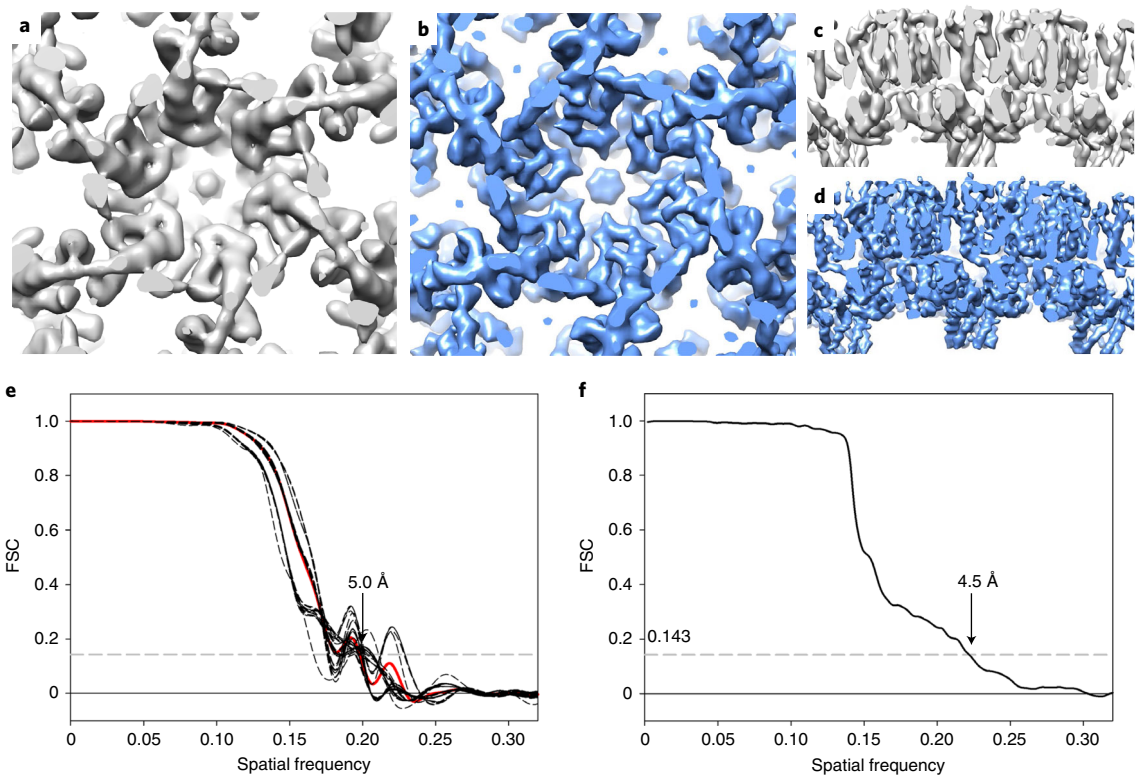

Fig. 4 | Subtomogram averages and conical FSC plots of HIV-1 Gag T8I assemblies (five tilt series). a,**c**, A subtomogram-averaged map of Gag T8I assemblies at 5.0 Å resolution, derived from seven classes (1–7) in Fig. 3, viewed from top (**a**) and side (**c**). **b**,**d**, Reconstruction of Gag T8I assemblies from projection images at 4.5 Å resolution using cisTEM with one cycle of additional translational refinement, viewed from top (**b**) and side (**d**). **e**, Conical FSC plots of the subtomogram-averaged map shown in **a**. The solid red curve represents the global FSC, and each dashed curve represents the FSC of a cone of 36° of half angle, with a 30° increment between each cone. **f**, FSC plot of the cisTEM-reconstructed map shown in **b**.

Further iterative cycles of STA, alignment and tomoCPR were carried out. The resulting final maps were generated using either subtomograms or 2D images with cisTEM, shown in Fig. 4, along with its corresponding FSC plots. cisTEM reconstruction and refinement resulted in a higher-resolution density map (4.5 Å) compared with averaging from subtomograms (5.0 Å) (Fig. 4).

Wild-type Gag

We also reprocessed a published five tilt series of wild-type Gag (EMPIAR-10164, TS_001, 003, 043, 045 and 054), which yielded a subtomogram-averaged map at 3.9 Å resolution previously[19]. The alignment procedure for this dataset is similar to that used for the Gag T8I dataset above, but does not include classification (Table 1 and Supplementary Information 2).

Given that the pixel size (1.35 Å) is slightly larger in this dataset, the iterative alignment step used in emClarity starts from bin4 tomograms and three rounds of tomoCPR were conducted at bin4, bin3 and bin2, respectively. The same alignment mask size `Ali_mRadius=[116,116,72]` encompassing seven hexamers as in the HIV-1 Gag T8I processing was used in the initial averaging/alignment steps. The size was changed to `[88,88,72]` in the last few iterations at bin1 to further improve the resolution. A final sixfold symmetrized map at a resolution of 3.3 Å was obtained, revealing clear side chains of Gag domains (Fig. 5).

Ribosomes

The emClarity processing of the ribosome dataset of isolated single particles (EMPIAR-10304) is included in the software tutorial (https://github.com/ffyr2w/emClarity-tutorial) along with emClarity installation. The tilt series were aligned with emClarity autoAlign function, and particles were picked through template search with bin6 tomograms. Subtomograms within the same subregion were split

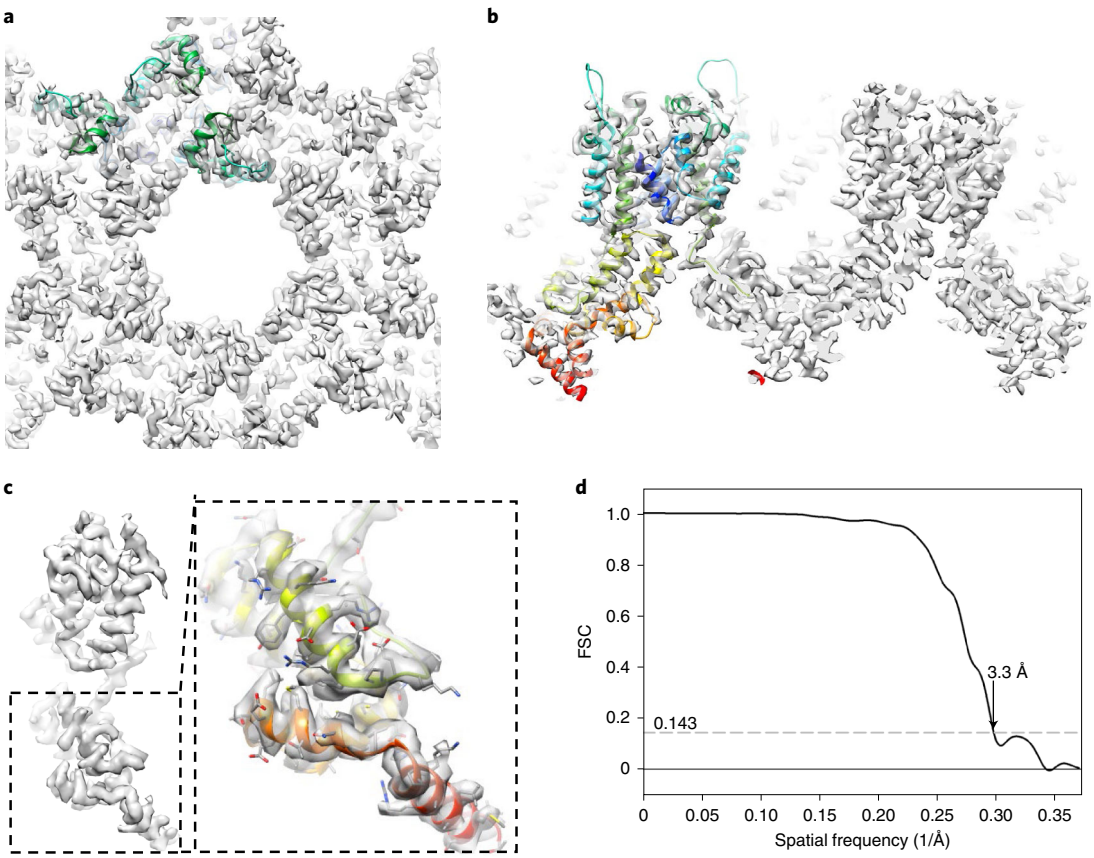

Fig. 5 | STA of WT Gag (five tilt series, EMPIAR-10164). a,**b**, A subtomogram-averaged map of Gag at 3.3 Å resolution. Top (**a**) and side (**b**) sectional views are shown. One asymmetric unit containing three Gag polypeptides is fitted with a Gag structure model (PDB 5l93), colored in rainbow (blue to red) from N-terminal to C-terminal of each polypeptide. **c**, A monomer extracted from the density map, with a close-up view of CTD (carboxy-terminal domain) region overlay with the atomic model (PDB 5l93). **d**, FSC plot of Gag subtomogram-averaged map.

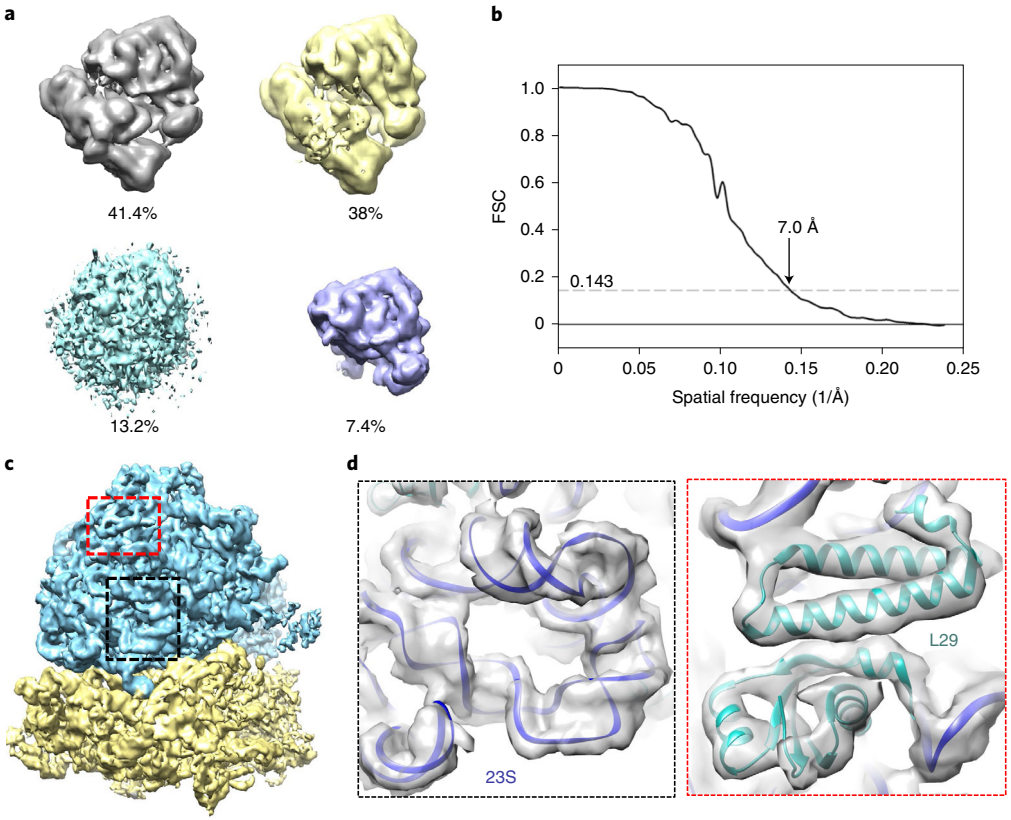

Fig. 6 | Subtomogram classification and averaging of ribosome (12 tilt series, EMPIAR-10304). a, 3D classification of ribosome revealing four classes, with two major classes having both 50S and 30S subunits (gray and yellow), one class with junk particles (cyan) and one with 50S only (light blue). **b**, FSC of final map reconstructed from two major classes. **c**, Density map of ribosome with 50S subunit colored in blue and 30S subunit in yellow. **d**, Zoom-in view of two local regions (dashed black and red boxed) with model fitted in (PDB code 5mdz), as boxed in **c**.

into two random halves since there is no overlap among them (`fscGoldSplitOnTomos=0`). The alignment and averaging were performed iteratively from bin5 to bin1 with one round of tomoCPR before transition to each lower binning. The classification was performed at bin3 to remove junk particles (Fig. 6). Four resolution bands were used for 3D classification (`pcaScaleSpace=[25,50,80,120]`), and several different numbers of classes were tried (2, 3, 4, 6, 8, 14, 18), all of which resulted in classes with junk particles (~13.2%) and a small class (7.4%) containing only the large subunits (Fig. 6a). The final reconstruction and refinement with cisTEM resulted in a 7.0 Å resolution map, showing clear secondary structure elements such as RNA groves and α-helices (Fig. 6b–d).

Apoferritin

The final example is the apoferritin cryoET sample, which was prepared using a graphene-coated EM grid, yielding a mono-dispersed thin layer of apoferritin (Fig. 7a). Tilt series were collected using the parameters presented in Table 1, and the emClarity commands are included in Supplementary Information 2. Six tilt series were aligned with Etomo by patch tracking (no fiducial gold beads) and imported into emClarity. Octahedral symmetry was applied throughout alignment. The final STA map was obtained from <5,000 subtomograms, with 2.86 Å resolution, approaching the Nyquist frequency (2.68 Å) (Fig. 7b–d).

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

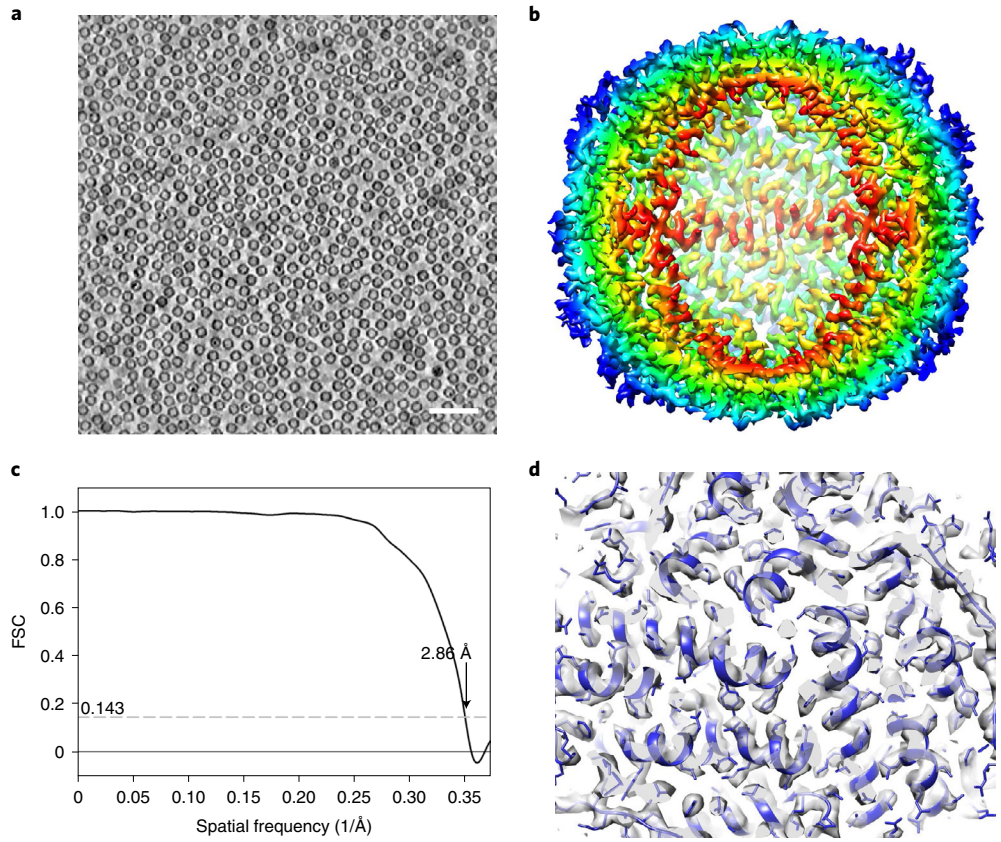

Fig. 7 | STA of apoferritin (six tilt series). a, A tomographic slice of apoferritin on graphene grids. Scale bar, 50 nm. **b**, Subtomogram-averaged map of apoferritin at 2.86 Å resolution. The density map is colored radially 40 Å (red) to 60 Å (blue) from the center. **c**, FSC plot of the averaged map. The resolution is 2.86 Å with an FSC cutoff of 0.143, approaching the Nyquist frequency, which is the edge of the plot. **d**, Representative density maps (fitted with PDB model 6s61).

Data availability

The Gag dataset (five tilt series) and apoferritin dataset (six tilt series) have been deposited in the EMPIAR database under accession codes EMPIAR-10643 and EMPIAR-10787, respectively. The resulting final reconstructions have been deposited in EMDB under the following accession codes: Gag-T8I, EMD-13390; Gag-WT, EMD-13354; apoferritin, EMD-13271; and ribosome, EMD-13270.

Code availability

The emClarity software is freely available at https://github.com/bHimes/emClarity/wiki. The tutorial documentation is available at https://github.com/ffyr2w/emClarity-tutorial.

References

1. Zhang, P. Advances in cryo-electron tomography and subtomogram averaging and classification. *Curr. Opin. Struct. Biol.* **58**, 249–258 (2019).
2. Kaplan, M. et al. In situ imaging and structure determination of biomolecular complexes using electron cryo-tomography. *Methods Mol. Biol.* **2215**, 83–111 (2021).
3. Turk, M. & Baumeister, W. The promise and the challenges of cryo-electron tomography. *FEBS Lett.* **594**, 3243–3261 (2020).
4. Mahamid, J. et al. Visualizing the molecular sociology at the HeLa cell nuclear periphery. *Science* **351**, 969–972 (2016).
5. Forster, F. & Hegerl, R. Structure determination in situ by averaging of tomograms. *Methods Cell Biol.* **79**, 741–767 (2007).
6. Bykov, Y. S. et al. The structure of the COPI coat determined within the cell. *eLife* https://doi.org/10.7554/eLife.32493 (2017).
7. Zhang, Y. et al. Molecular architecture of the luminal ring of the *Xenopus laevis* nuclear pore complex. *Cell Res.* **30**, 532–540 (2020).

8. Pfeffer, S. et al. Structure of the native Sec61 protein-conducting channel. *Nat. Commun.* **6**, 8403 (2015).
9. Cassidy, C. K. et al. CryoEM and computer simulations reveal a novel kinase conformational switch in bacterial chemotaxis signaling. *eLife* https://doi.org/10.7554/eLife.08419 (2015).
10. Dodonova, S. O., Prinz, S., Bilanchone, V., Sandmeyer, S. & Briggs, J. A. G. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. *Proc. Natl Acad. Sci. USA* **116**, 10048–10057 (2019).
11. Mattei, S., Glass, B., Hagen, W. J., Krausslich, H. G. & Briggs, J. A. The structure and flexibility of conical HIV-1 capsids determined within intact virions. *Science* **354**, 1434–1437 (2016).
12. Dick, R. A. et al. Structures of immature EIAV Gag lattices reveal a conserved role for IP6 in lentivirus assembly. *PLoS Pathog.* **16**, e1008277 (2020).
13. Schur, F. K. et al. An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation. *Science* **353**, 506–508 (2016).
14. Qu, K. et al. Structure and architecture of immature and mature murine leukemia virus capsids. *Proc. Natl Acad. Sci. USA* **115**, E11751–E11760 (2018).
15. von Kugelgen, A. et al. In situ structure of an intact lipopolysaccharide-bound bacterial surface layer. *Cell* **180**, 348–358 e315 (2020).
16. Tegunov, D., Xue, L., Dienemann, C., Cramer, P. & Mahamid, J. Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 A in cells. *Nat. Methods* **18**, 186–193 (2021).
17. Lucic, V., Rigort, A. & Baumeister, W. Cryo-electron tomography: the challenge of doing structural biology in situ. *J. Cell Biol.* **202**, 407–419 (2013).
18. Wan, W. & Briggs, J. A. Cryo-electron tomography and subtomogram averaging. *Methods Enzymol.* **579**, 329–367 (2016).
19. Turonova, 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 3.4A. *J. Struct. Biol.* **199**, 187–195 (2017).
20. 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).
21. Nicastro, D. et al. The molecular architecture of axonemes revealed by cryoelectron tomography. *Science* **313**, 944–948 (2006).
22. Chen, M. et al. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. *Nat. Methods* **14**, 983–985 (2017).
23. Galaz-Montoya, J. G., Flanagan, J., Schmid, M. F. & Ludtke, S. J. Single particle tomography in EMAN2. *J. Struct. Biol.* **190**, 279–290 (2015).
24. Galaz-Montoya, J. G. et al. Alignment algorithms and per-particle CTF correction for single particle cryo-electron tomography. *J. Struct. Biol.* **194**, 383–394 (2016).
25. Bharat, T. A. & Scheres, S. H. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. *Nat. Protoc.* **11**, 2054–2065 (2016).
26. Bharat, T. A. M., Russo, C. J., Lowe, 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).
27. Castano-Diez, 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).
28. 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).
29. Forster, F., Pruggnaller, S., Seybert, A. & Frangakis, A. S. Classification of cryo-electron sub-tomograms using constrained correlation. *J. Struct. Biol.* **161**, 276–286 (2008).
30. 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).
31. Winkler, H. 3D reconstruction and processing of volumetric data in cryo-electron tomography. *J. Struct. Biol.* **157**, 126–137 (2007).
32. Himes, B. A. & Zhang, P. emClarity: software for high-resolution cryo-electron tomography and sub-tomogram averaging. *Nat. Methods* **15**, 955–961 (2018).
33. 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 e1214 (2020).
34. Watanabe, R. et al. The in situ structure of Parkinson's disease-linked LRRK2. *Cell* **182**, 1508–1518 e1516 (2020).
35. Sutton, G. et al. Assembly intermediates of orthoreovirus captured in the cell. *Nat. Commun.* **11**, 4445 (2020).
36. Tan, T. Y. et al. Capsid protein structure in Zika virus reveals the flavivirus assembly process. *Nat. Commun.* **11**, 895 (2020).
37. Unchwaniwala, N. et al. Subdomain cryo-EM structure of nodaviral replication protein A crown complex provides mechanistic insights into RNA genome replication. *Proc. Natl Acad. Sci. USA* **117**, 18680–18691 (2020).
38. Gibson, K. H. et al. An asymmetric sheath controls flagellar supercoiling and motility in the leptospira spirochete. *eLife* https://doi.org/10.7554/eLife.53672 (2020).
39. Cassidy, C. K. et al. Structure and dynamics of the *E. coli* chemotaxis core signaling complex by cryo-electron tomography and molecular simulations. *Commun. Biol.* **3**, 24 (2020).

40. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
41. Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* **9**, 853–854 (2012).
42. 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).
43. Mastronarde, D. N. & Held, S. R. Automated tilt series alignment and tomographic reconstruction in IMOD. *J. Struct. Biol.* **197**, 102–113 (2017).
44. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
45. Grant, T., Rohou, A. & Grigorieff, N. cisTEM, user-friendly software for single-particle image processing. *eLife* https://doi.org/10.7554/eLife.35383 (2018).
46. Heymann, J. B. Guidelines for using Bsoft for high resolution reconstruction and validation of biomolecular structures from electron micrographs. *Protein Sci.* **27**, 159–171 (2018).
47. Mendonca, L. et al. CryoET structures of immature HIV Gag reveal six-helix bundle. *Commun. Biol.* **4**, 481 (2021).

Acknowledgements

We are grateful to Y. Zhu for discussion and critical reading of the manuscript. We acknowledge Diamond for access and support of the CryoEM facilities at the UK national Electron Bio-Imaging Centre (eBIC, proposal CM26464), funded by the Wellcome Trust, Medical Research Council (MRC) and Biotechnology and Biological Sciences Research Council (BBSRC). The computational aspects of this research were supported by the Wellcome Trust Core Award grant number 203141/Z/16/Z and the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). This work was supported by the National Institutes of Health grants AI150481, the UK Wellcome Trust Investigator Award 206422/Z/17/Z, the UK Biotechnology and Biological Sciences Research Council grant BB/S003339/1, and the European Research Council Advanced Grant (ERC AdG) grant 101021133.

Author contributions

P.Z. conceived the research and designed the experiments. Y.S. prepared the apoferritin on graphene grids, and D.C. collected data. T.N., L.M. and Y.S. performed tomography reconstruction and STA and classification. T.F. wrote the emClarity tutorial. B.A.H. and T.F. updated code/binaries with new features in later versions of emClarity. T.N. and P.Z. wrote the manuscript with support from all the authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41596-021-00648-5.

Correspondence and requests for materials should be addressed to Peijun Zhang.

Peer review information *Nature Protocols* thanks Peter J. Peters and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 23 February 2021; Accepted: 8 October 2021;
Published online: 12 January 2022

Related links

Key references using this protocol

Liu, C. et al. *Structure* **28**, 1218–1224.e1214 (2020): https://doi.org/10.1016/j.str.2020.10.001
Watanabe, R. et al. *Cell* **182**, 1508–1518.e1516 (2020): https://doi.org/10.1016/j.cell.2020.08.004
Sutton, G. et al. *Nat. Commun.* **11**, 4445 (2020): https://doi.org/10.1038/s41467-020-18243-9
Tan, T. Y. et al. *Nat. Commun.* **11**, 895 (2020): https://doi.org/10.1038/s41467-020-14647-9
Unchwaniwala, N. et al. *Proc. Natl Acad. Sci. USA* **117**, 18680–18691 (2020): https://doi.org/10.1073/pnas.2006165117
Gibson, K. H. et al. *eLife* **9**, e53672 (2020): https://doi.org/10.7554/eLife.53672
Cassidy, C. K. et al. *Commun. Biol.* **3**, 24 (2020): https://doi.org/10.1038/s42003-019-0748-0

Key data used in this protocol

Eisenstein, F. et al. *J. Struct. Biol.* **208**, 107–114 (2019): https://doi.org/10.1016/j.jsb.2019.08.006
Schur, F. K. et al. *Science* **353** 506–508 (2016): https://doi.org/10.1126/science.aaf9620

nature research

Corresponding author(s): Peijun Zhang

Last updated by author(s): September 19, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	CryoEM data were collected using Thermo Fisher Tomo5
Data analysis	MotionCor2, eTomo, emClarity

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The Gag T8I dataset (5 tilt-series) and apoferritin dataset (6 tilt-series) have been deposited in EMPIAR database under accession codes EMPIAR-10643 and EMPIAR-10787, respectively. The resulting final reconstructions have been deposited in EMDB under the following accession codes: Gag-T8I, EMD-13390; Gag-WT, EMD-13354; apoferritin, EMD-13271; and ribosome, EMD-13270.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For cryoEM structure determination, sample sizes were those required for the resolution. The details of datasets, including sample sizes, are listed in table 1.
Data exclusions	Subtomograms closer than half the particle size were excluded on the basis that they could represent duplicate particles.
Replication	For cryoEM, two randomly divided half datasets were processed independently, and combined to give rise to the final structures. The resolution of the structure is assessed by comparing the two independent maps.
Randomization	CryoEM particles were randomly divided into ODD and EVEN datasets, as standard approach implemented in emClarity.
Blinding	No blinding

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging