Toward an integrated pipeline for CryoET data processing

Muyuan Chen
Baylor College of Medicine
2018-11

CryoET data processing pipeline

- Tilt series alignment / tomogram reconstruction
- Tomogram annotation / particle extraction
- CTF determination / CTF correction
- Initial model generation
- Subtomogram refinement
- Heterogeneity analysis

CryoET data processing pipeline

- Tilt series alignment / tomogram reconstruction
  IMOD, RAPTOR, … **IMOD (WBP, SIRT)**
- Tomogram annotation / particle extraction
  Amira, EMAN2 **PyTom, EMAN2, …**
- CTF determination / CTF correction
  CTFFIND,… **IMOD,…**
- Initial model generation
  PyTom?
- Subtomogram refinement
  PEET, Relion, PyTom, emClarity, etc…
- Heterogeneity analysis
  Relion, emClarity…

Integrated CryoET data processing pipeline

- Tilt series alignment / tomogram reconstruction
  EMAN2 **EMAN2**
- Tomogram annotation / particle extraction
  EMAN2 **EMAN2**
- CTF determination / CTF correction
  EMAN2 **EMAN2**
- Initial model generation
  EMAN2
- Subtomogram refinement
  EMAN2
- Heterogeneity analysis
  EMAN2

Benefit of integration:

- User friendly
- Book-keeping
- Automation
- Performance

Tilt series alignment

Start: Unaligned tiltseries

Coarse alignment

Generate tomogram

Pick landmarks (3D)

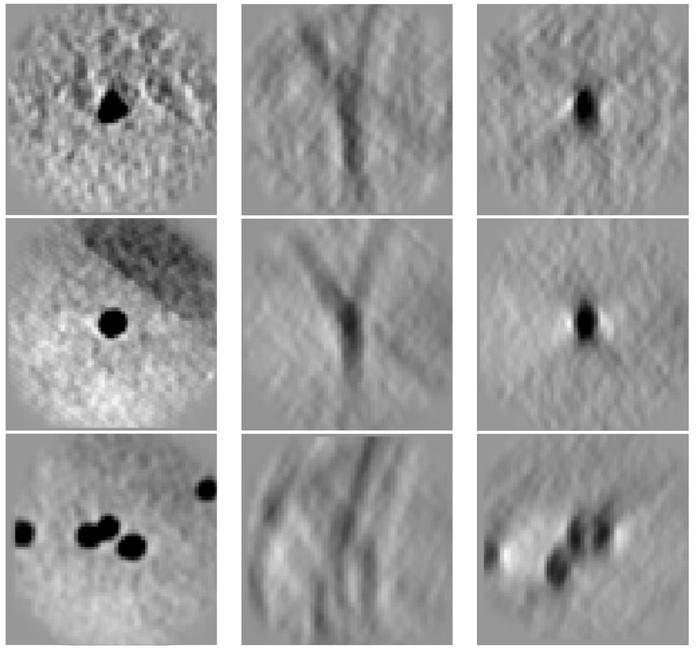
Refine landmark coordinates

Refine alignent parameters

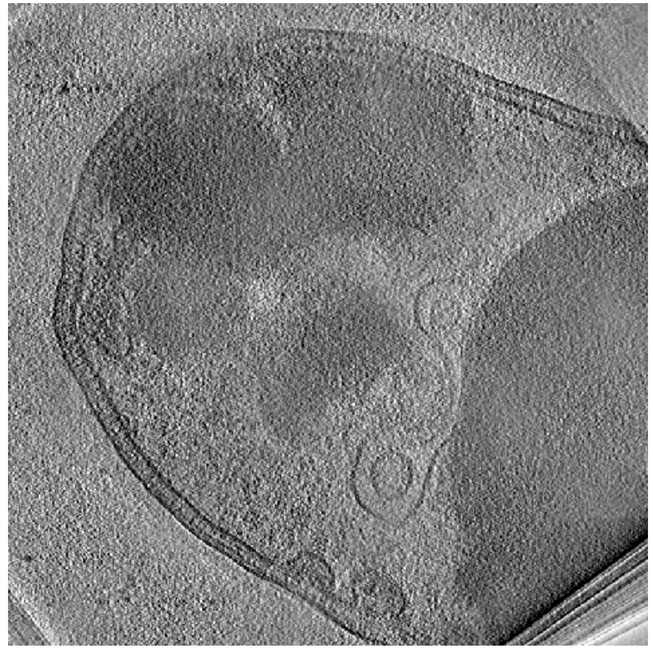
Bin by 8x 4x 2x 1x

Landmarks

Top view | **Side view iteration #1** | **Side view iteration #4**



Tomogram slice view

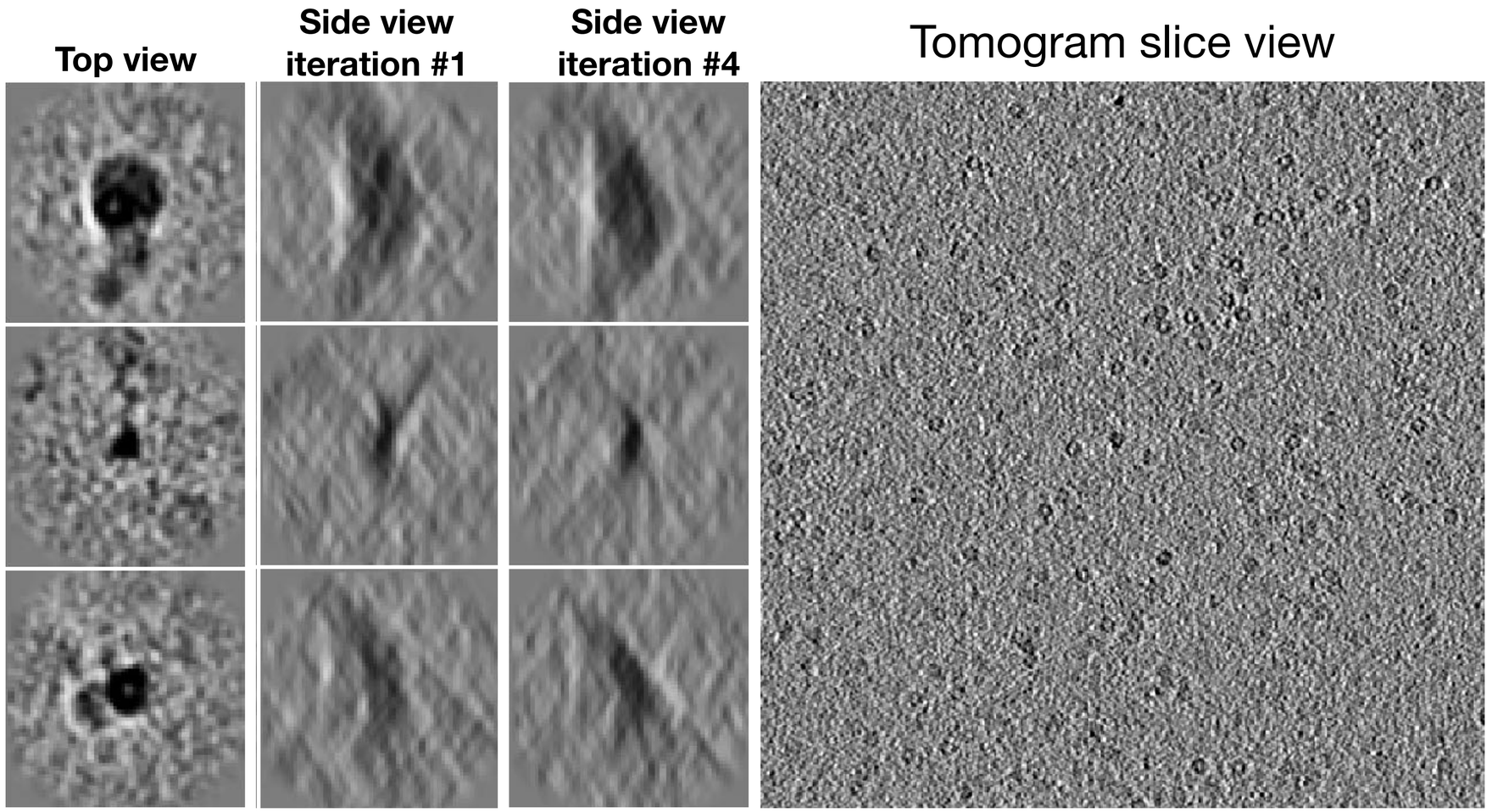


E.coli over expressing AcrAB-TolC pump

Z. Wang, X. Shi, BCM

~5-10 mins per tomogram

Fiducial-less alignment

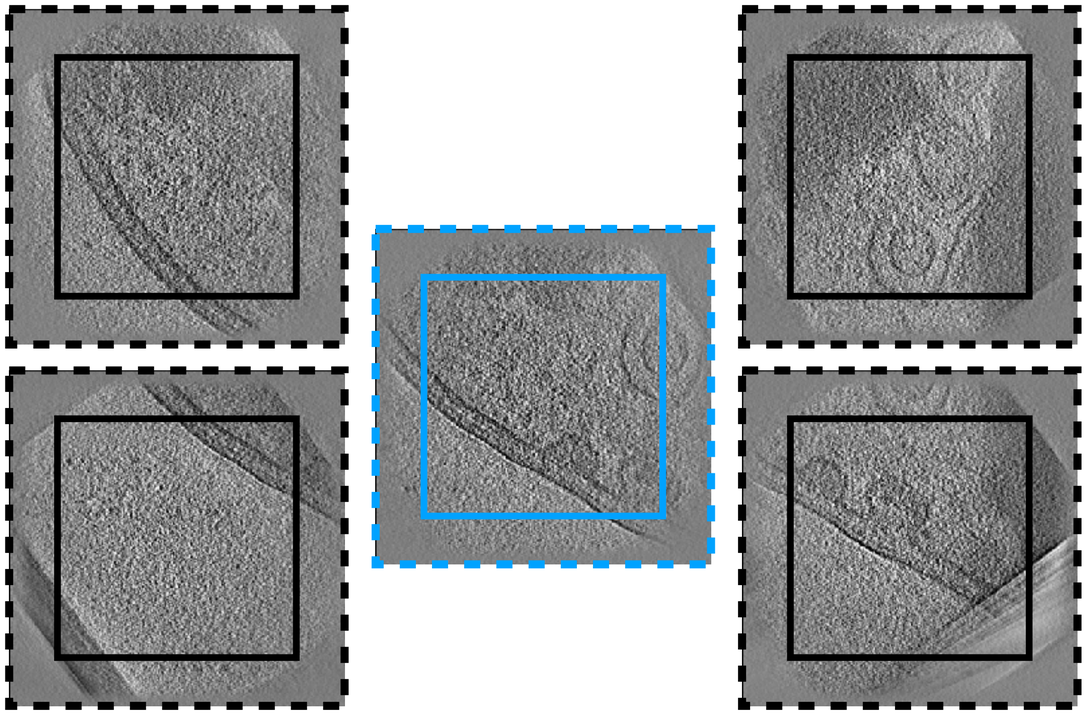


EMPIAR-10171, Apoferritin

Reconstruction via tiled direct Fourier inversion

- Reconstruction via tiled direct Fourier inversion

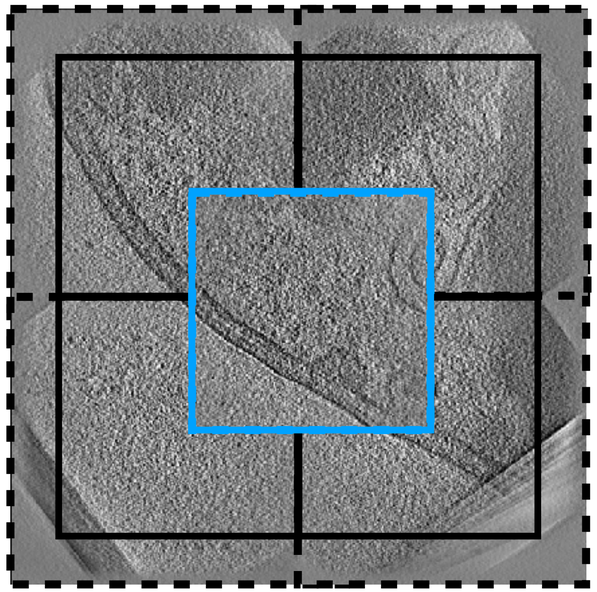
- Normally only generate 1K or 2K output
  -> for visualization and annotation

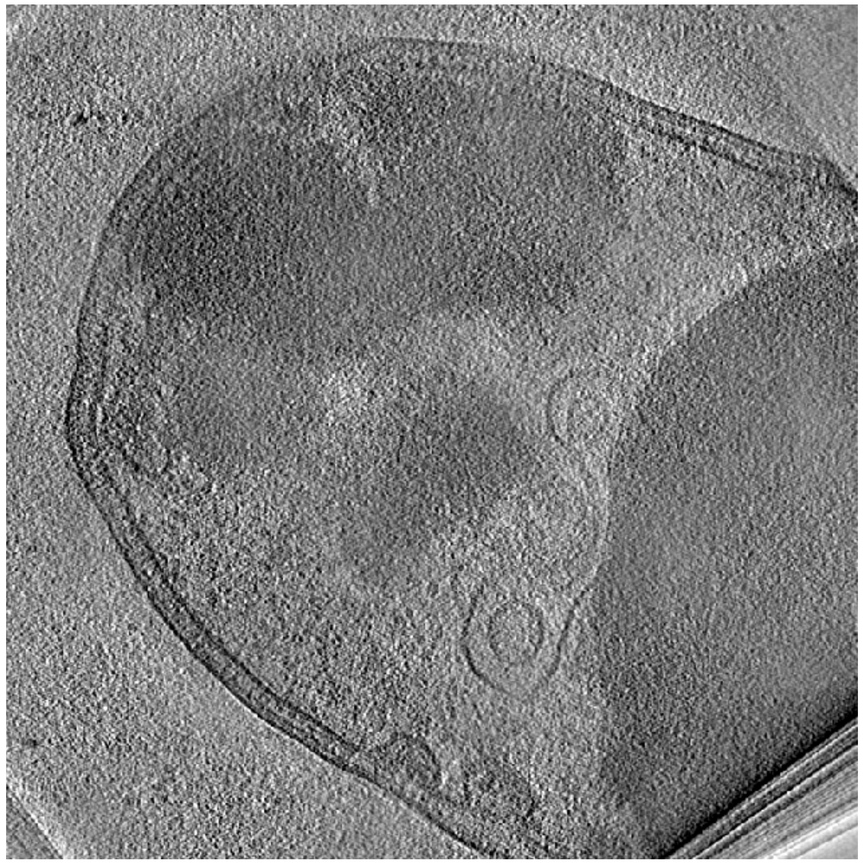


Reconstruction via tiled direct Fourier inversion

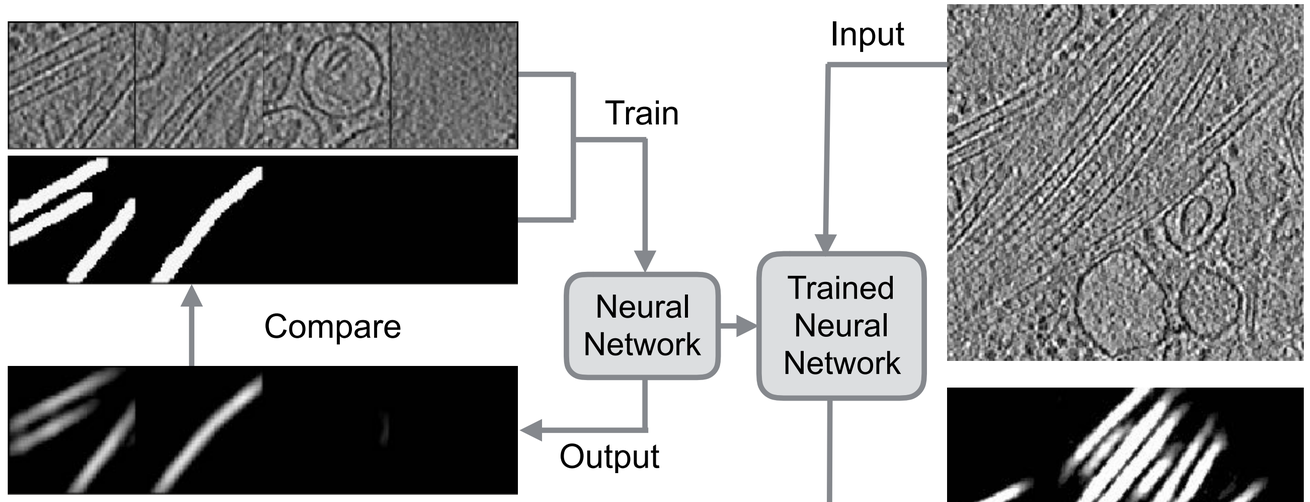
- Reconstruction via tiled direct Fourier inversion

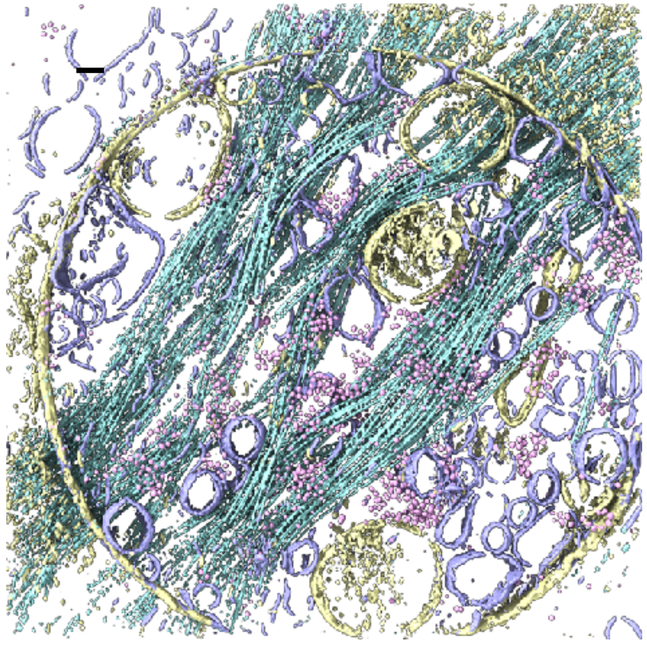
- Normally only generate 1K or 2K output
  -> for visualization and annotation





Tomogram annotation





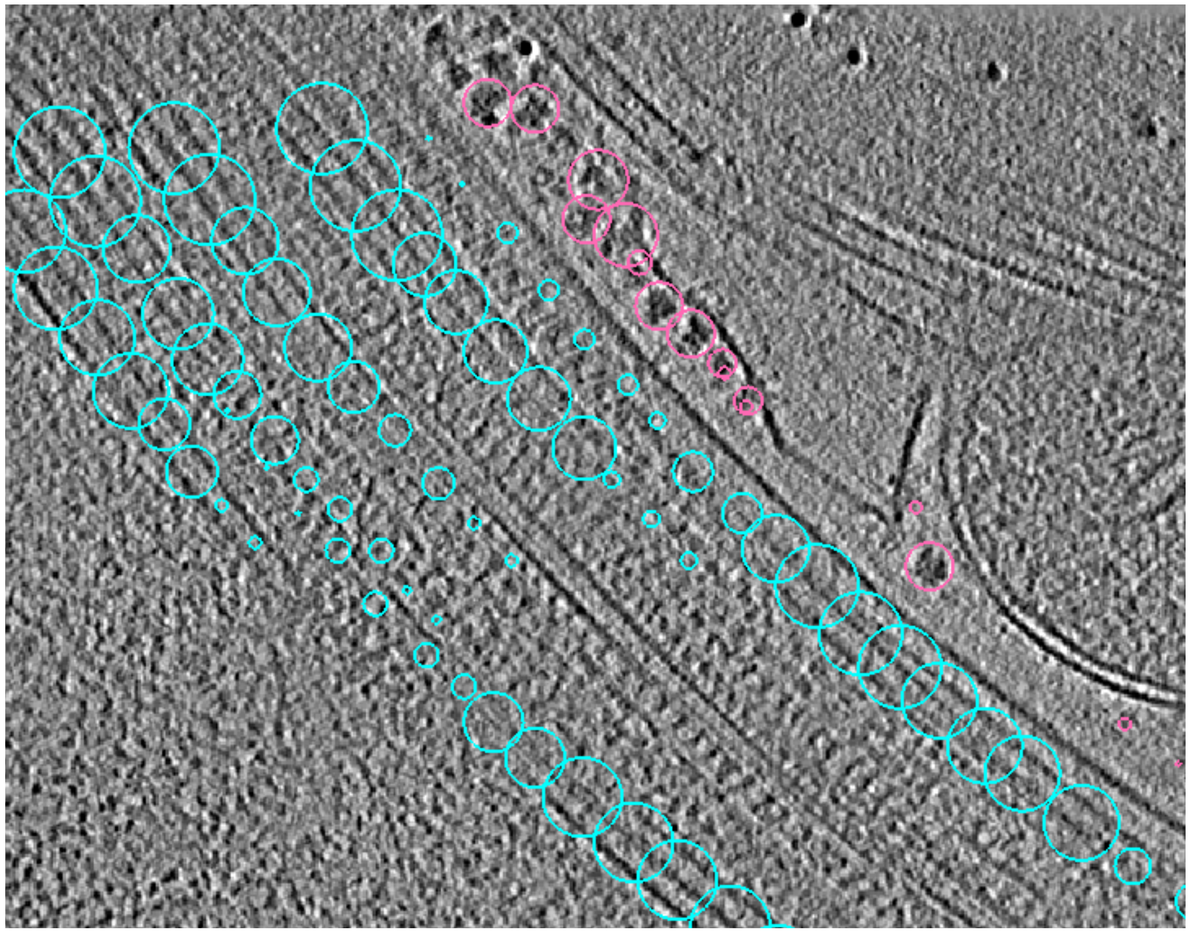
EMD-8594, PC12 cell

CNN based auto-annotation

What's new?

- Integration with tomogram reconstruction
- New particle selection interface
- Better bookkeeping of trained networks and segmentations
- **Directly extract particle coordinates from annotation**

Particle extraction



Trypanosome flagella and cell body

S. Y. Sun, Stanford

To locate particles:

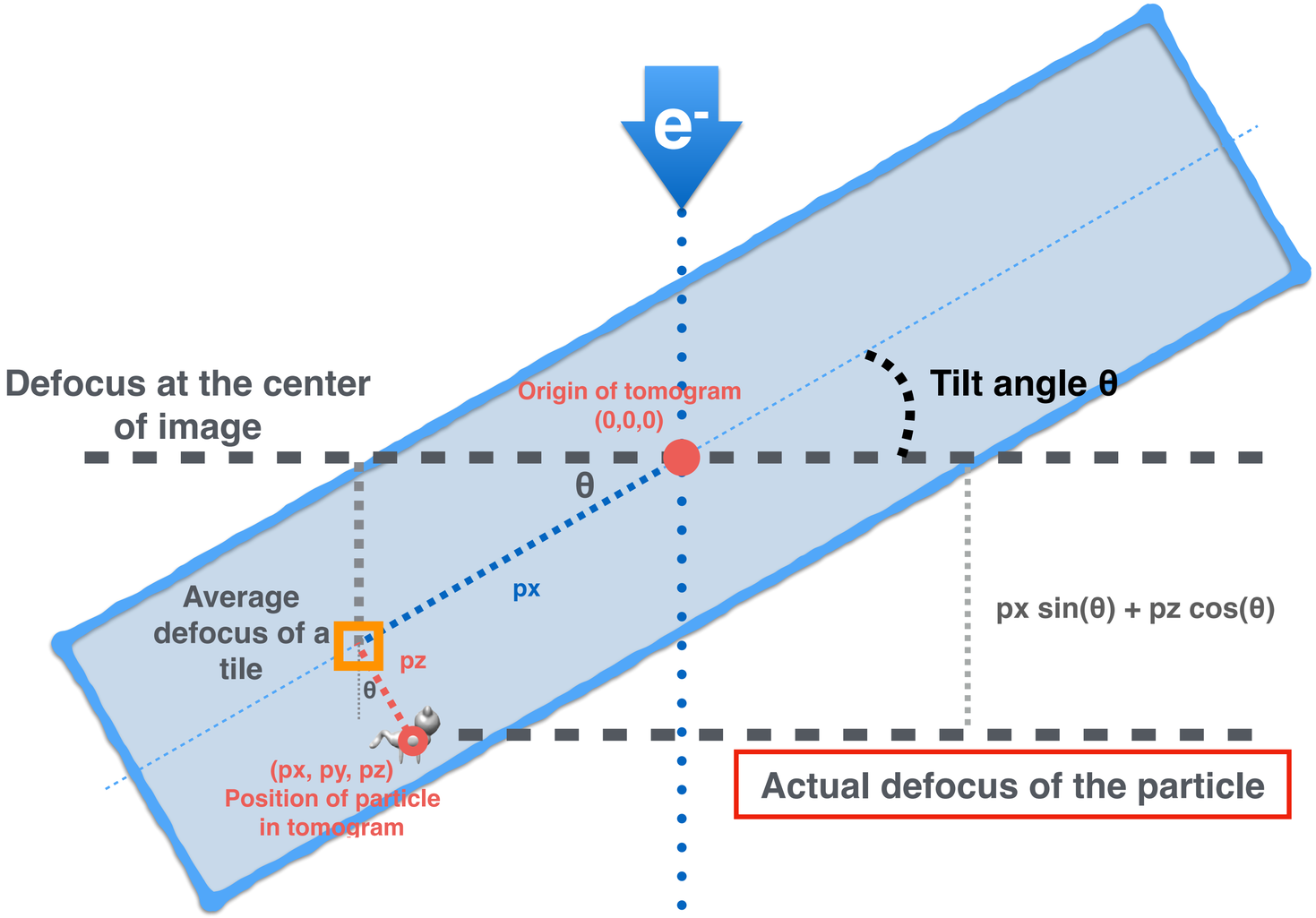
- Manual particle picking / extract coordinates from annotation / template matching
- Select different types of particles in the same tomogram at different scaling/filtration
- How to get unbinned particles since we only make binned tomograms?

Particle extraction

To generate 3D particles:

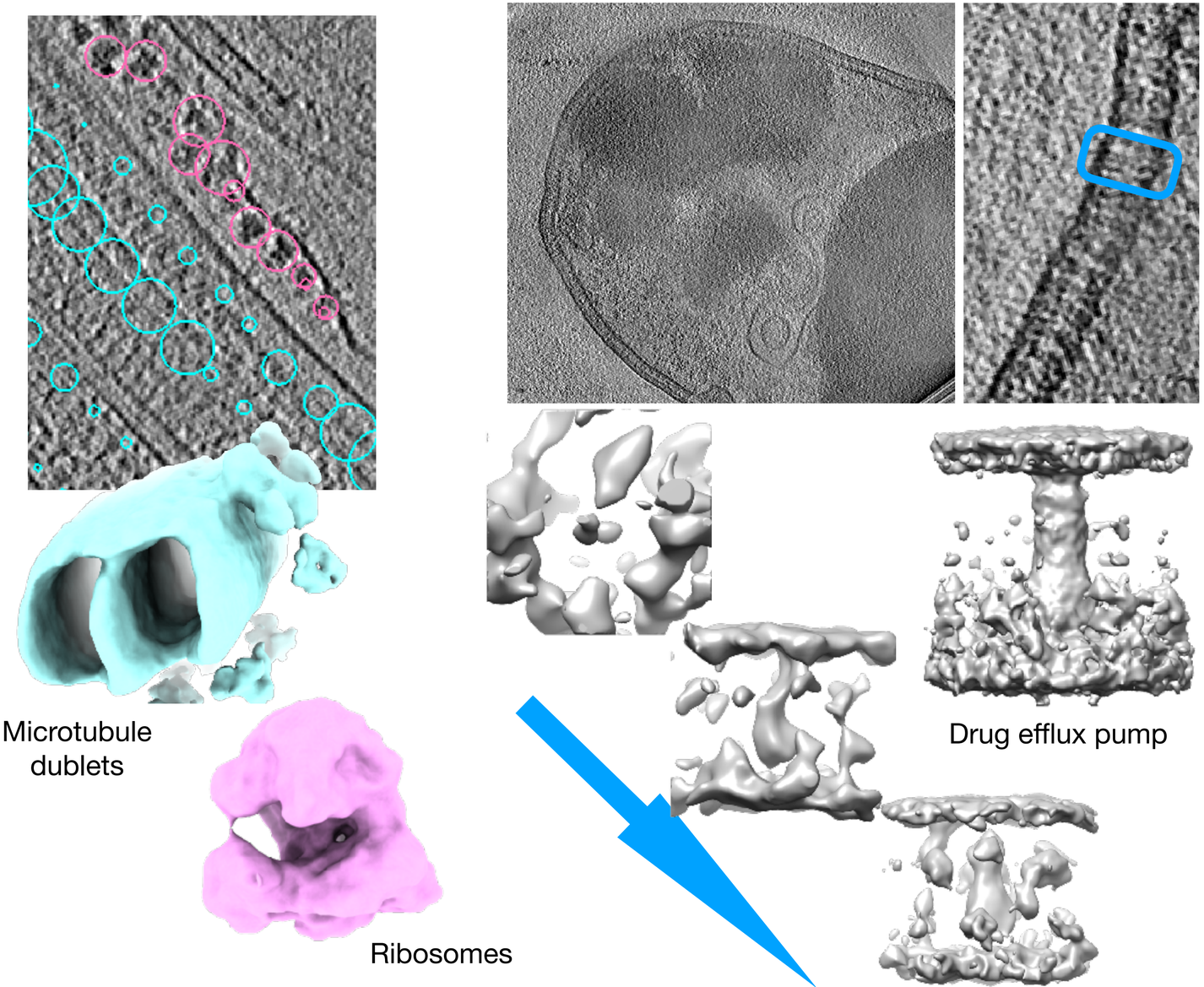
- Map particle coordinates back to raw tilt series and extract sub-tilt series for each 3D particle
- Reconstruct 3D particles from the sub-tilt series via direct Fourier inversion
- **But before this, CTF correction…**

CTF determination and correction



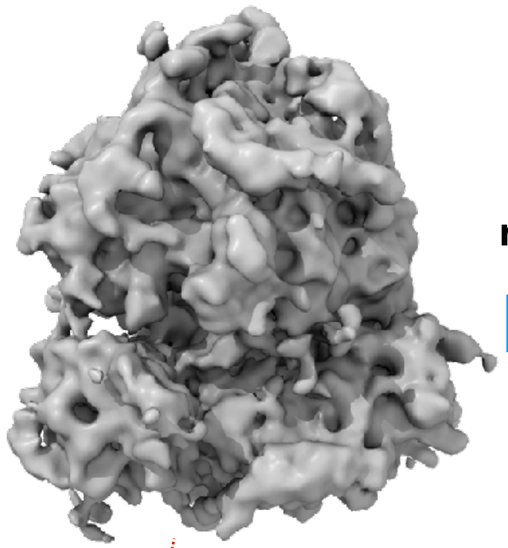
- Estimate the defocus at the center of image using all information from the micrograph
- Calculate defocus of each particle
- Phase flip and Wiener filter the 2D sub-tilt series of each particle before reconstructing them into the 3D particle

Initial model generation

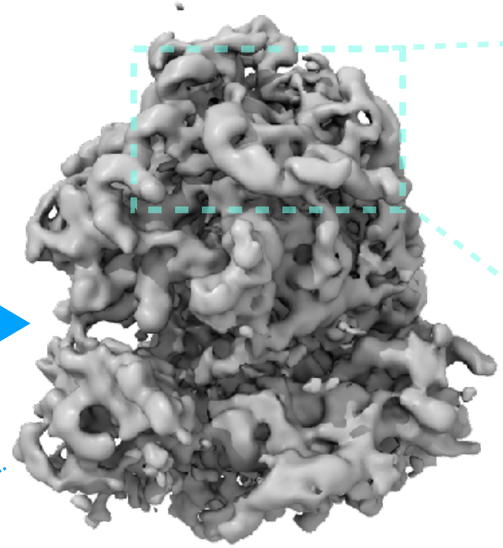


- Stochastic gradient descent
- Build good initial model from ~30 raw particles
- Works on proteins of various shapes without external information

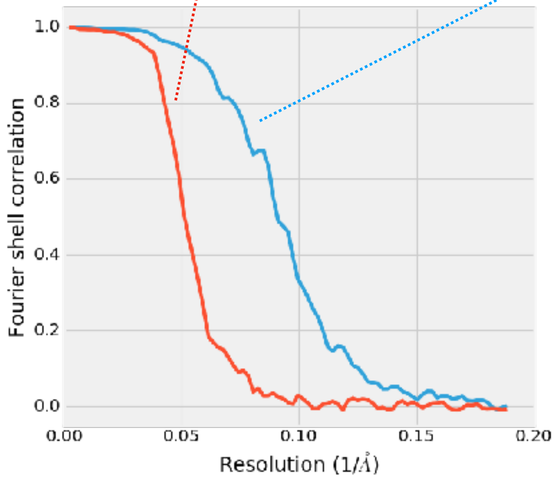
Subtomogram refinement



sub-tilt refinement



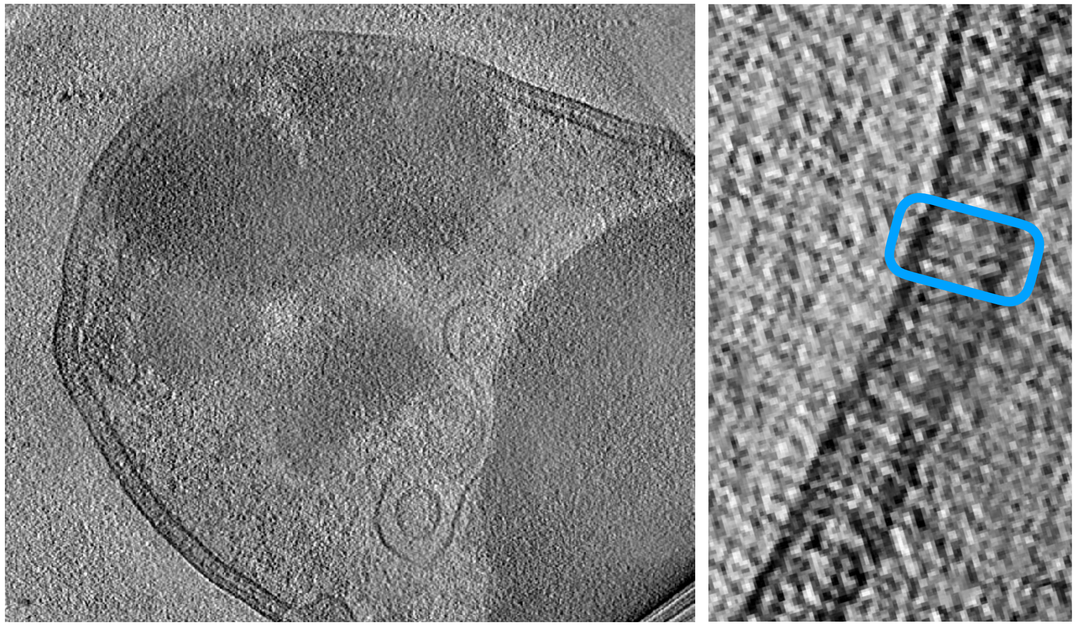
EMPIAR 10064: 3239 particles, 13Å -> 8.5Å

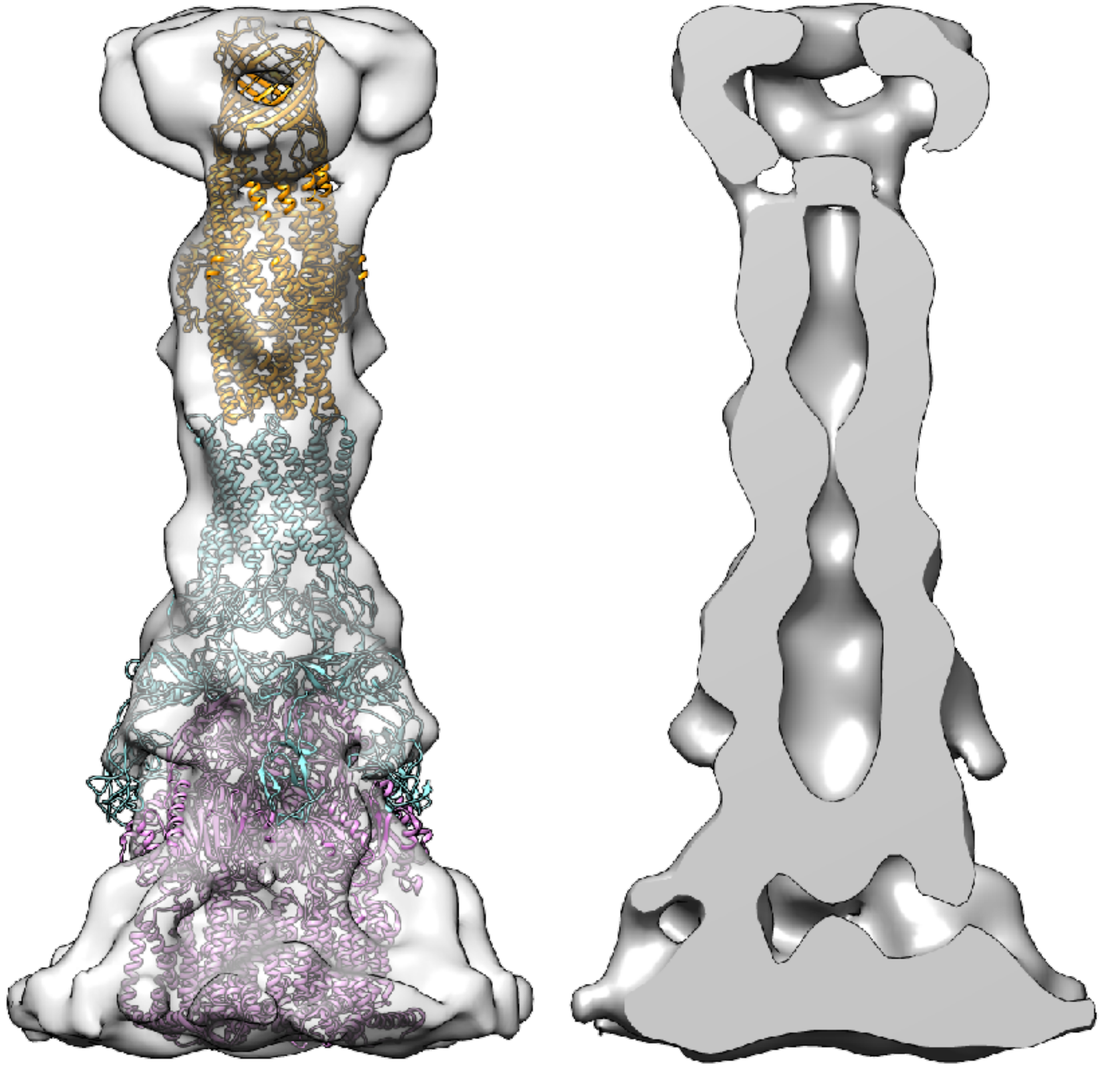


Two-step approach:

- Determine and refine the transform of each subtomogram
- From the subtomogram orientation, locally refine the transforms of the 2D sub-tilt series corresponding to each particle (per-particle-per-tilt refinement)

Subtomogram refinement

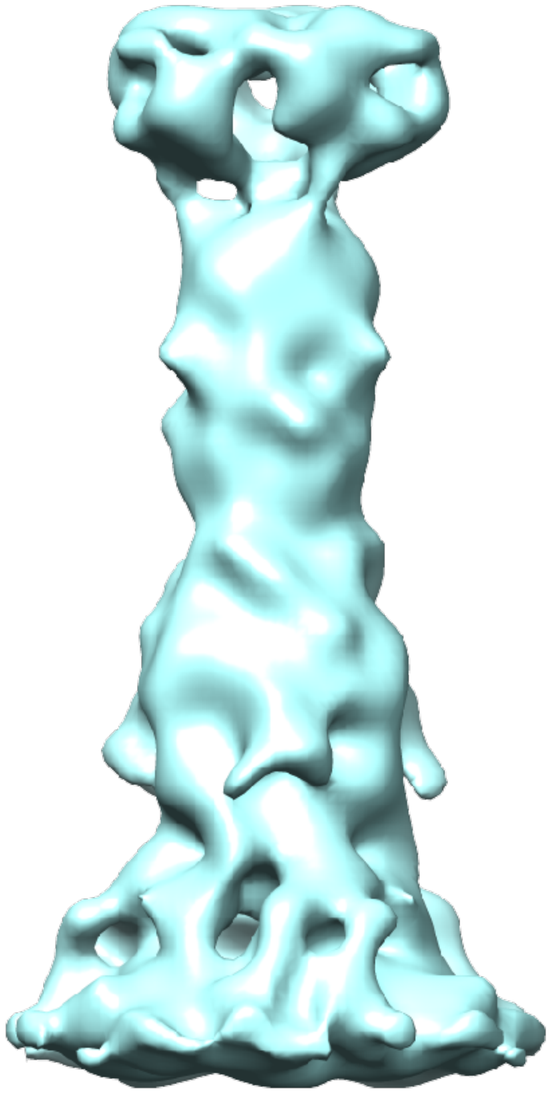


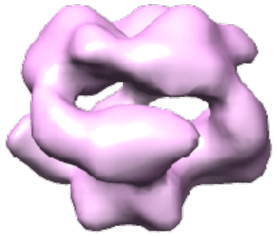


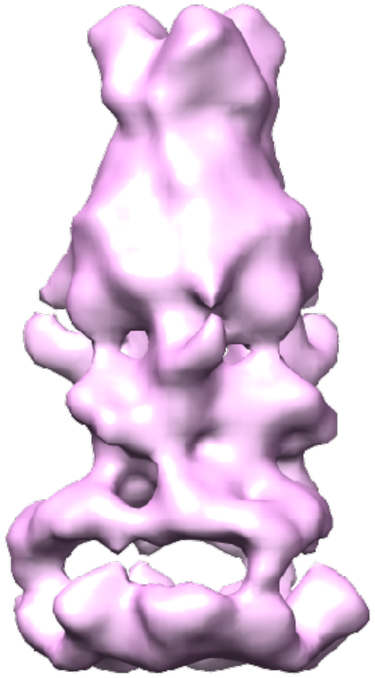
AcrAB-TolC pump on e.coli:

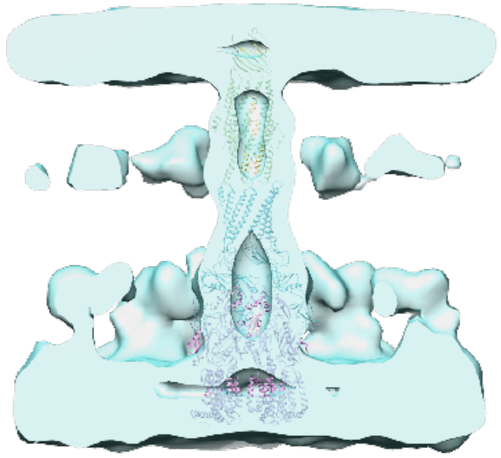
1321 particles, c3 symmetry, 19Å -> 14Å

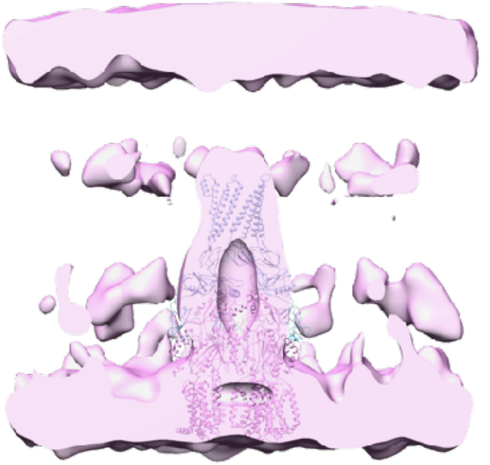
Heterogeneity analysis

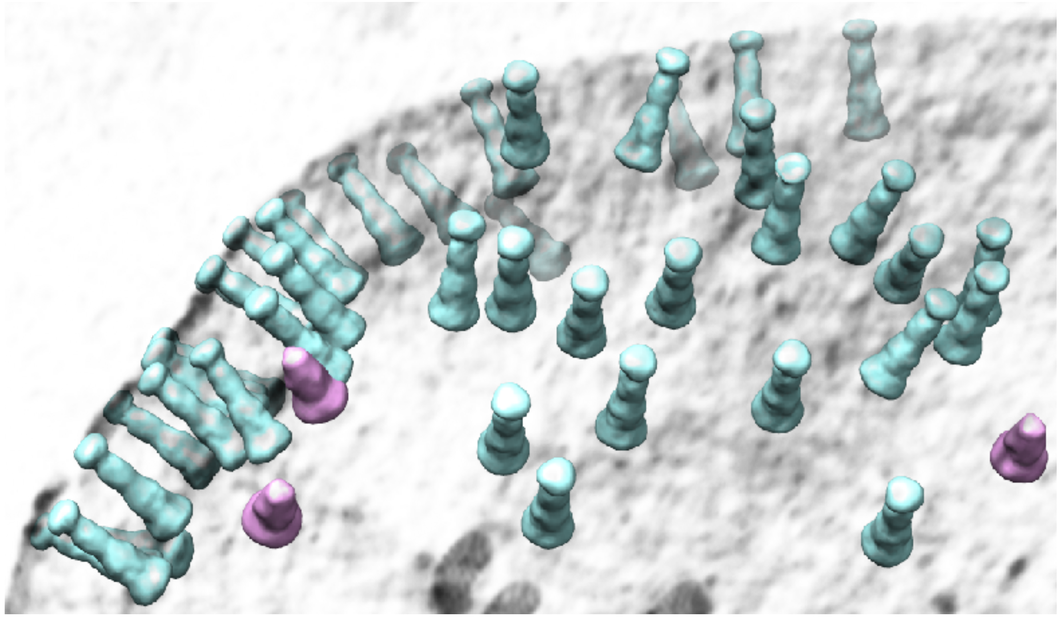




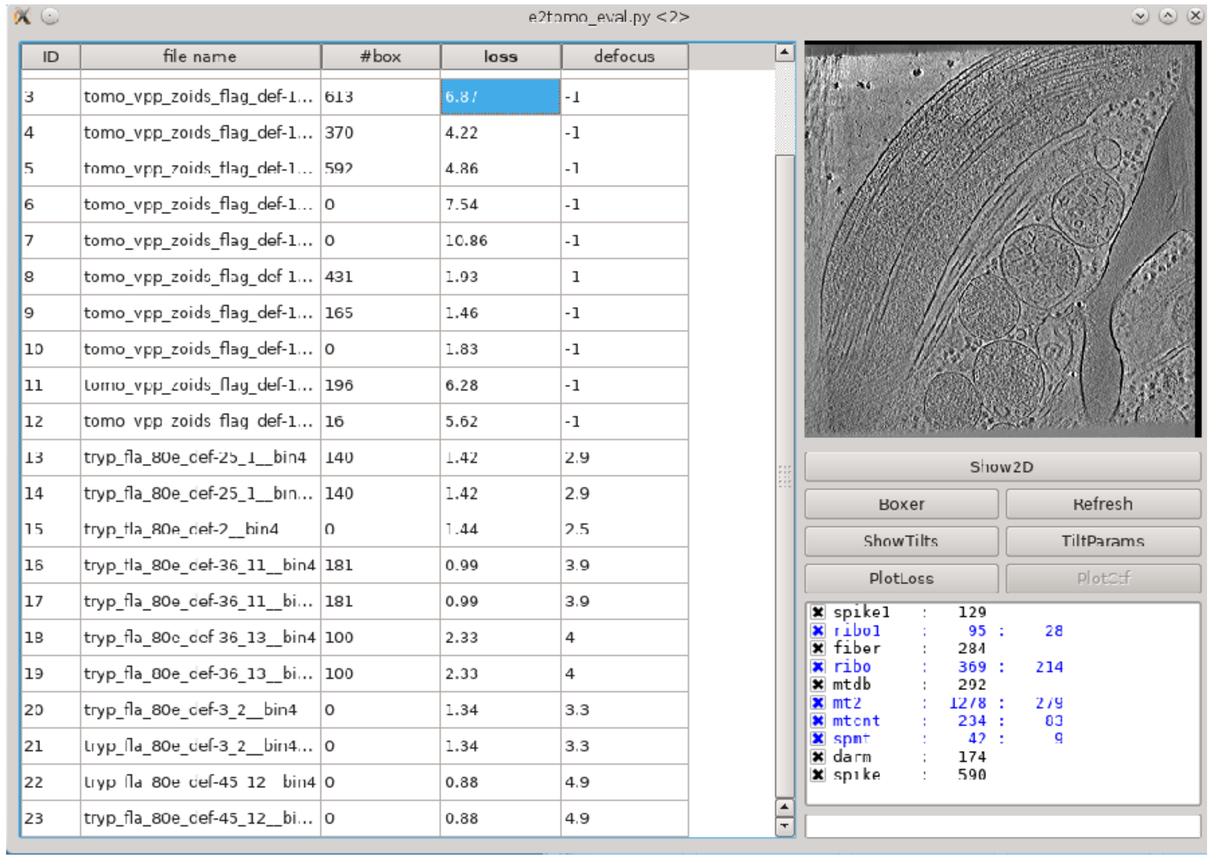


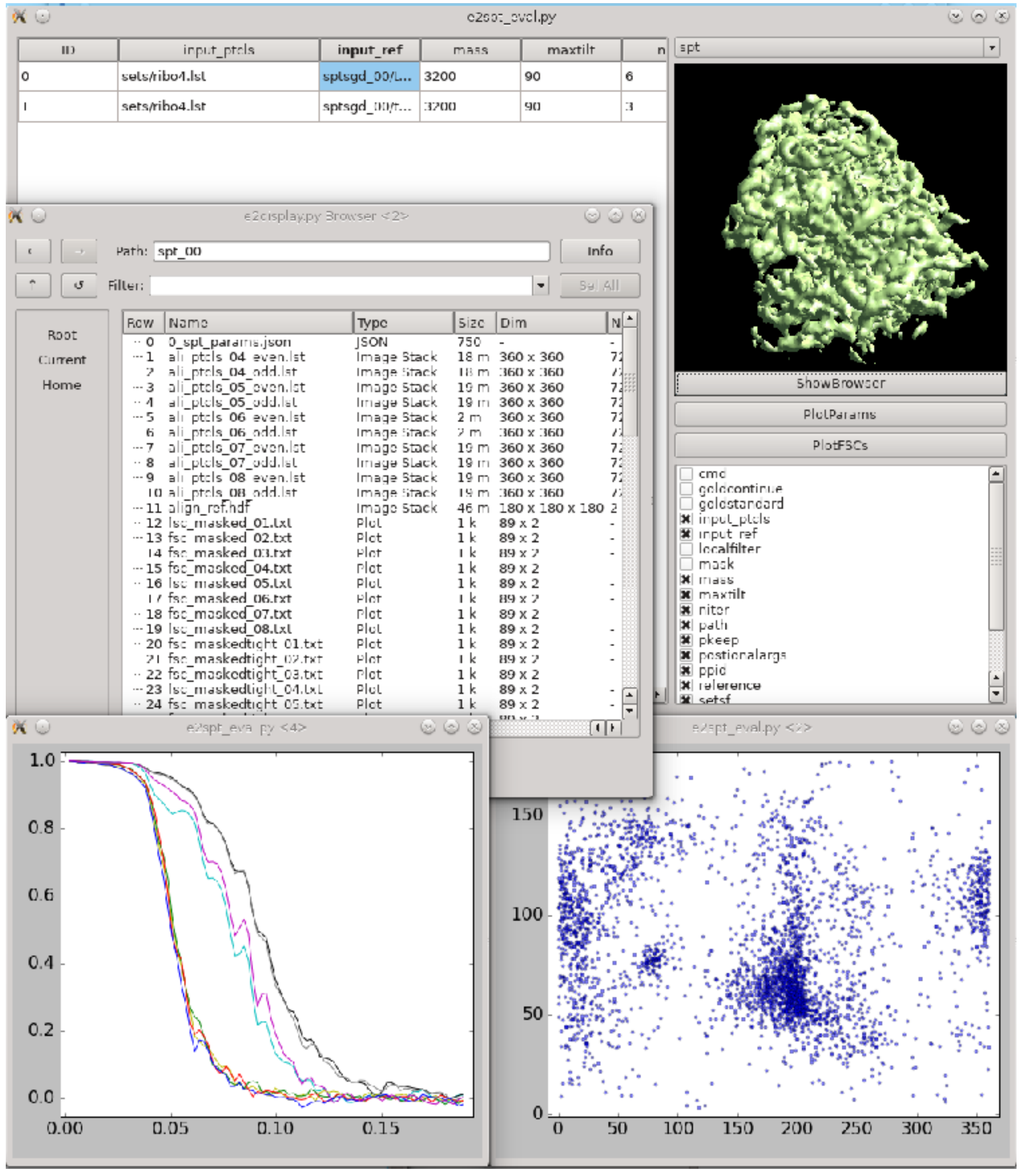






Project management utilities





Many features per tomogram,

many tomograms per projects…

Future directions

- Integrate to data collection
- Smaller proteins, higher resolution
- Identify unknown protein from cellular tomograms
- Solve continuous heterogeneity *in situ*

Acknowledgement

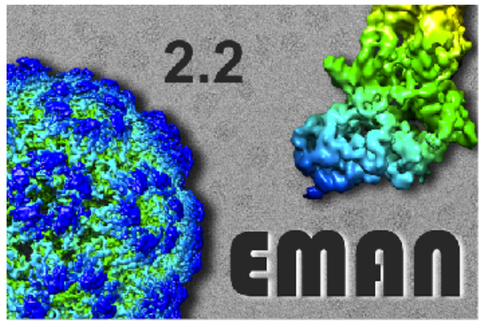
Developers

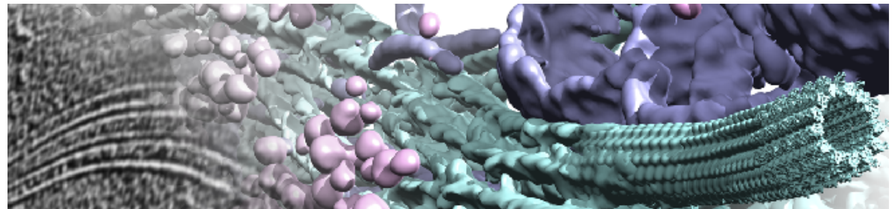
- Ludtke lab (BCM):
  - Steven Ludtke
  - Michael Bell
- NIH: R01GM080139

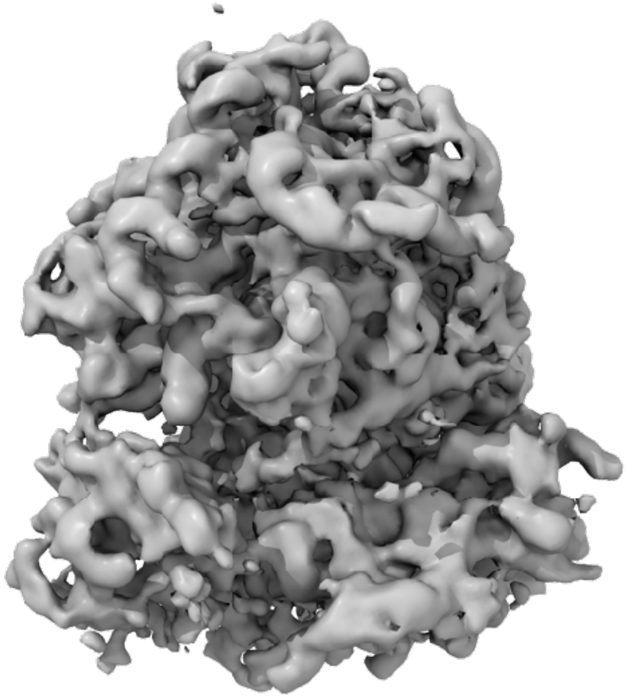
Data providers:

- Wang lab (BCM):
  - Zhao Wang
  - Xiaodong Shi
- Chiu lab (Stanford):
  - Stella Y. Sun
- NIH: P41GM103832

Also special thanks to all early users and their valuable feedbacks…







Task	Program name	# Cores	Walltime (min)	Iterations
Raw data import	e2import.py		1	
Tomographic reconstruction	e2tomogram.py †	12	9	2,1,1,1
Reference-based particle picking	e2spt_tempmatch.py		7	
CTF correction	e2spt_tomoctf.py		2	
Subtomogram extraction	e2spt_extract.py †	1	31	
Initial model generation	e2spt_sgd.py †	12	41	3
Subtomogram refinement	e2spt_refine.py †	12	181	3
Subtilt refinement	e2spt_tiltrefine.py †*	96	308	6

Parallelism: * = MPI, † = Thread. Note, e2spt_sgd.py is parallelized by batch, so running the program with a batch size of 12 will use 12 threads.

EMPIAR 10064: 3239 particles, 13Å -> 8.5Å