A Template-free Pipeline for Recovering Structures in Cryo-electron tomography

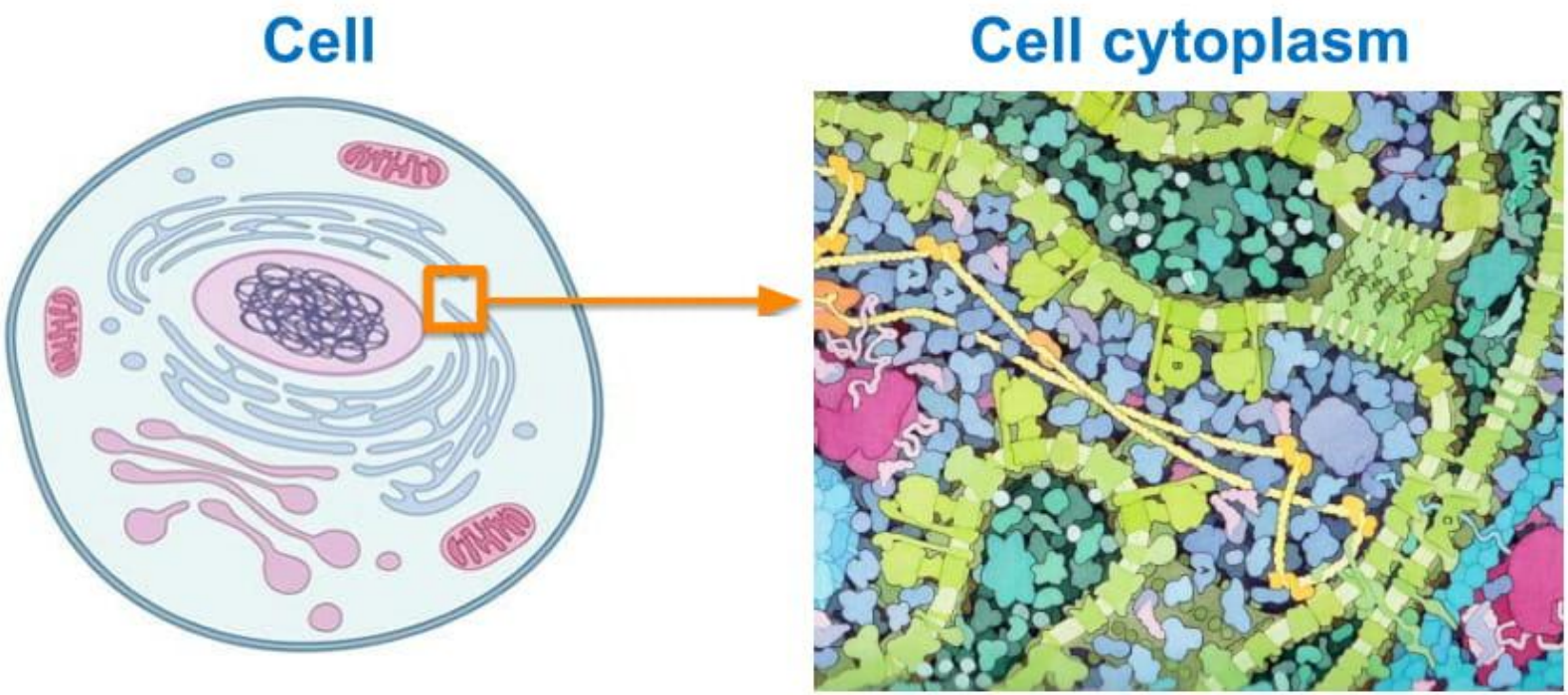
Xiangrui Zeng

Computational Biology Department
School of Computer Science
Carnegie Mellon University

MMBioS Monthly Scientific Meeting

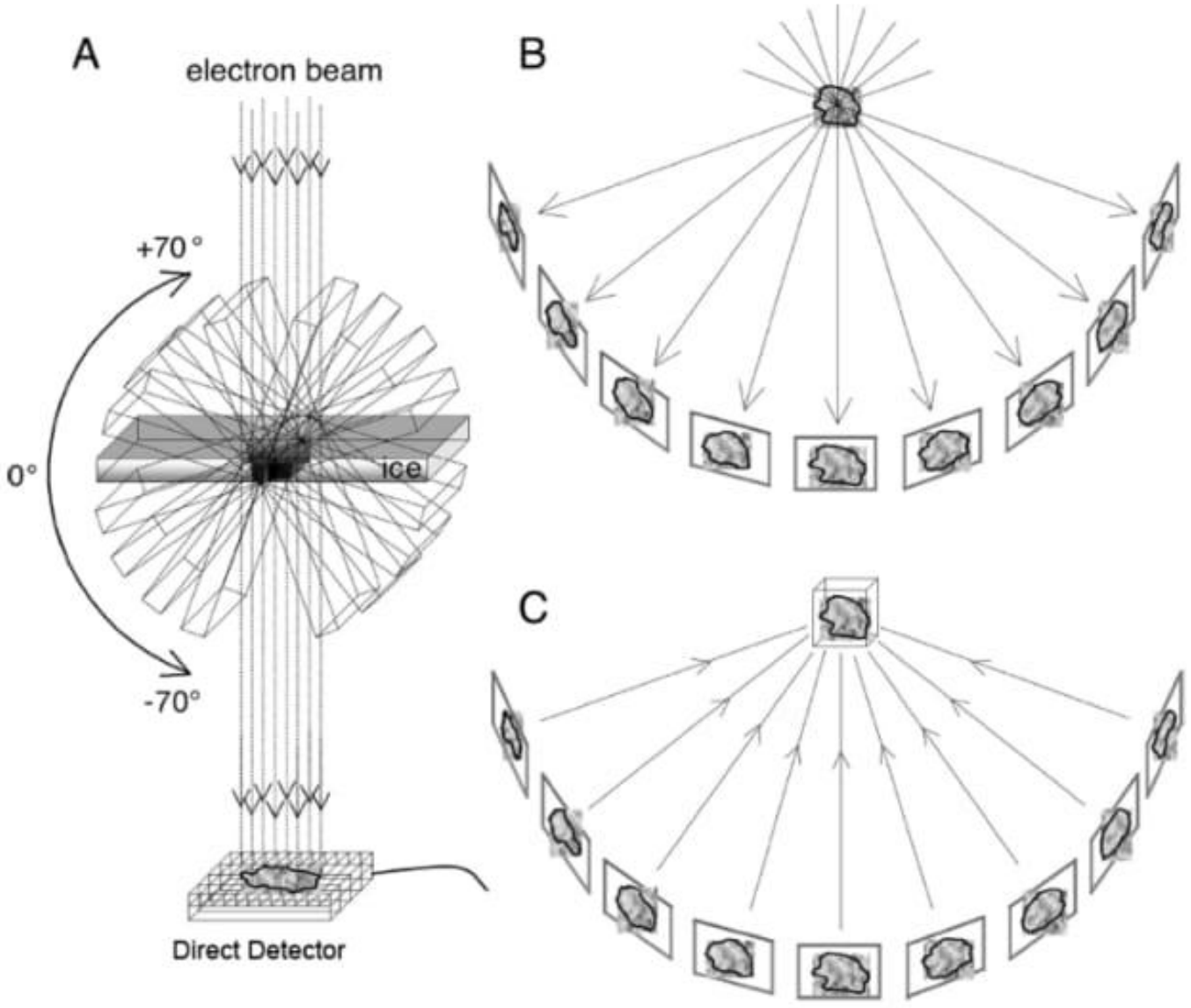
November 1, 2018 2:00-3:00 pm

Macromolecules



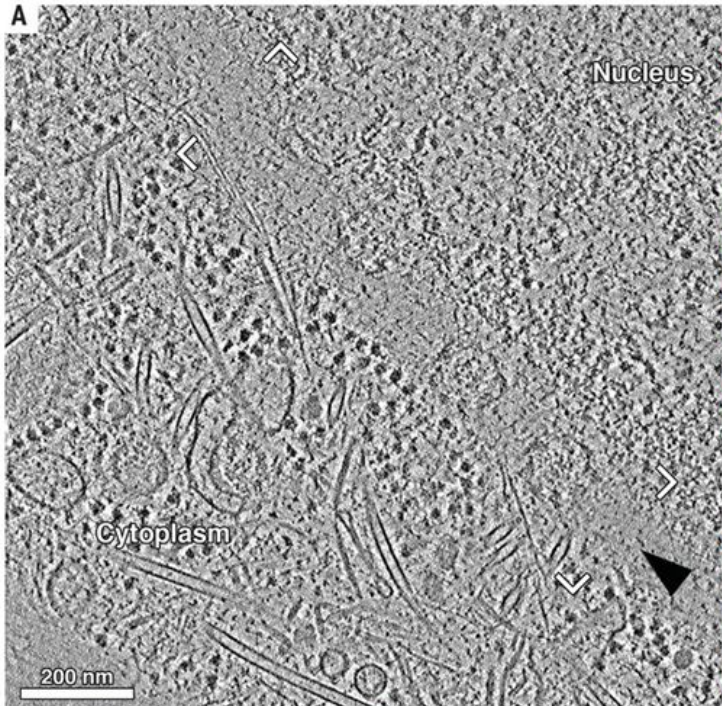
Hypothetical painting by David Goodsell

Tilt series and reconstruction

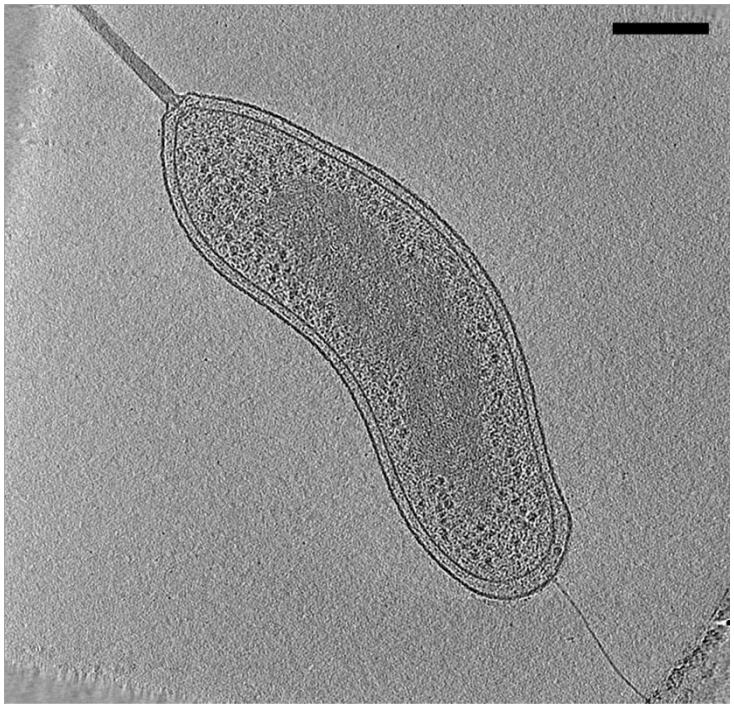


Steven and Belnap, *Current Protocols in Protein Science*, 2005

Example slices of 3D tomogram

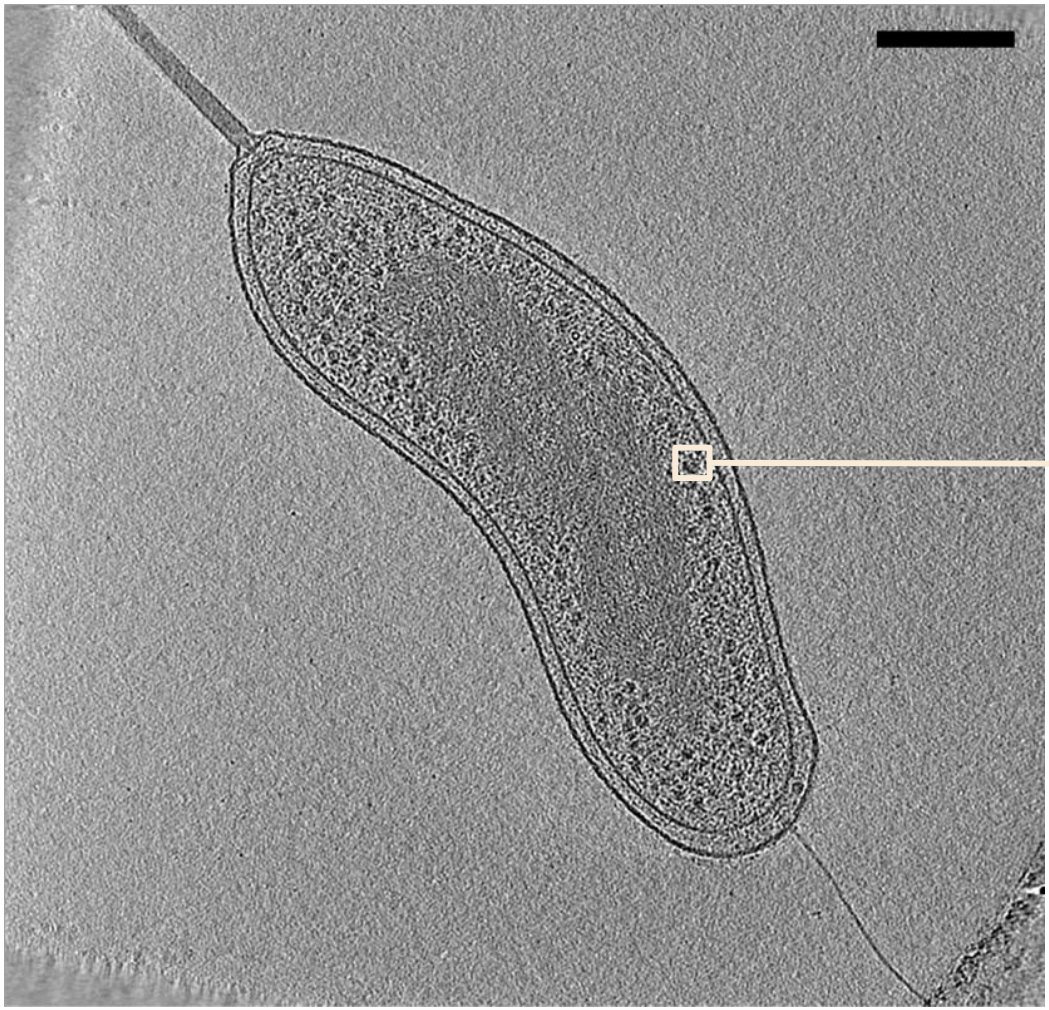


Mahamid et al, Science, 2016



Grant Jensen Lab at Caltech

Subtomogram extraction

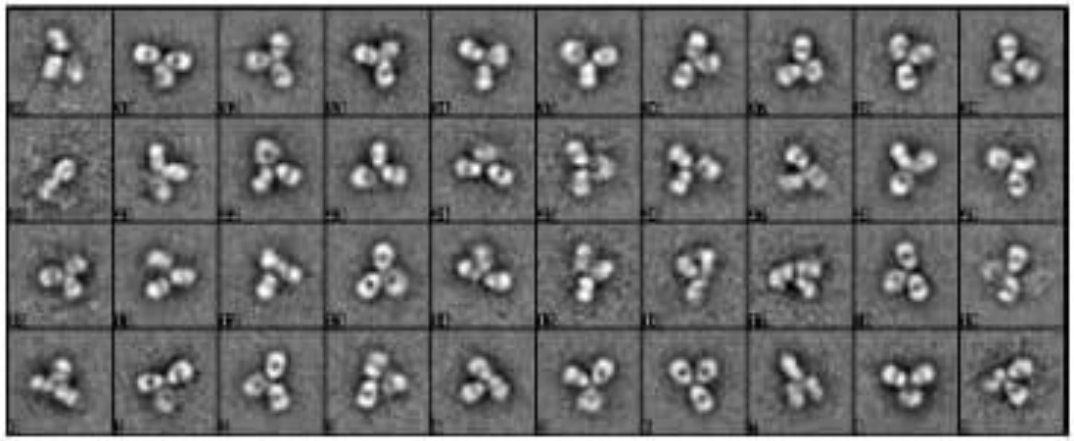


A subtomogram containing one macromolecule

Image from Grant Jensen Lab at Caltech

Challenge 1: Orientation

Particles can adopt all different orientations in 3D space

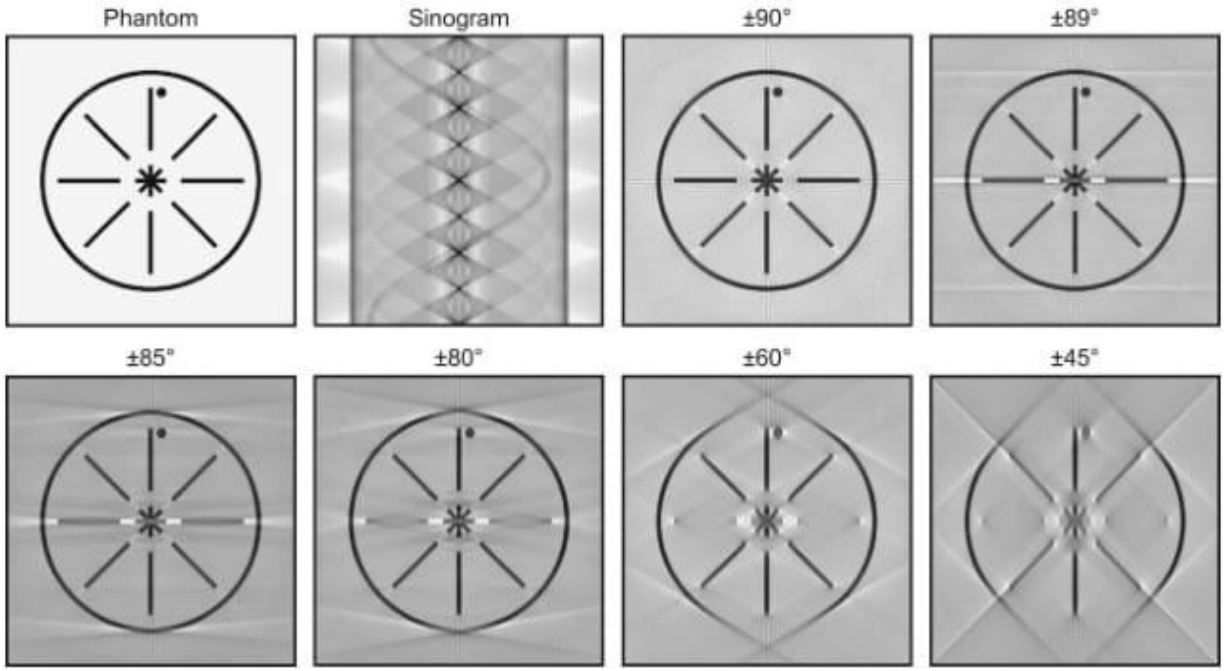


IgG antibodies

Image from Philippe Ringler lab at C-CINA

Challenges 2: Missing wedge effect

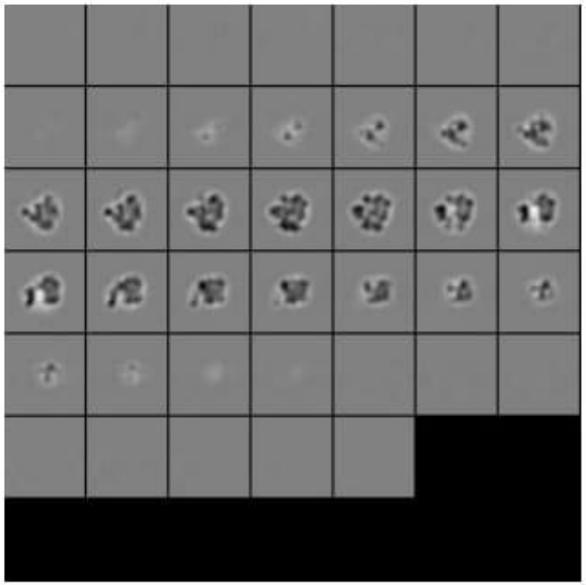
Reconstruction distortion due to limited range of tilt angles



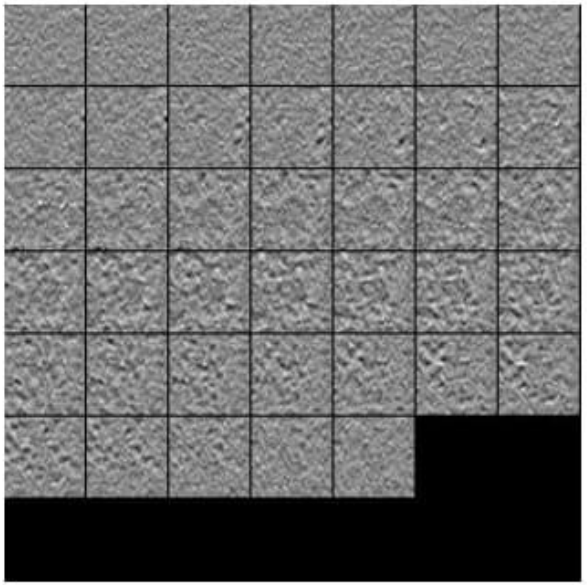
Palmer and lowe, *Ultramicroscopy*, 2014

Challenges 3: Noise

Subtomograms are very noisy



Mammalian ribosome ground truth (40^3)



Mammalian ribosome experimental data

Challenges 4: Molecular crowding

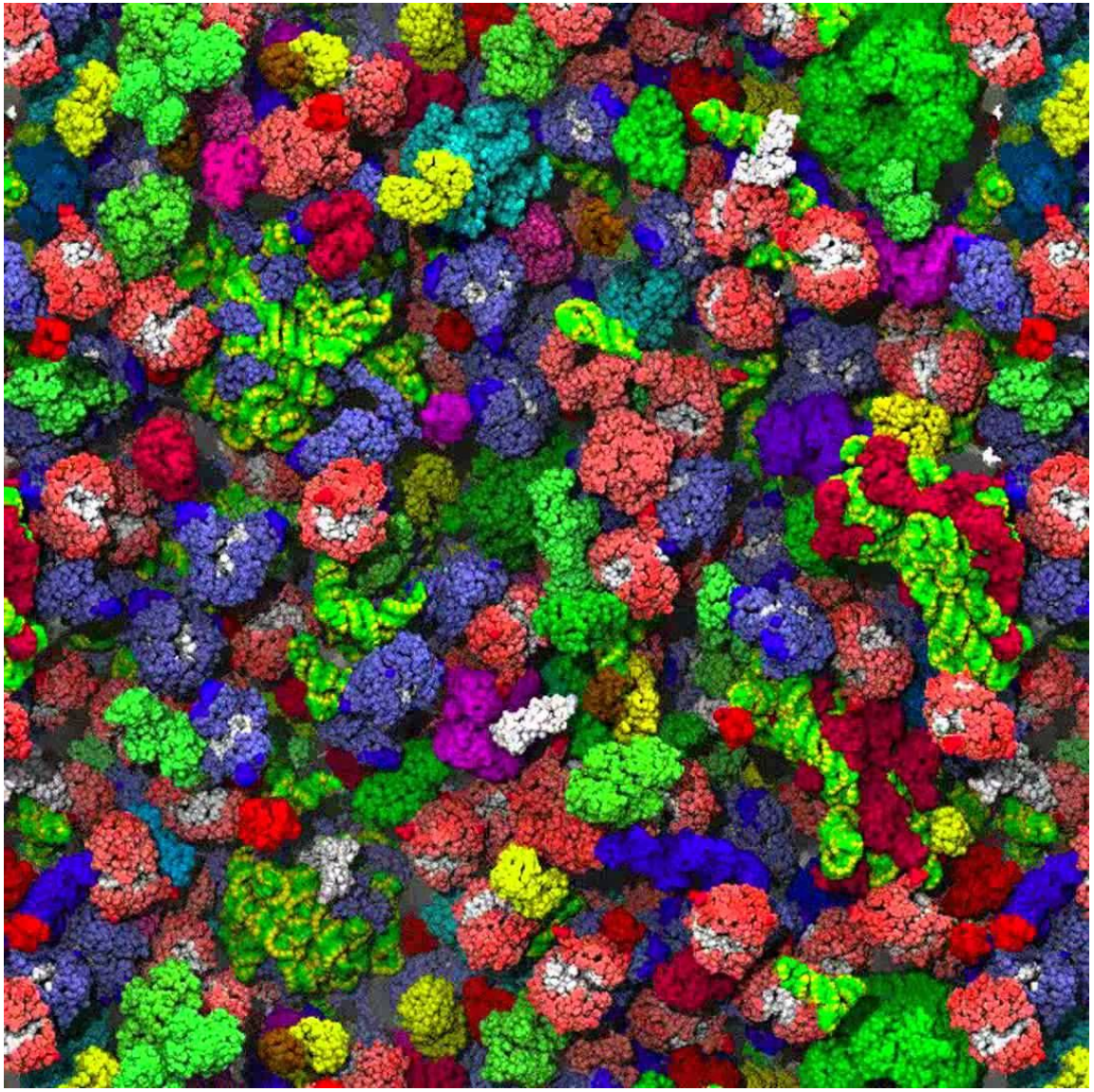
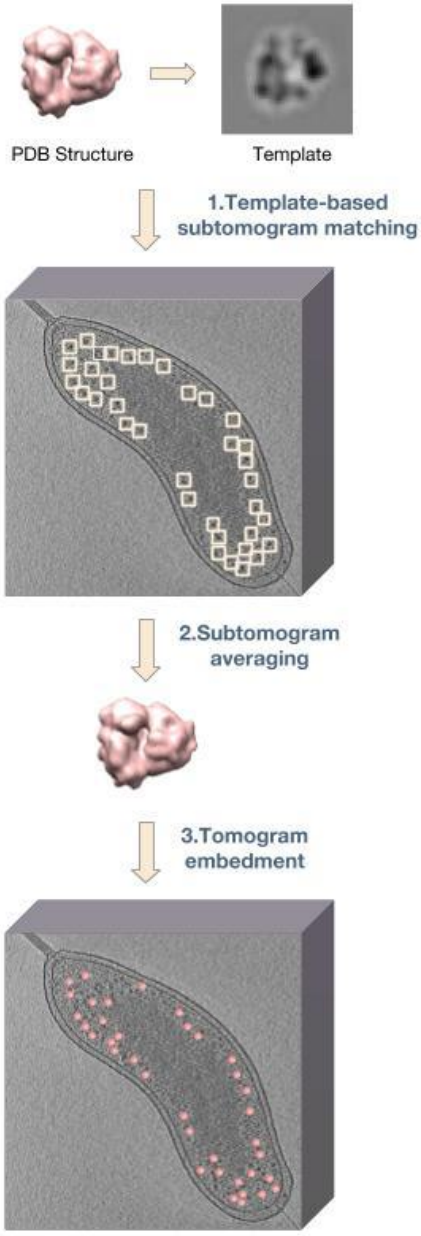


Image of simulated bacterial cytoplasm from McGuffee & Elcock, PLoS Comput Biol

Standard template-based structure recovery pipeline

Drawbacks

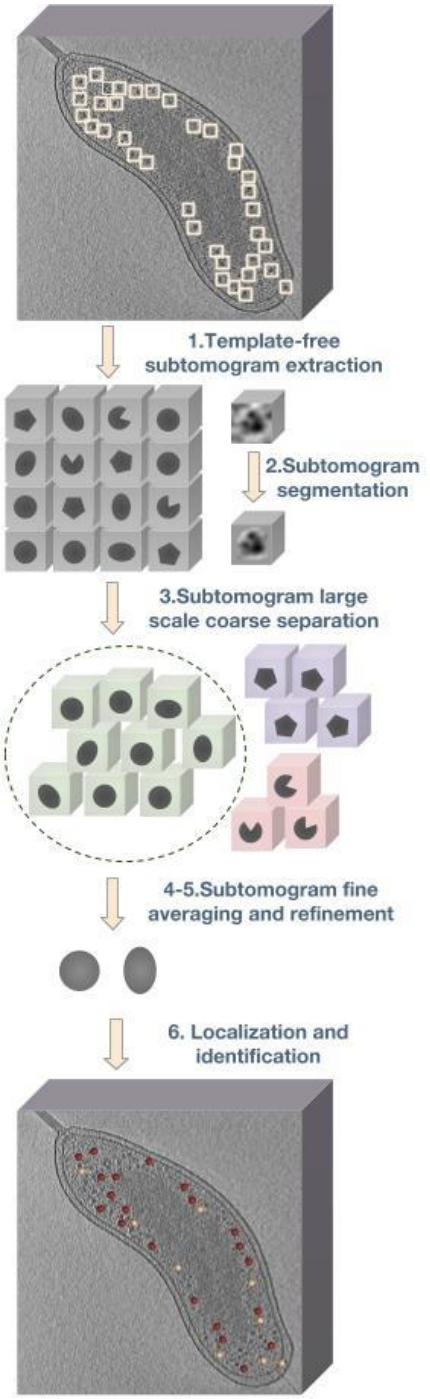
1. Rely on prior knowledge, unable to detect novel structures
2. Could be biased
3. Unsystematic for large datasets



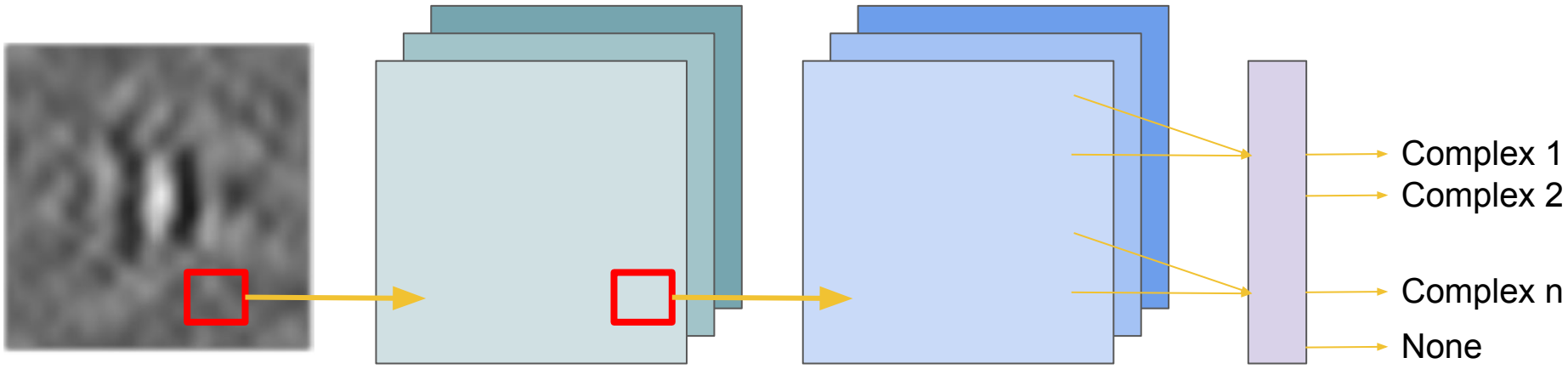
A template-free pipeline

Overview of techniques we developed

- Subtomogram segmentation
  - Semantic segmentation
- Coarse structural separation
  - Pose normalization
  - Deep structural feature extraction
  - Convolutional Autoencoder
  - CNN classifier
    - Deeper models
    - Model compression
- Averaging and classification
  - Fast alignment + maximum likelihood averaging
- Others
  - Saliency detection
  - Generative models of pseudo molecular structures
  - Multi-task learning



CNN based subtomogram classification

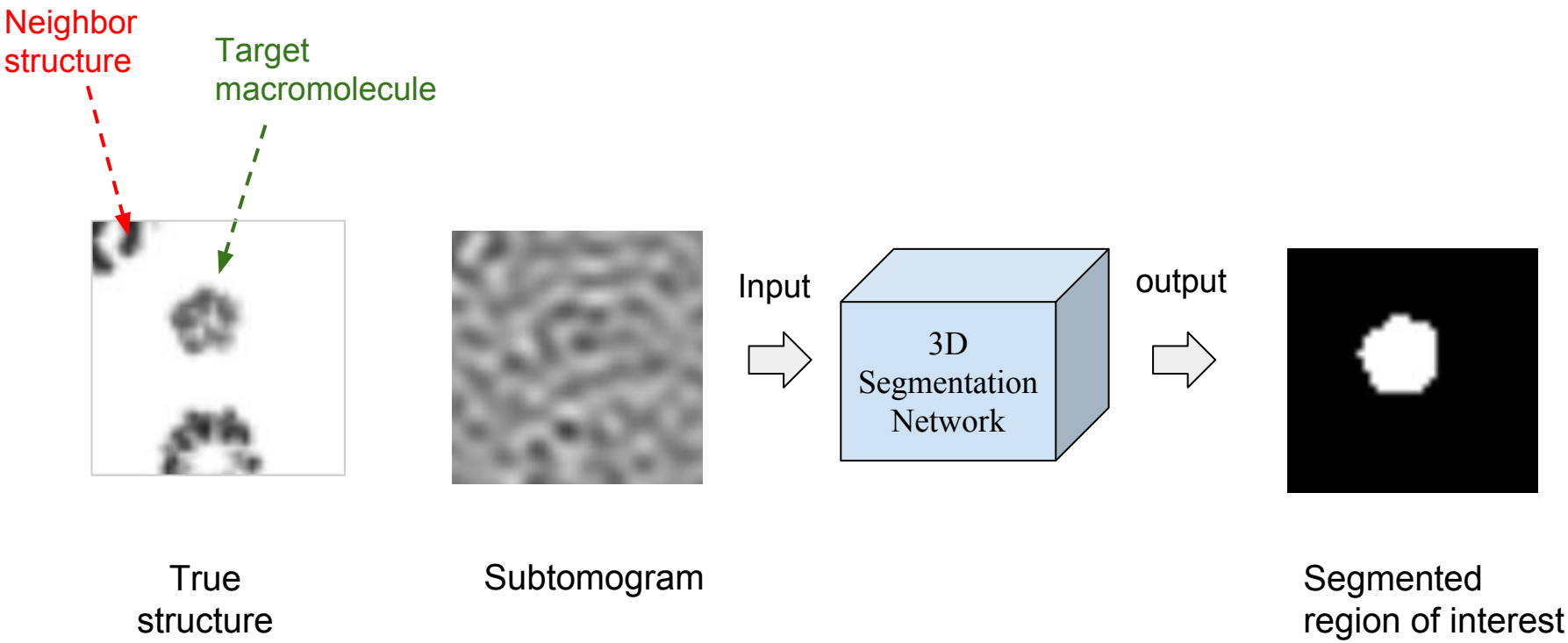


Input subtomogram

Output classes

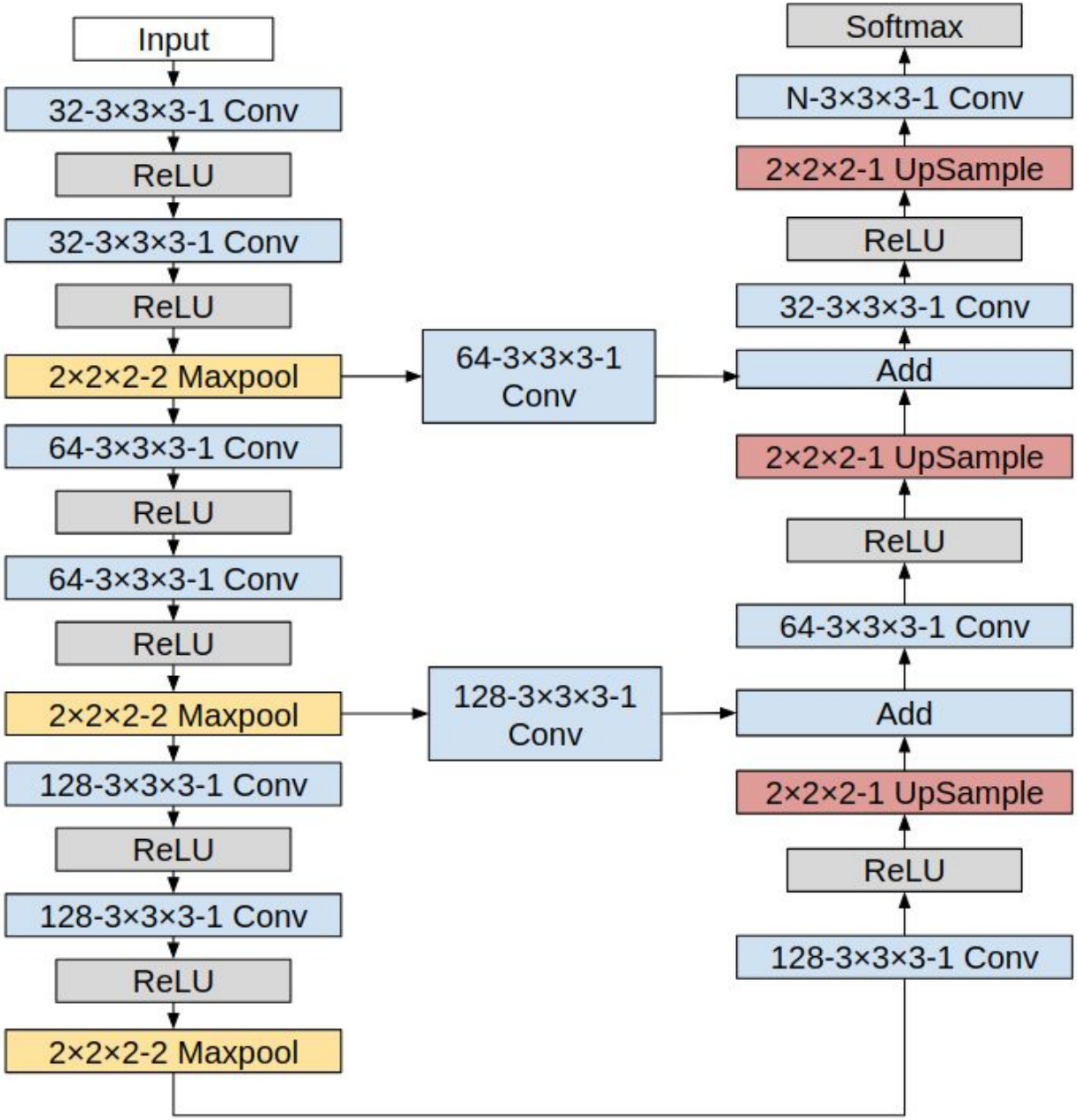
Xu et al. 2017, Che et al. 2018, Guo et al. 2018

Subtomogram segmentation

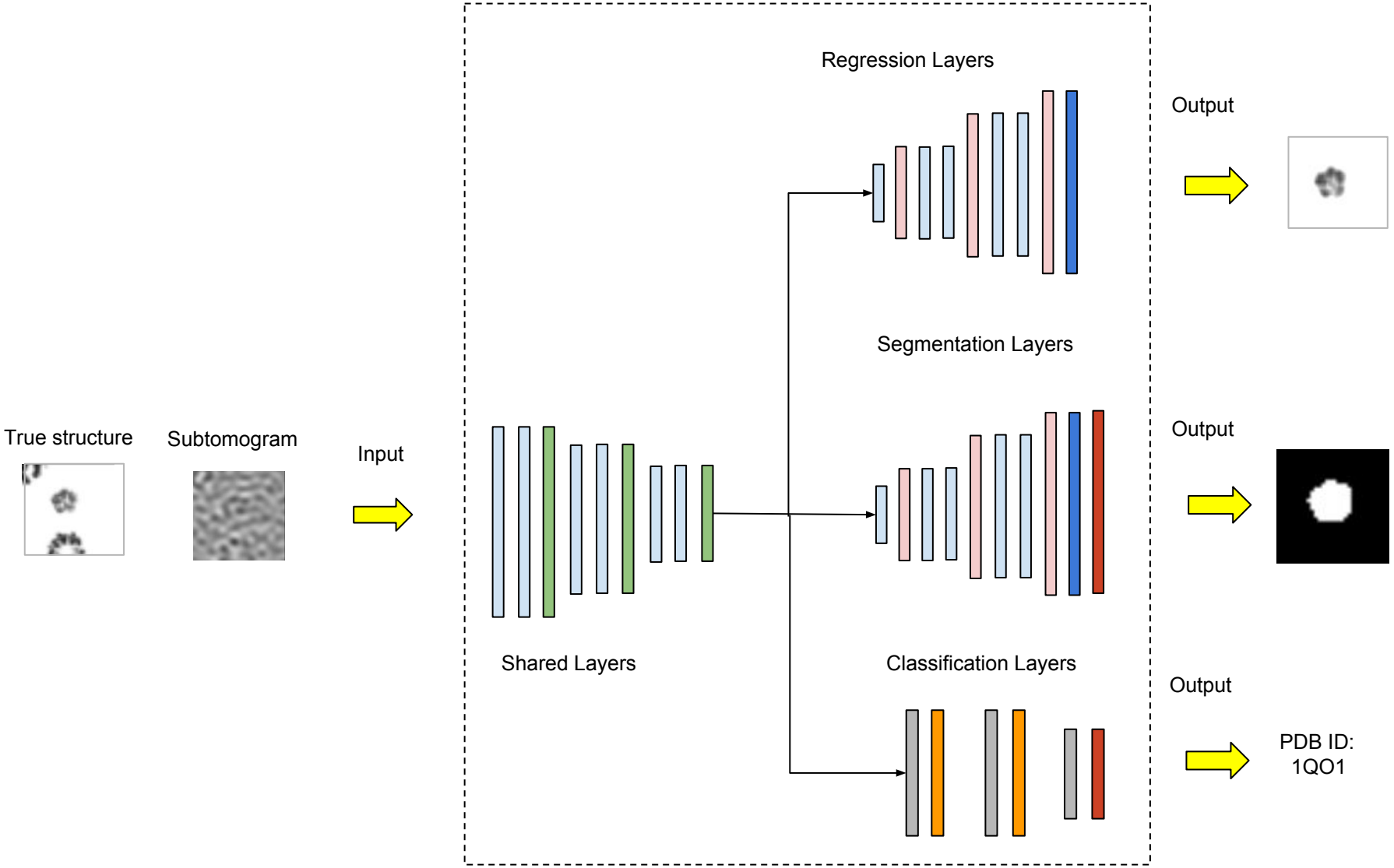


Liu et al. ICIP 2018

Segmentation network model

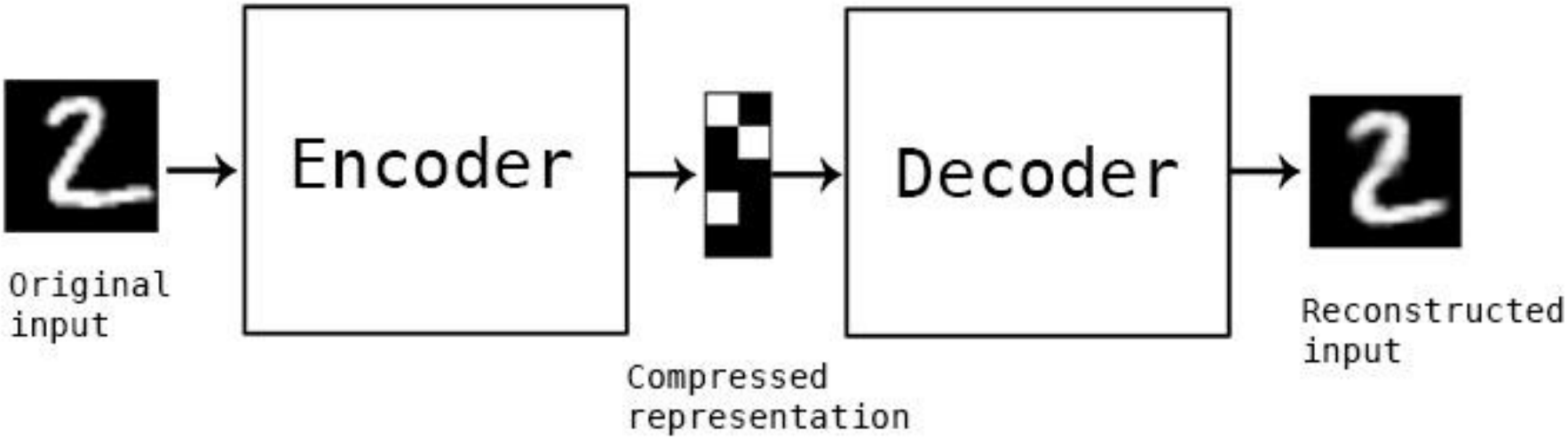


Multi-task learning model

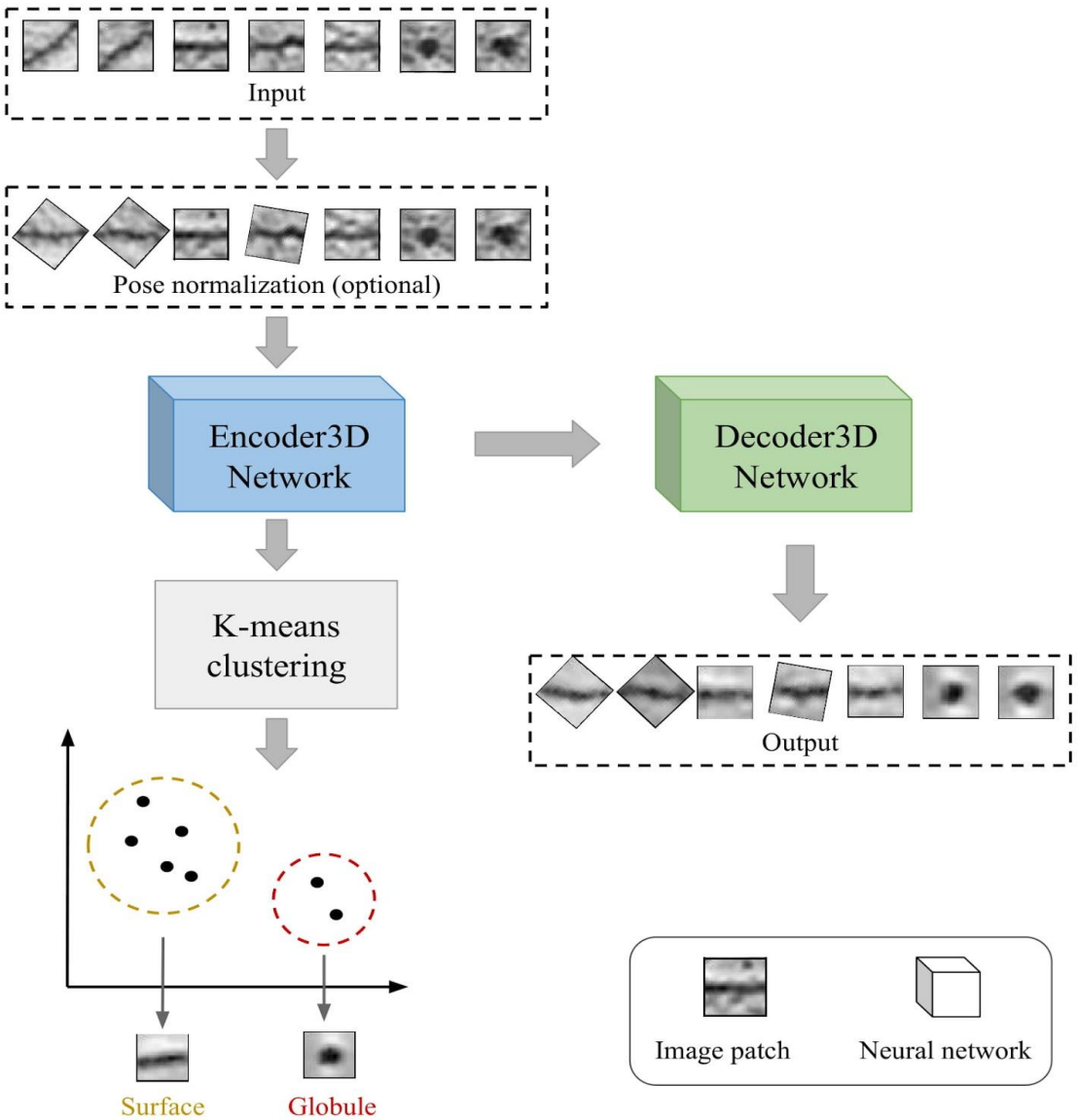


Liu et al. BMVC 2018

Autoencoder based pattern compression

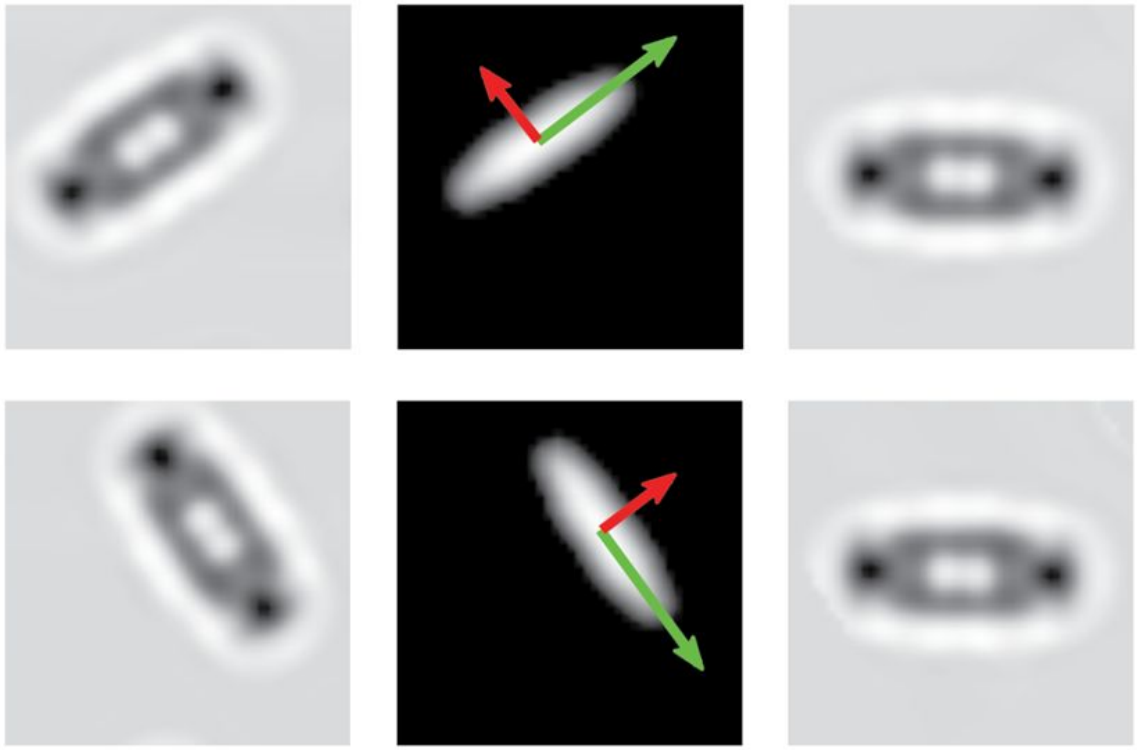


Autoencoder : large scale subtomogram separation



Zeng et al. Journal of Structural Biology

Pose normalization



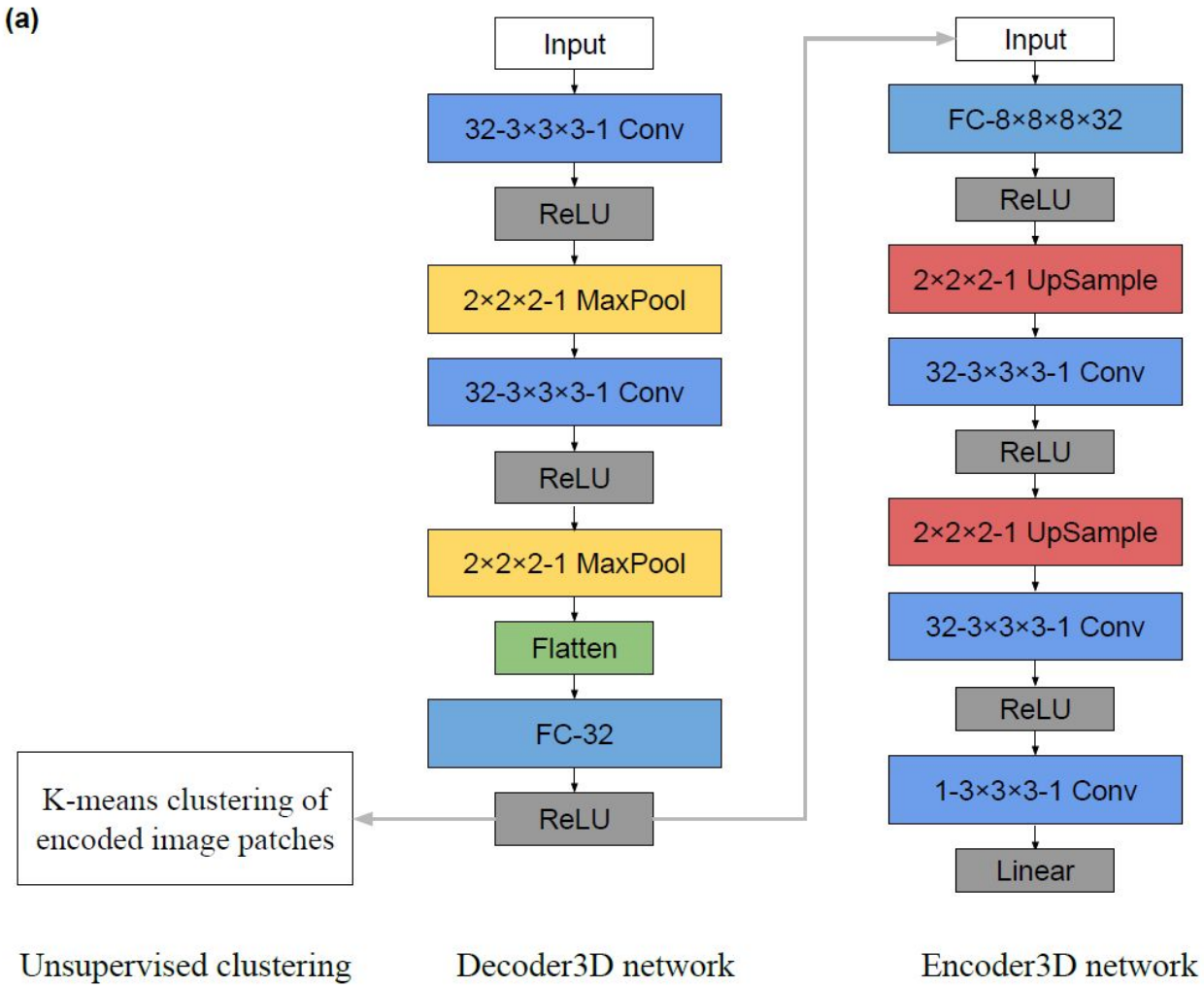
Standard PCA

$$\mathbf{W} = \sum_i \mathbf{X}_i \mathbf{X}_i^\top$$

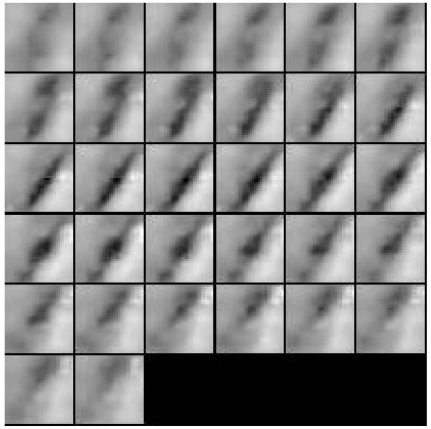
PCA on weighted covariance matrix

$$\mathbf{W} = \int_{\mathbf{x}} \phi(\mathbf{x})^2 \, \mathbf{x}\mathbf{x}^\top$$

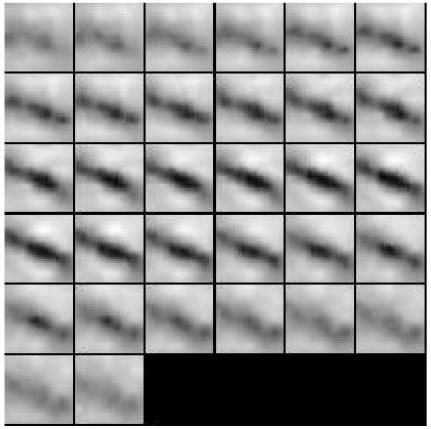
Autoencoder model network



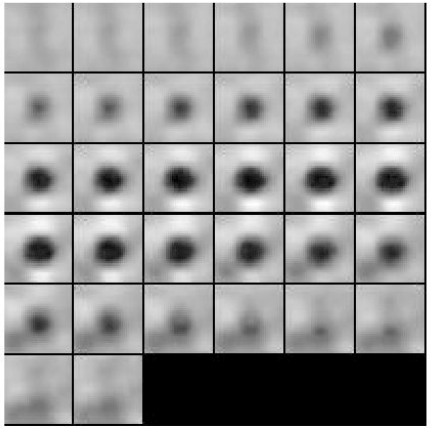
Pattern detected



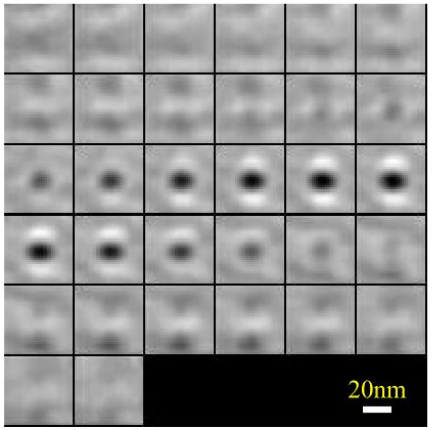
Surface patch



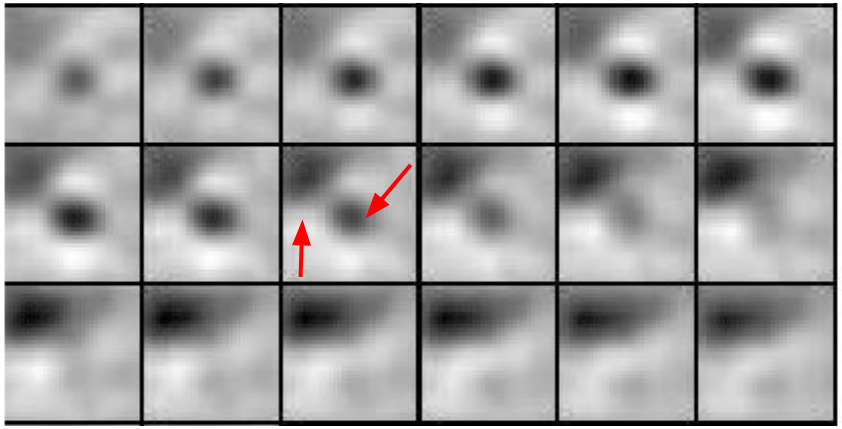
Surface patch



Large globule

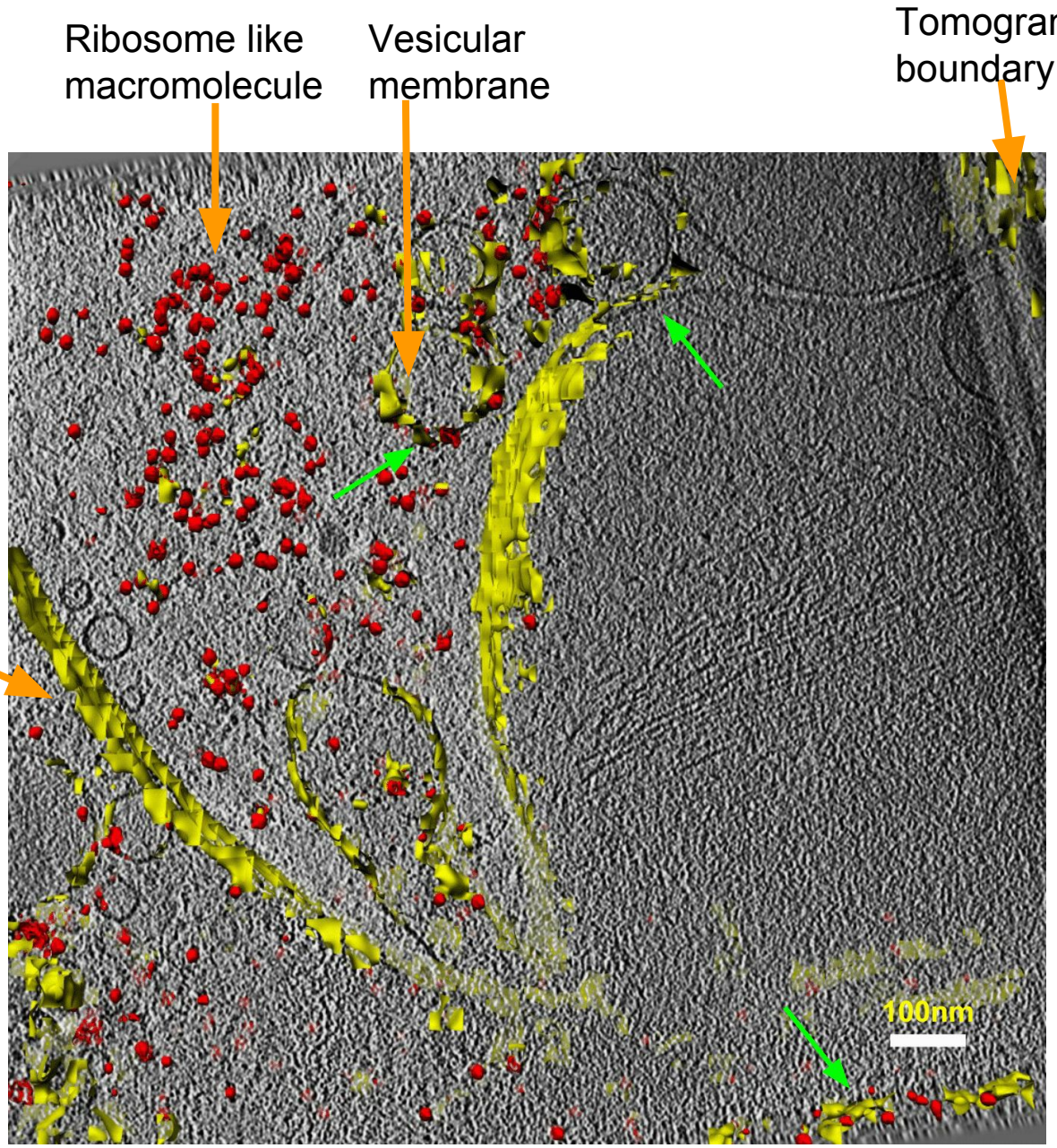


Small globule

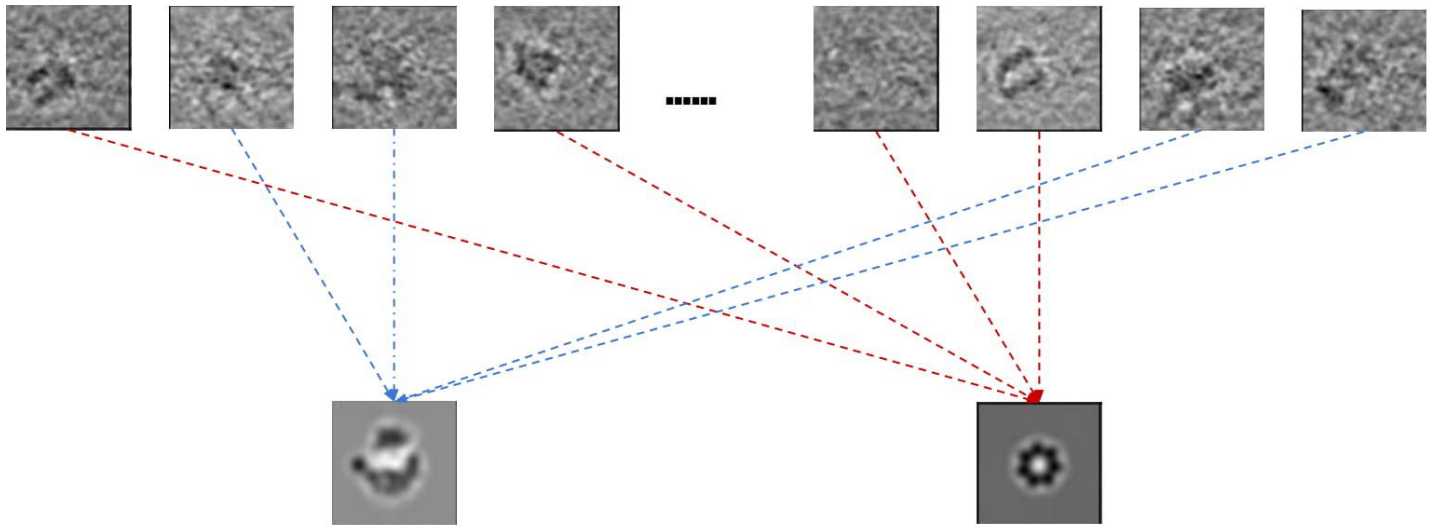


Interaction between cellular components

Pattern embedded



Subtomogram averaging



N subtomograms are transformed (rotated and translated), clustered, and averaged into K different classes.

Existing method 1 : Maximum likelihood

- Optimizes the probability of observing the data given a data model

$$X_i = R_{\phi_i} A_{\kappa_i} + G_i \quad \forall i = 1, \ldots, N$$

- Uses the EM algorithm
- Exhaustive scanning in 6D parametric space of rigid transformation is needed
  - In principle, computationally infeasible

Existing method 2 : Fast alignment

- Optimizes a correlation score between a subtomogram X and an average A:

$$c(\phi^{\text{ro}}, \phi^{\text{tr}}) = \frac{\sum_j w_j^2 X_j \ \exp(2\pi i \xi_j^\top \phi^{\text{tr}}) \ \overline{(R_{\phi^{\text{ro}}} A)_j}}{\sqrt{\sum_j \ w_j^2 \ [R_{\phi^{\text{ro}}}(A \circ A)]_j}}$$

- Uses the fast rotational matching algorithm
- Searches sub-optimal rigid transformations using local maximum under constraints
- Not sufficiently robust to low SNR and missing wedge effects

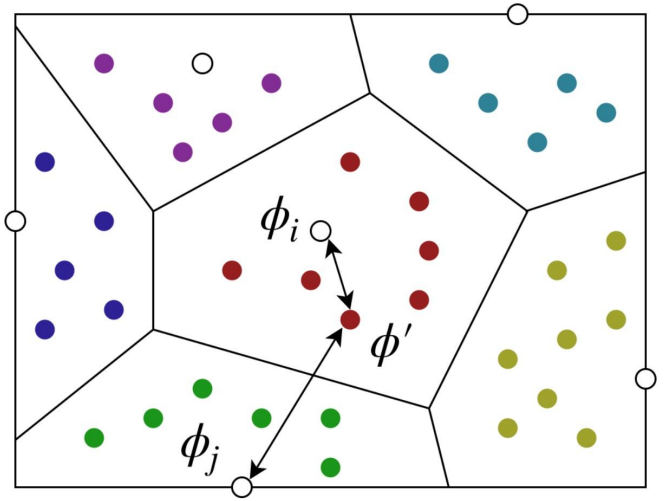
FAML: Integration of two methods

Approximate the integral with sub-optimal transformations:

$$\int_{\phi} f(\phi, X, A)\, d\phi \approx \sum_{\phi \in \oplus} f(\phi, X, A)\, \tilde{v}(\phi, \oplus)$$

Where $\tilde{v}(\phi, \oplus) := \frac{|v(\phi, \oplus)|}{\sum_{\phi' \in \oplus} |v(\phi', \oplus)|}$ is the normalized hypervolume of Ø.

Voronoi weights are approximated using Monte-Carlo sampling



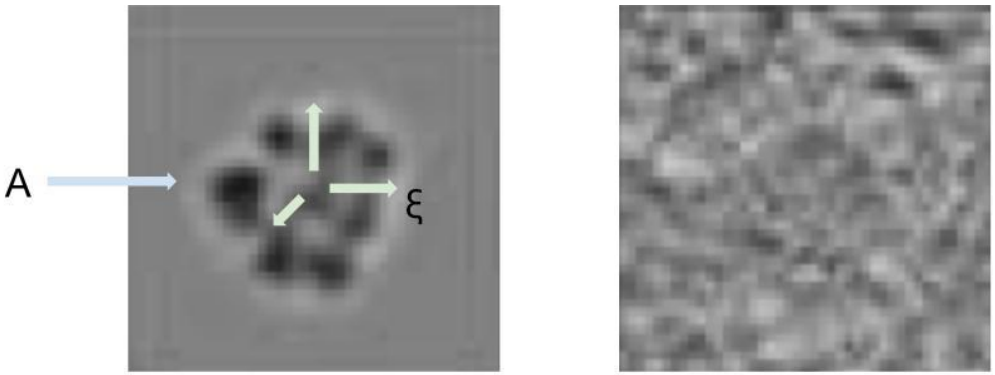
Zhao et al. Bioinformatics

FAML: Algorithm

Initialize model parameters $\Theta = (A, \alpha, \sigma, \xi)$ from the distribution of X

For M iterations:

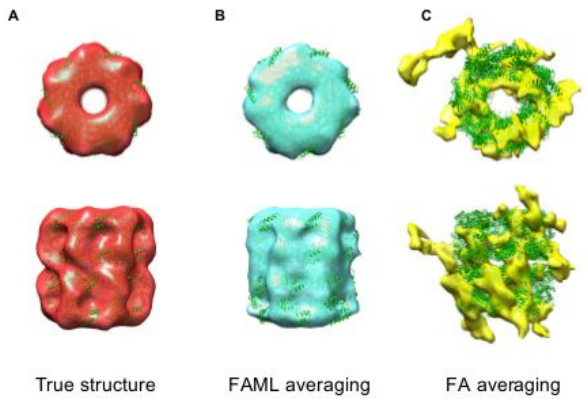
- Compute a list of sub-optimal rigid transformations using FA
- Compute their Voronoi weights
- Update α, the proportion of particles belonging to different classes
- Update σ, the standard deviation of Gaussian noise
- Update ξ, the standard deviation of translation parameters (3D Gaussian)
- Update A, the underlying true structure (subtomogram average)



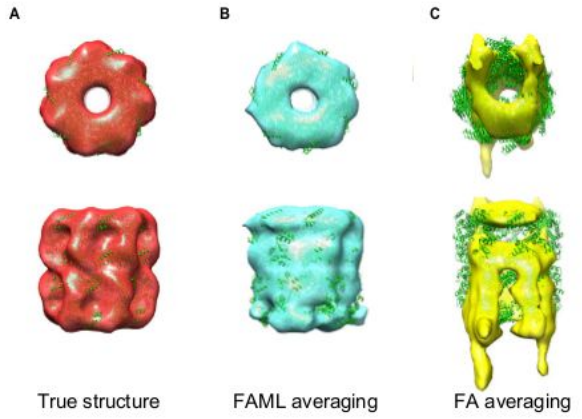
Proportion of A: α

Noise: σ

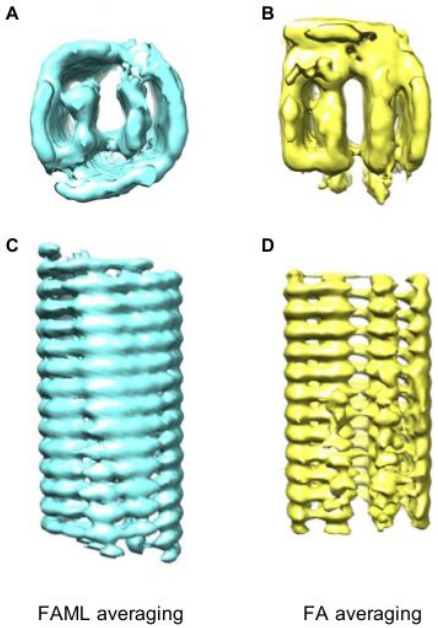
Results



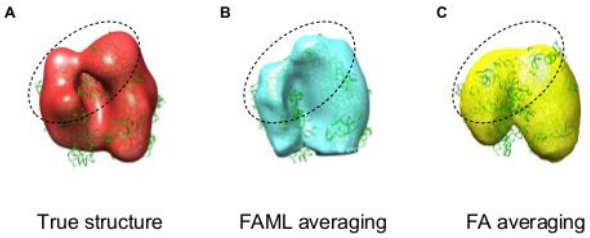
Low SNR



Orientation preferred



TMV (Kunz et al 2015)



Ribosome (Guo et al 2018)

Results

We randomly selected 400 GroEL/GroES subtomograms from an original dataset of 780.

FAML:

GroEL: $r = 0.87$

GroEL-GroES: $r = 0.78$

FA:

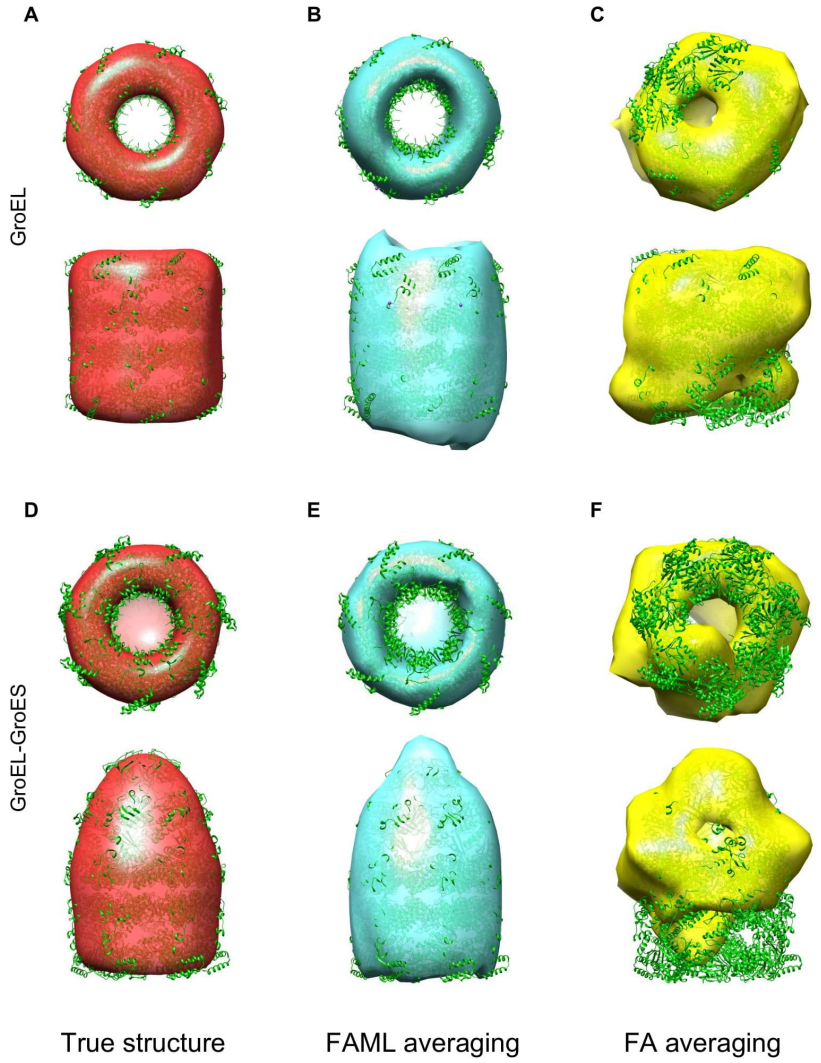
GroEL: $r = 0.40$

GroEL-GroES: $r = 0.24$

ML (on original dataset):

GroEL: $r = 0.88$

GroEL-GroES: $r = 0.81$



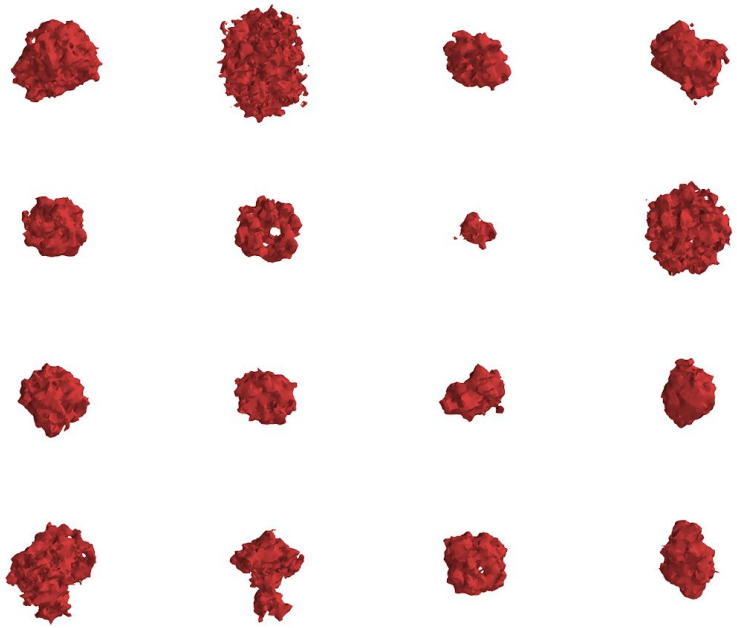
Computing time

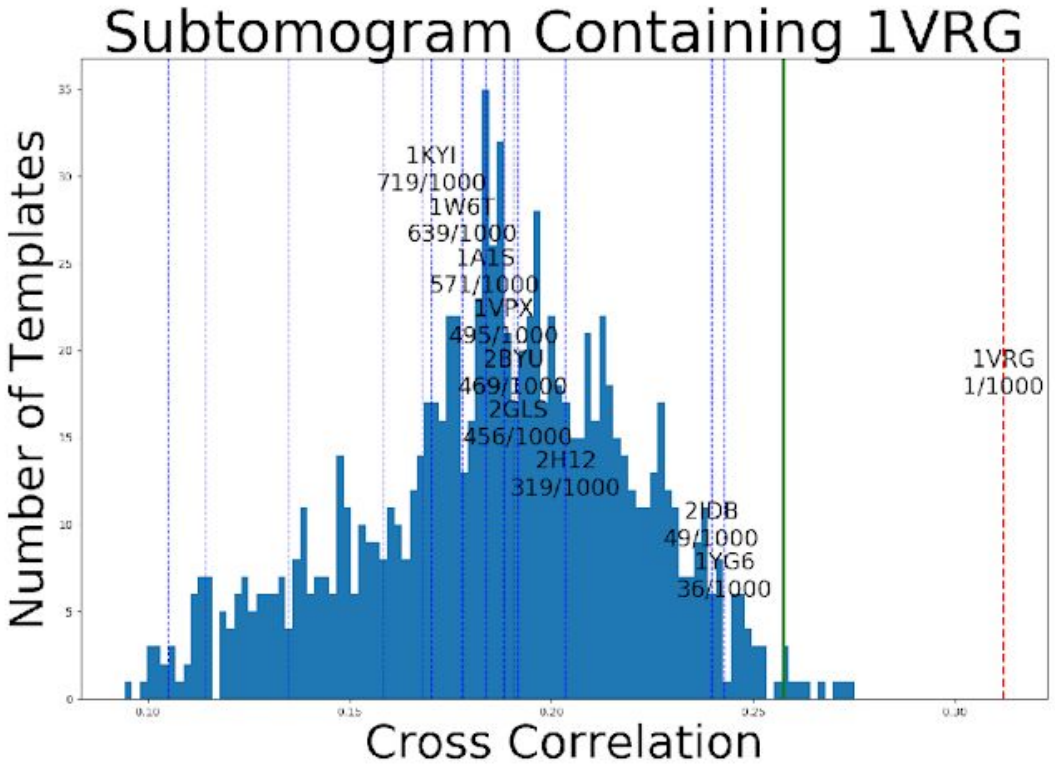
Table 1: Computing time used for FA, FAML and RELION methods. 32^3 in parenthesis denotes the testing subtomograms are of size 32^3.

	Iterations to converge	Mean time per iteration	Total time
FA (32^3)	11	7 s	78 s
FAML (32^3)	10	56 s	562 s
RELION (32^3)	8	340 s	2720 s
FA (64^3)	4	27 s	106 s
FAML (64^3)	3	150 s	451 s
RELION (64^3)	6	340 s	2041 s
FA (128^3)	5	143 s	717 s
FAML (128^3)	4	449 s	1794 s
RELION (128^3)	3	921 s	2764 s

- FAML is slower than FA but takes less iterations to converge
- FAML is 2 to 5 times faster than Relion using their default parameters

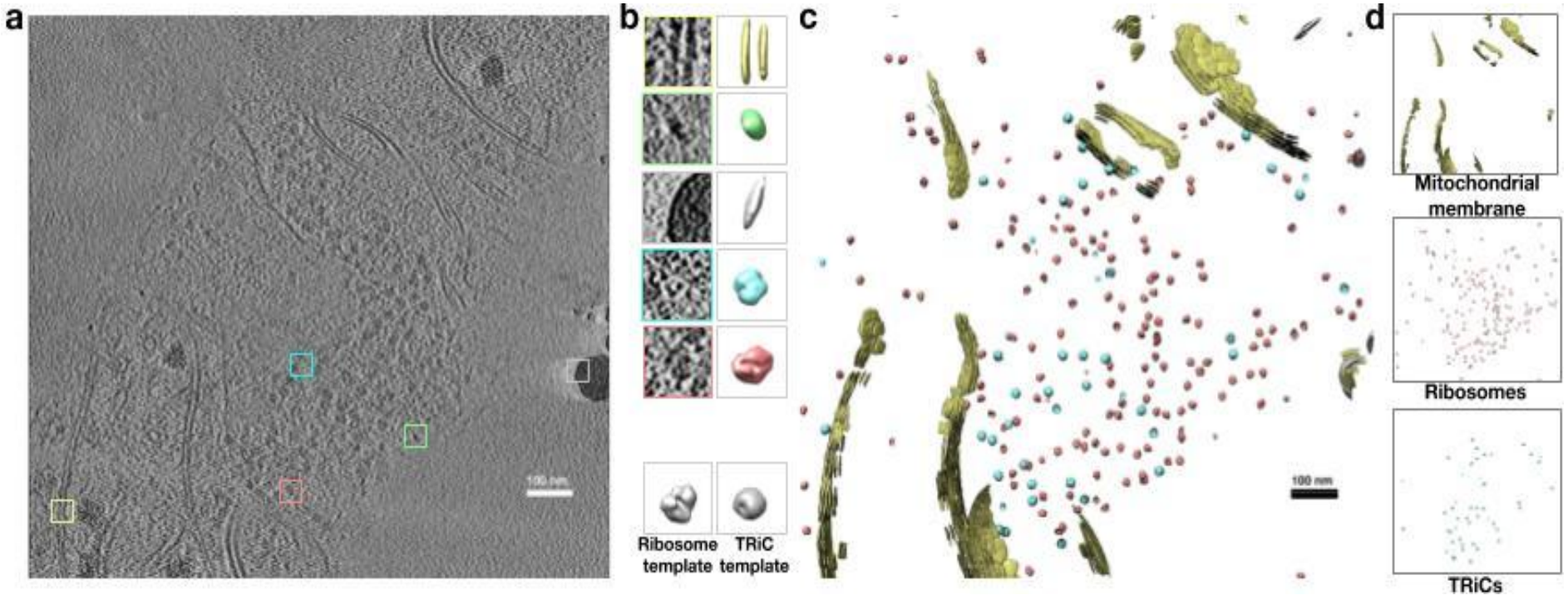
Hypothesis testing for structure identification



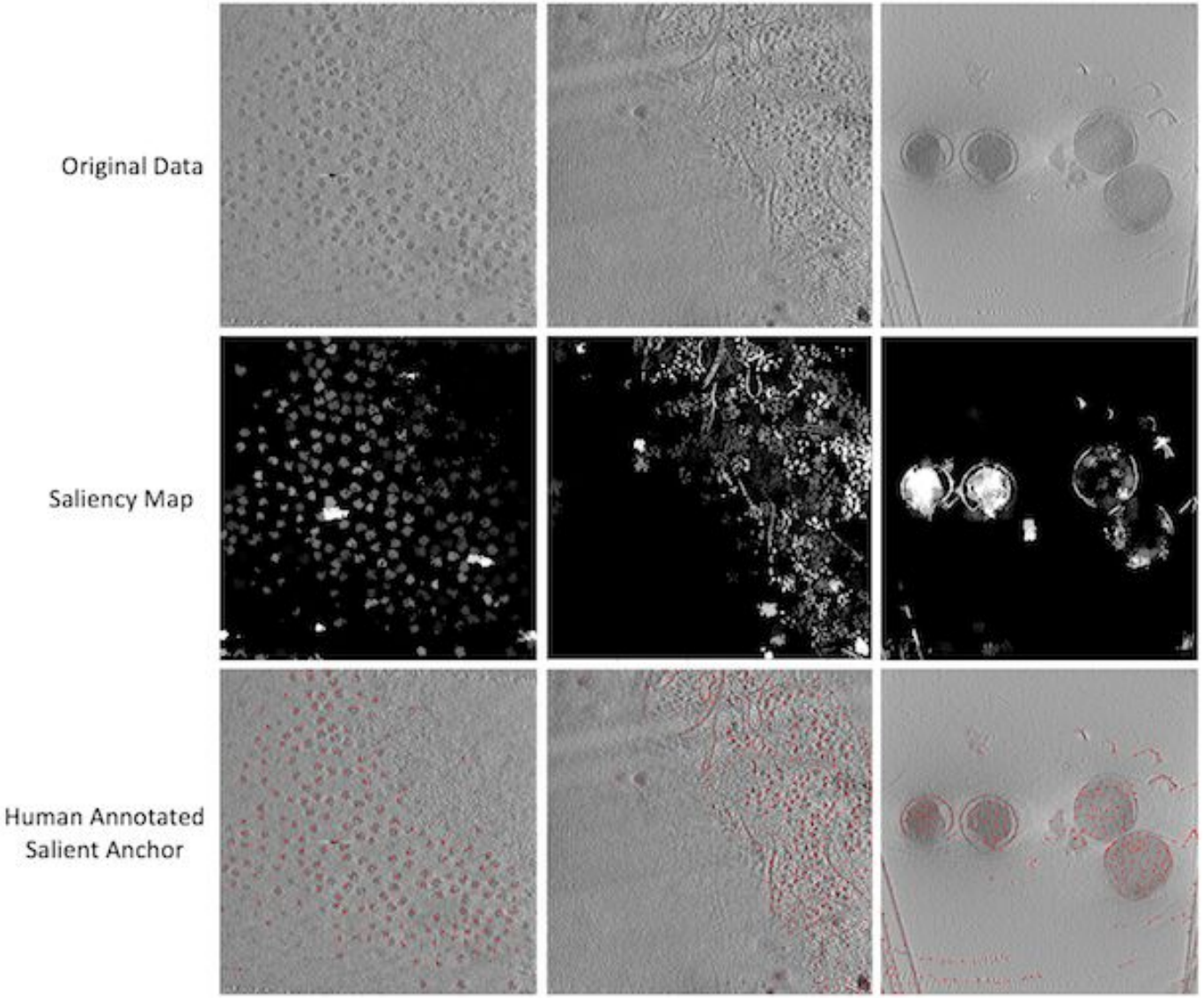


Wang et al. BMVC 2018

Pipeline results on a rat neuron tomogram



Future works



Zhou et al. arXiv

Future works

1. Saliency based particle picking method
2. Better tomogram reconstruction algorithm
3. Fast systematic structure identification method
4. Models for recovered structures
5. Spatial statistical models for macromolecule distribution

Acknowledgement

Funding support:

NIH P41 GM103712

Samuel and Emma Winters Foundation

People

Dr. Min Xu

Dr. Robert Murphy

Dr. Zachary Freyberg

Dr. Tzviya Zeev-Ben-Mordehai

References

Zeng, X., Leung, M.R., Zeev-Ben-Mordehai, T. and Xu, M., 2018. A convolutional autoencoder approach for mining features in cellular electron cryo-tomograms and weakly supervised coarse segmentation. *Journal of structural biology*, *202*(2), pp.150-160.

Liu, C., **Zeng, X**., Lin, R., Liang, X., Freyberg, Z., Xing, E. and Xu, M., 2018. Deep learning based supervised semantic segmentation of Electron Cryo-Subtomograms. *ICIP 2018*.

Wang, K.W., **Zeng, X.**, Liang, X., Huo, Z., Xing, E.P. and Xu, M., 2018. Image-derived generative modeling of pseudo-macromolecular structures-towards the statistical assessment of Electron CryoTomography template matching. *BMVC 2018*.

Zhao, Y., **Zeng, X.**, Guo, Q. and Xu, M., 2018. An integration of fast alignment and maximum-likelihood methods for electron subtomogram averaging and classification. *Bioinformatics 34(13):i227–i236, 2018*.

Zhou, B., Guo, Q., **Zeng, X.** and Xu, M., 2018. Feature Decomposition Based Saliency Detection in Electron Cryo-Tomograms. *arXiv preprint arXiv:1801.10562*.

Guo, J., Zhou, B., **Zeng, X.**, Freyberg, Z. and Xu, M., 2018, June. Model compression for faster structural separation of macromolecules captured by Cellular Electron Cryo-Tomography. In International Conference Image Analysis and Recognition (pp. 144-152). Springer, Cham.

Liu, C., **Zeng, X.**, Wang, K., Guo, Q. and Xu, M., 2018. Multi-task Learning for Macromolecule Classification, Segmentation and Coarse Structural Recovery in Cryo-Tomography. BMVC 2018.

Che, C., Lin, R., **Zeng, X.**, Elmaaroufi, K., Galeotti, J. and Xu, M., 2017. Improved deep learning based macromolecules structure classification from electron cryo tomograms. *Journal of Machine Vision and Application*.

Xu, M., Beck, M. and Alber, F., 2012. High-throughput subtomogram alignment and classification by Fourier space constrained fast volumetric matching. *Journal of structural biology*, *178*(2), pp.152-164.

Scheres, S.H., Melero, R., Valle, M. and Carazo, J.M., 2009. Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization. *Structure*, *17*(12), pp.1563-1572.