# Automatic detection of mitochondria and cross-domain macromolecule structure classification in cellular electron cryo-tomograms

Xu Lab
Computational Biology Department
Carnegie Mellon University

# Overview

**Cross-domain macromolecule classification in cellular tomograms**

Ruogu Lin*, Xiangrui Zeng*, et al, *ISMB* 2019

1st and 2nd year CPCB PhD students.

**Automatic detection of mitochondria in cellular tomograms**

Ran Li*, Xiangrui Zeng*, et al, *BMC Bioinformatics*

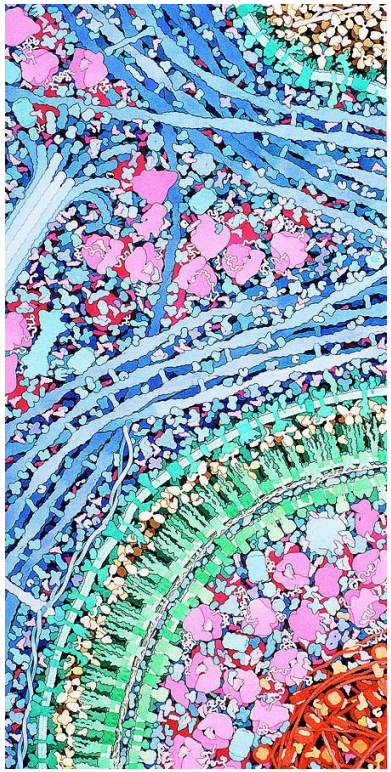
In collaboration with Freyberg lab at Pitt and Jiang lab at THU

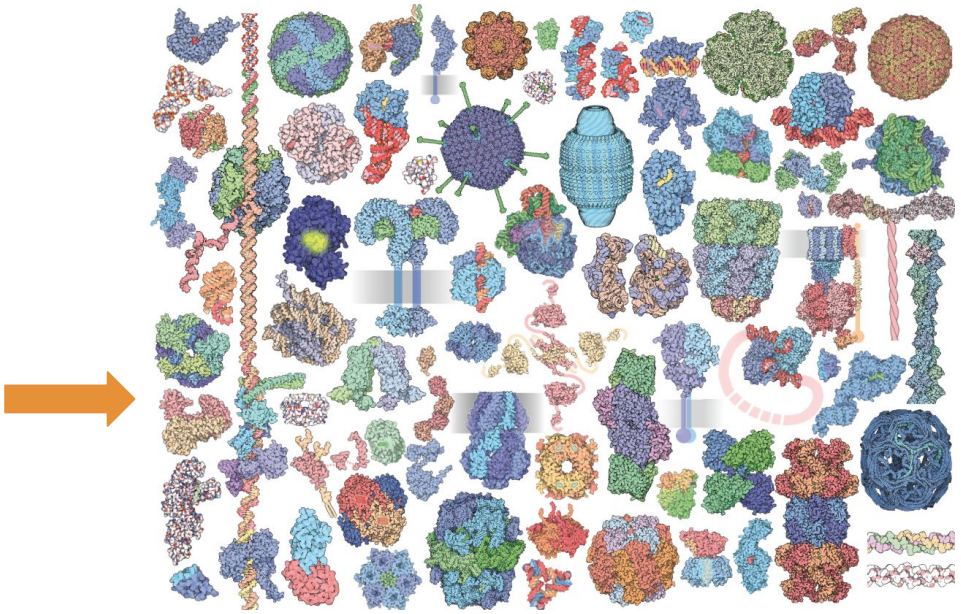
**Funding support:** NIH P41 GM103712

# Adversarial domain adaptation for cross data source macromolecule in situ structural classification in cellular electron cryo-tomograms

R Lin*, X Zeng*, K Kitani, M Xu. *ISMB* 2019

# Macromolecule



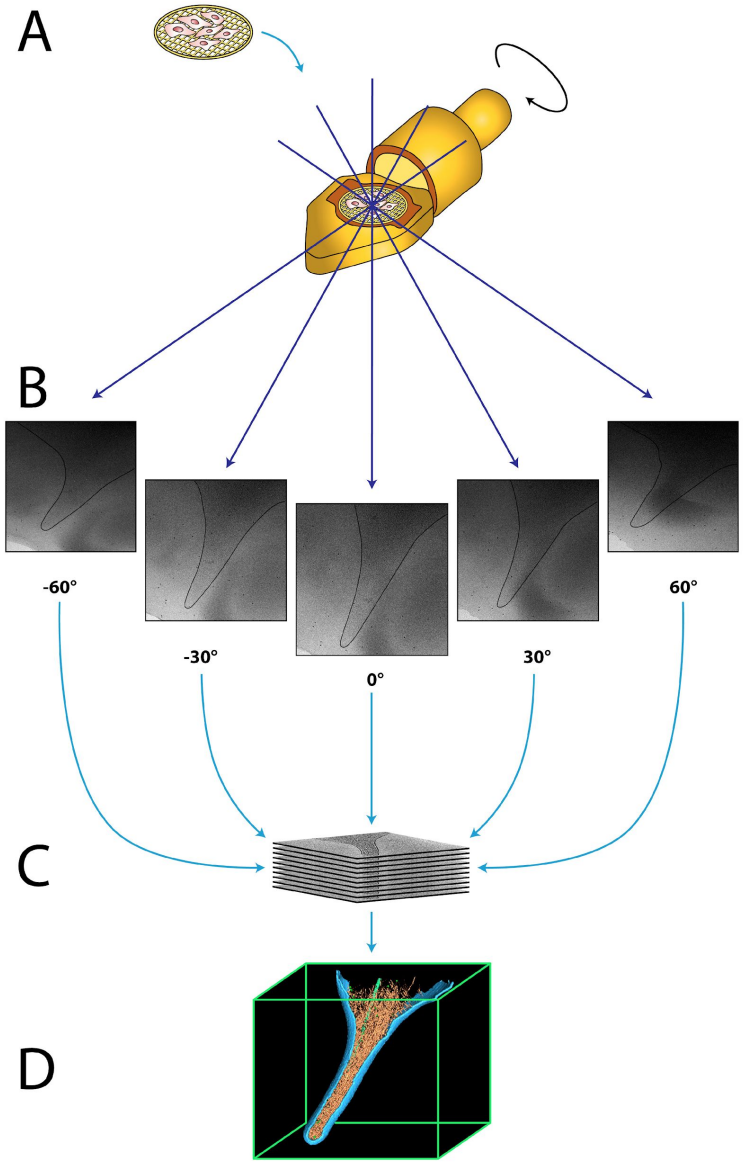


**Macromolecules**

**Cell cytoplasm**

Hypothetical Painting by David Goodsell

# Cryo-electron tomography




A: Sample preparation

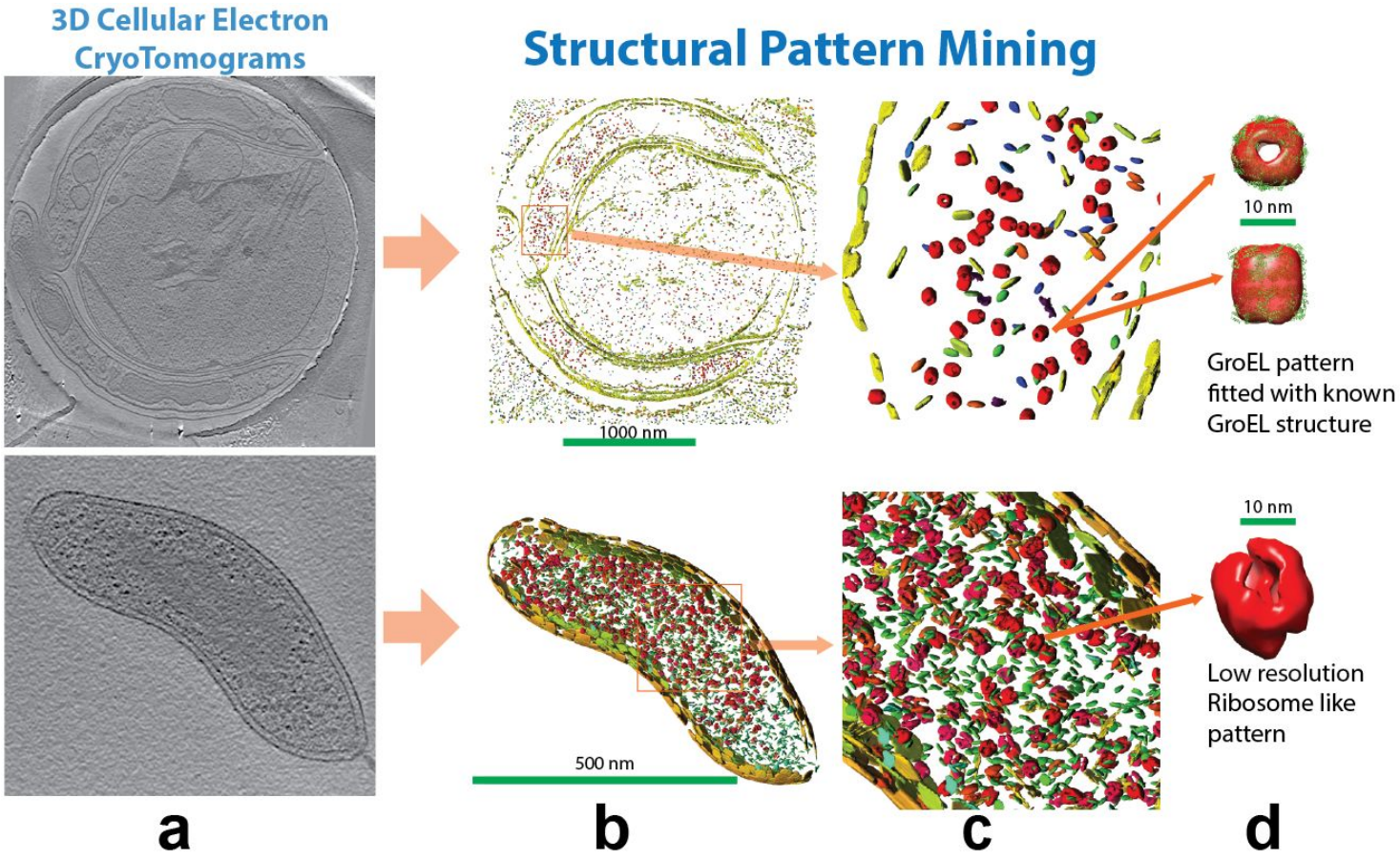
B: Imaging through tilt-series

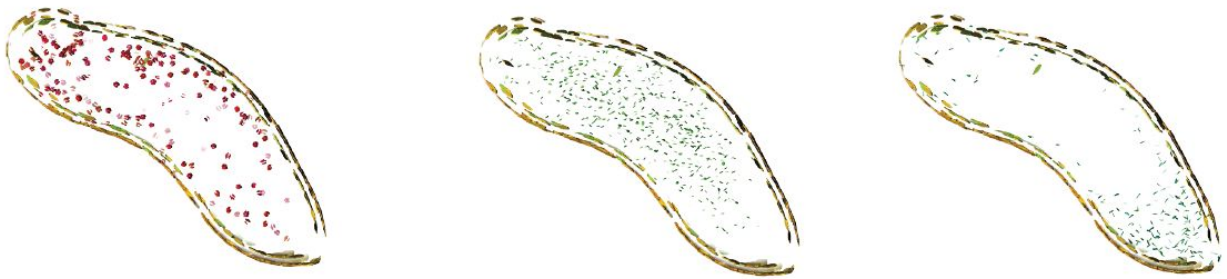
C: Data collection

D: 3D reconstruction & analysis

Weber et al, *Cells*, 2019

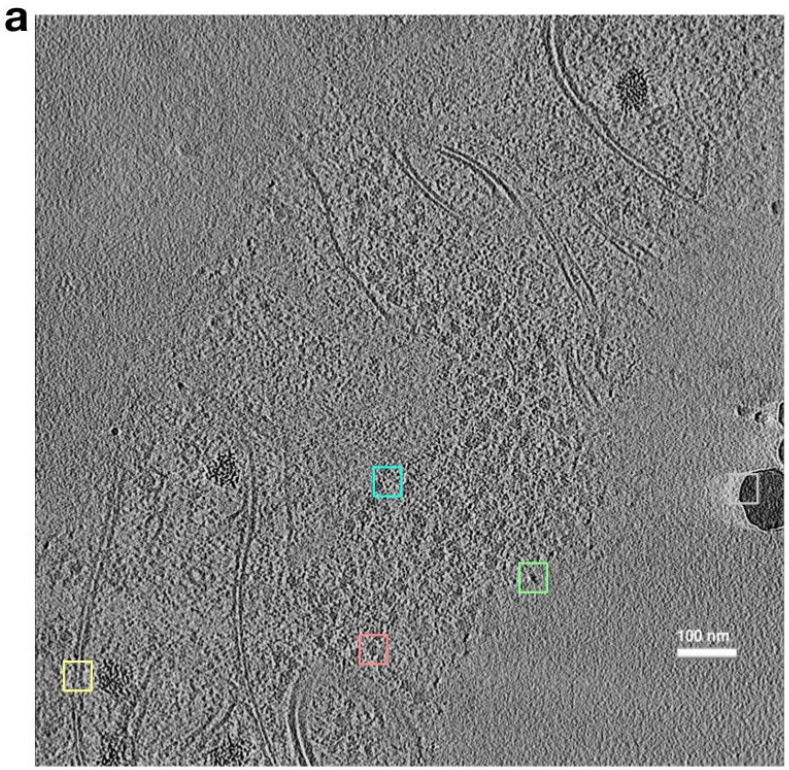
# Structural pattern mining

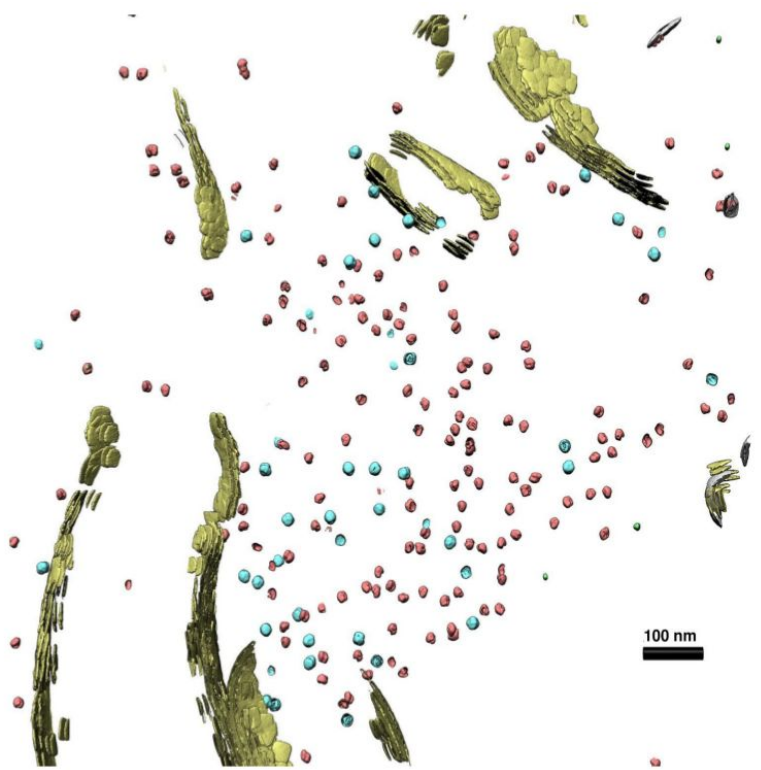


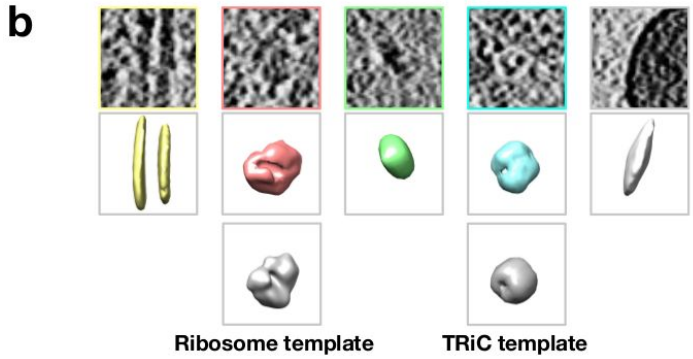




Xu et al 2019

# Classifying macromoleule structures in Cryo-ET



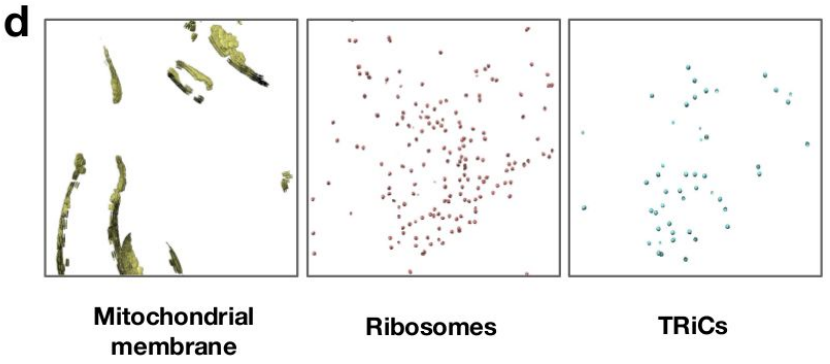




Image data from Guo et al 2018

# Subtomogram classification

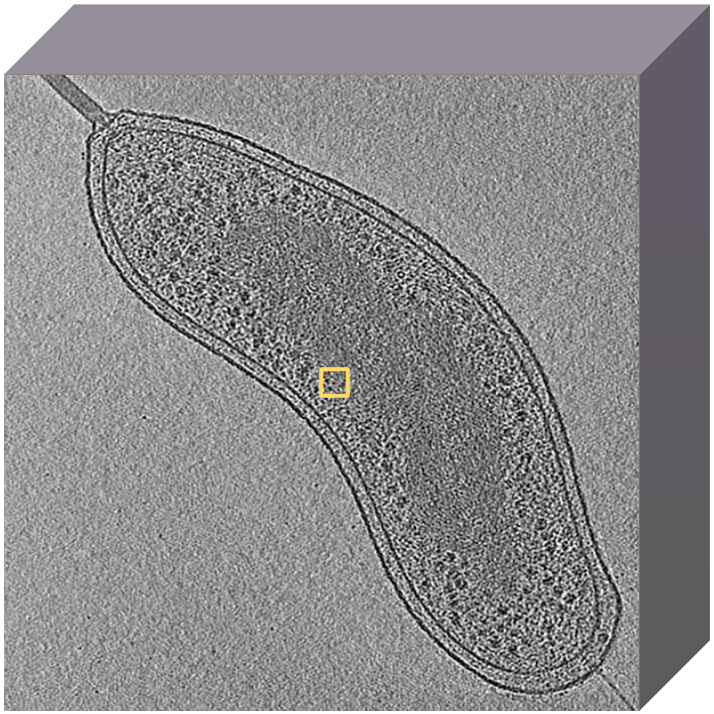
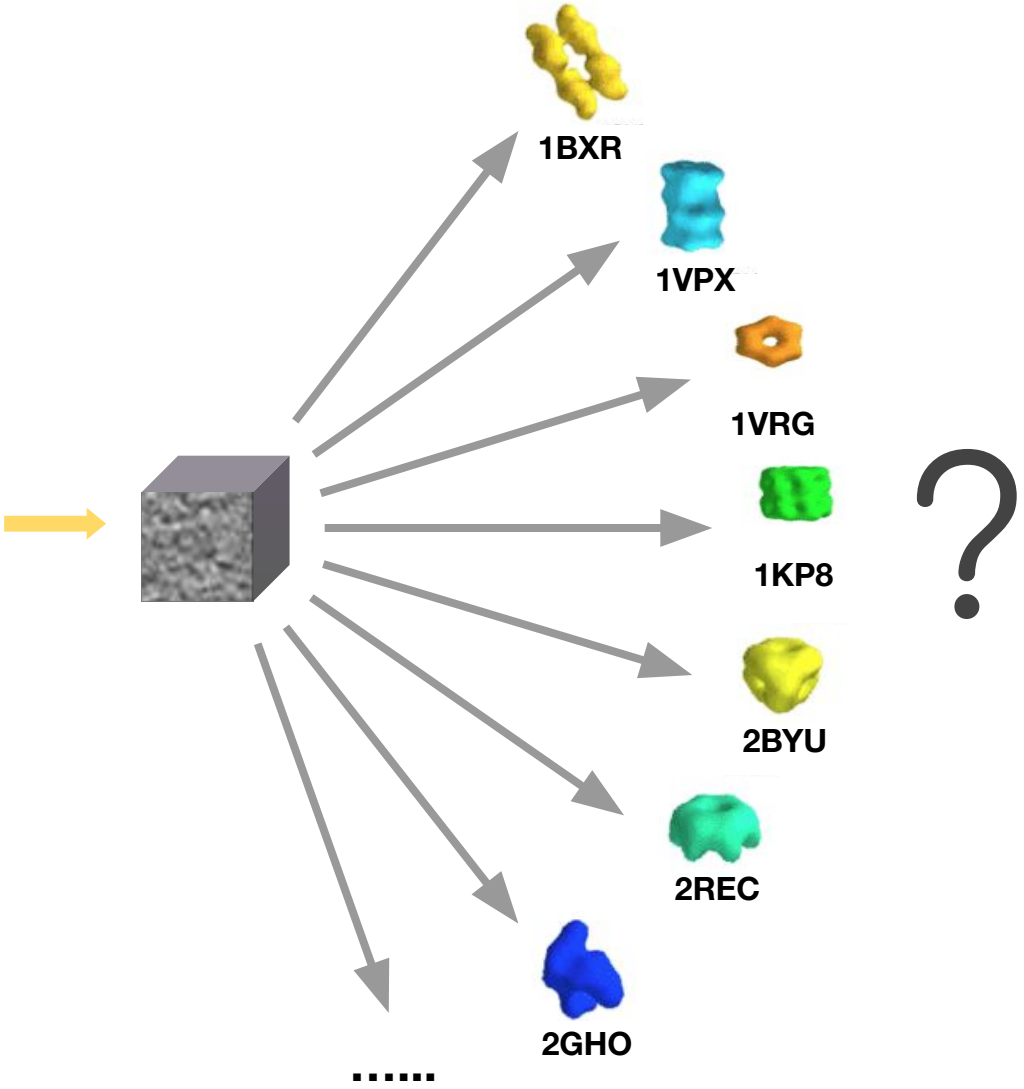


Image data from Jensen Lab




Xu et al, *Bioinformatics*, 2017

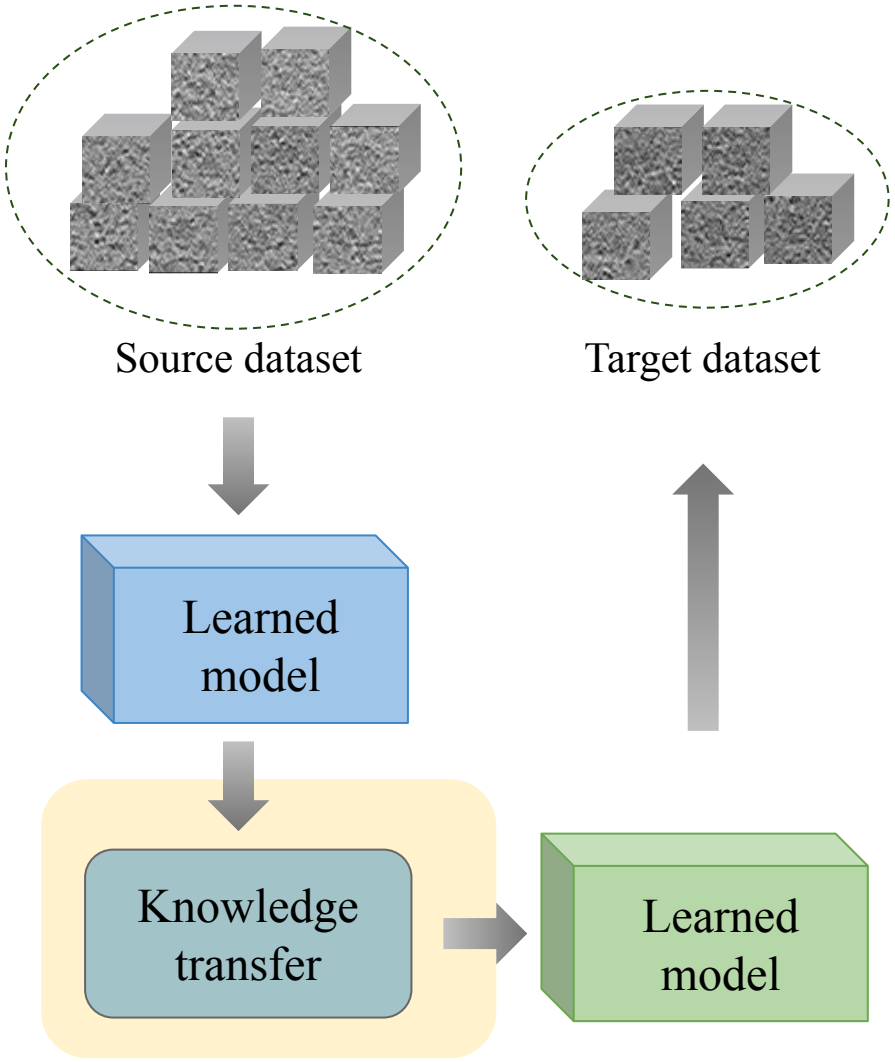
# Cross-domain classification problem

1. Deep learning based classification achieved signicant improvement in accuracy and throughput. However it requires large amount of training data.

2. Annotating a dataset for training is laborious

3. It is ideal to transfer the knowledge from simulated dataset or already annotated dataset to new dataset.

4. However different datasets have different image intensity distributions due to different experimental conditions.



# Domain shift

A subtomogram: $x$ Class label: $y$

Source domain: $D_s$ Target domain: $D_t$

Covariate shift:

$$\mathbb{P}_{x\sim D_s}(y|x) = \mathbb{P}_{x\sim D_t}(y|x), \text{ but } \mathbb{P}_{x\sim D_s}(x) \neq \mathbb{P}_{x\sim D_t}(x)$$

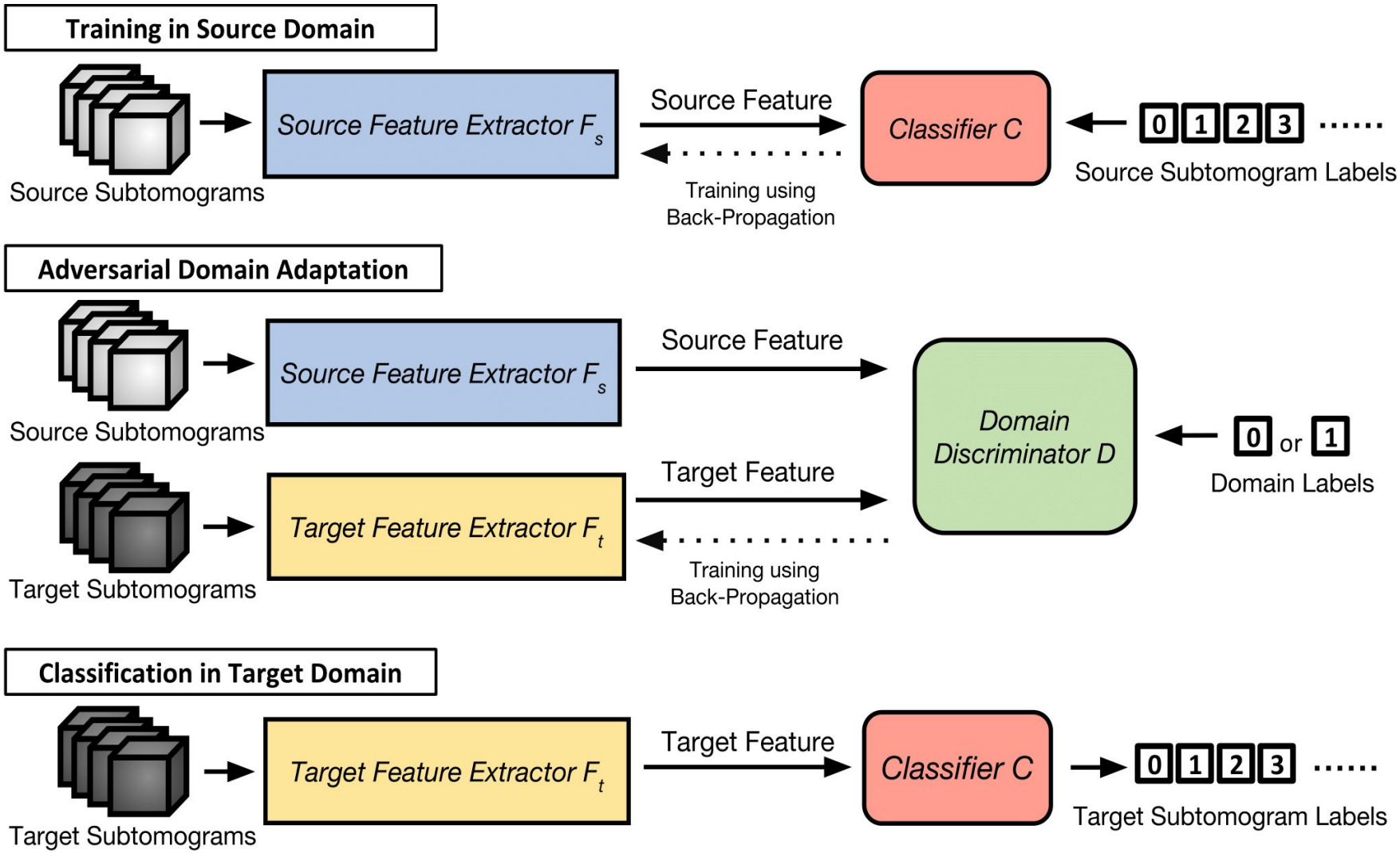
Prior probability shift:

$$\mathbb{P}_{x\sim D_s}(y|x) = \mathbb{P}_{x\sim D_t}(y|x), \text{ but } \mathbb{P}_{x\sim D_s}(y) \neq \mathbb{P}_{x\sim D_t}(y)$$

Concept shift:

$$\mathbb{P}_{x\sim D_s}(y|x) \neq \mathbb{P}_{x\sim D_t}(y|x)$$

# 3D Adversarial Domain Adaptation (ADA)




Adapted from Tzeng et al, *CVPR*, 2017

# Algorithm & Model architecture

**Algorithm 1** Adversarial Domain Adaptation Training

**Input:**

Set of subtomograms from source domain: $X_s$

Set of subtomograms from target domain: $X_t$

Domain labels: $L_s = 0$ and $L_t = 1$

Trained Source Feature Extractor: $F_s$

**Output:**

Trained domain discriminator: $D$

Trained target feature extractor :$F_t$

1: **for** $n$ training iterations **do**

2: **for** $k$ steps **do**

3: Sample minibatch of $m$ samples $\{x_s^1, \ldots, x_s^m\}$ from $X_s$.

4: Sample minibatch of $m$ samples $\{x_t^1, \ldots, x_t^m\}$ from $X_t$.

5: Update $D$ by ascending stochastic gradient of $L_D$, with $F_t$ fixed:

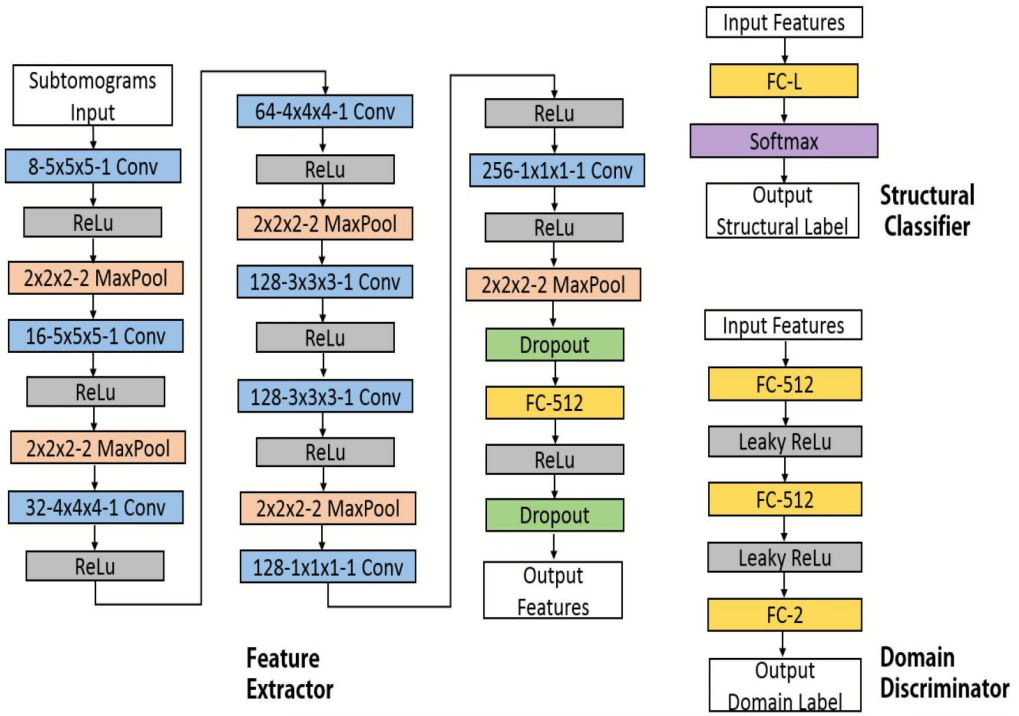
$$\nabla_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} \left[ -\log\left(D\left(F_t\left(x_t^i\right)\right) - L_s\right) - \log\left(L_t - D\left(F_s\left(x_s^i\right)\right)\right)\right]$$

6: Sample minibatch of $m$ target samples $\{x_t^1, \ldots, x_t^m\}$ from $X_t$.

7: Update $F_t$ by descending stochastic gradient of $L_F$ with the $D$ fixed:

$$\nabla_{\theta_{F_t}} \frac{1}{m} \sum_{i=1}^{m} \left[ -\log\left(L_t - D\left(F_t\left(x_t^i\right)\right)\right)\right]$$

8: **return** $D$, $F_t$



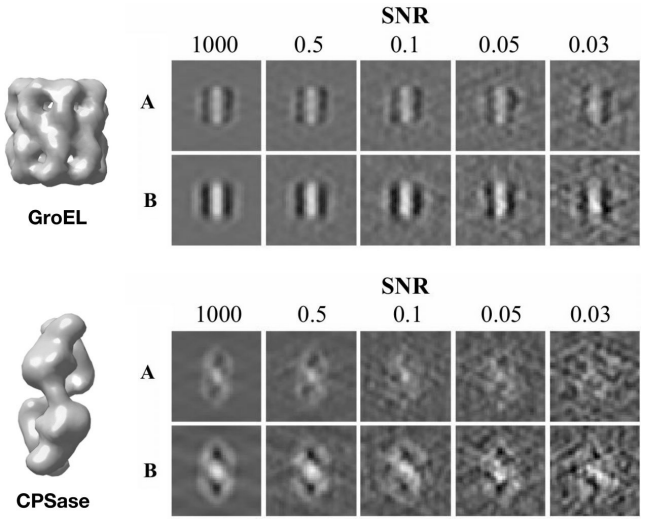
# Simulated datasets

Dataset batch A and B, each contains 5 datasets with different SNR

23*1000 subtomograms in each dataset

Different imaging condition: spherical aberration and defocus

| PDB ID | Macromolecular complex |
|---|---|
| 1A1S | Ornithine carbamoyltransferase |
| 1BXR | Carbamoyl phosphate synthetase |
| 1EQR | Aspartyl tRNA-synthetase |
| 1F1B | E. coli aspartate transcarbamoylase |
| 1FNT | Yeast 20S proteasome with activator |
| 1GYT | Aminopeptidase a |
| 1KP8 | GroEL-KMgATP 14 |
| 1LB3 | Mouse L-chain ferritin |
| 1QO1 | Rotary motor in ATP synthase |
| 1VPX | Transaldolase |
| 1VRG | Propionyl-CoA carboxylase, beta subunit |
| 1W6T | Octameric enolase |
| 1YG6 | ClpP |
| 2BO9 | Human carboxypeptidase A4 |
| 2BYU | Small heat shock protein Acr1 |
| 2GHO | Recombinant thermus aquaticus RNA polymerase |
| 2GLS | Glutamine synthetase |
| 2H12 | Acetobacter aceti citrate synthase |
| 2IDB | 3-octaprenyl-4-hydroxybenzoate decarboxylase |
| 2REC | RecA hexamer |
| 3DY4 | Yeast 20S proteasome |
| 4V4Q | Bacterial ribosome |
| NULL | (No particle) |



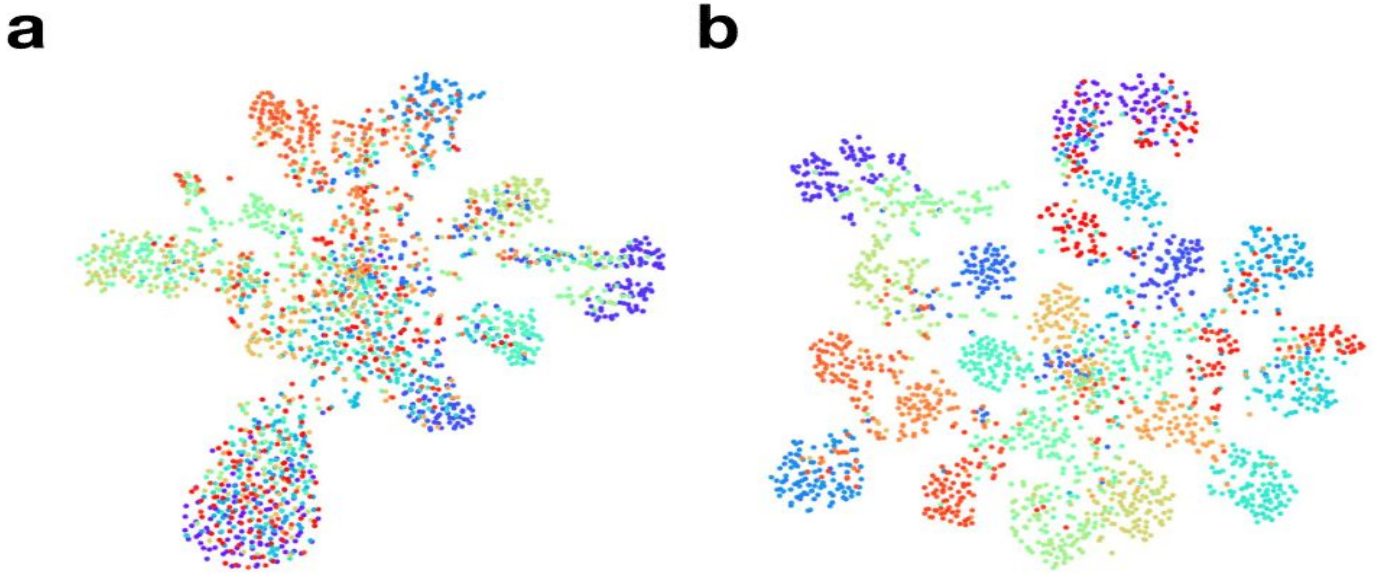
# Cross-domain prediction accuracy

| Accuracy | | SNR of Target Domain ($S_A$) | | | | |
|---|---|---|---|---|---|---|
| SNR of Source Domain ($S_B$) | | **1000** | **0.5** | **0.1** | **0.05** | **0.03** |
| | **1000** | 0.855 | 0.687 | 0.385 | 0.235 | 0.157 |
| | | 0.739 | 0.620 | 0.287 | 0.159 | 0.111 |
| | | 0.760 | 0.638 | 0.289 | 0.162 | 0.114 |
| | | **0.991** | **0.923** | **0.737** | **0.499** | **0.326** |
| | **0.5** | 0.779 | 0.757 | 0.547 | 0.366 | 0.258 |
| | | 0.806 | 0.710 | 0.479 | 0.372 | 0.291 |
| | | 0.819 | 0.723 | 0.486 | 0.373 | 0.291 |
| | | **0.978** | **0.970** | **0.835** | **0.628** | **0.464** |
| | **0.1** | 0.902 | 0.922 | 0.894 | 0.726 | 0.503 |
| | | 0.864 | 0.881 | 0.776 | 0.637 | 0.475 |
| | | **0.905** | 0.920 | 0.826 | 0.650 | 0.479 |
| | | 0.894 | **0.932** | **0.901** | **0.760** | **0.626** |
| | **0.05** | 0.946 | 0.950 | 0.911 | 0.766 | 0.563 |
| | | 0.937 | 0.929 | 0.897 | 0.758 | 0.575 |
| | | 0.948 | 0.951 | 0.907 | 0.774 | 0.583 |
| | | **0.967** | **0.971** | **0.928** | **0.825** | **0.628** |
| | **0.03** | 0.938 | 0.924 | 0.903 | 0.844 | 0.704 |
| | | 0.903 | 0.891 | 0.864 | 0.775 | 0.609 |
| | | 0.907 | 0.893 | 0.865 | 0.778 | 0.613 |
| | | **0.976** | **0.972** | **0.952** | **0.891** | **0.773** |

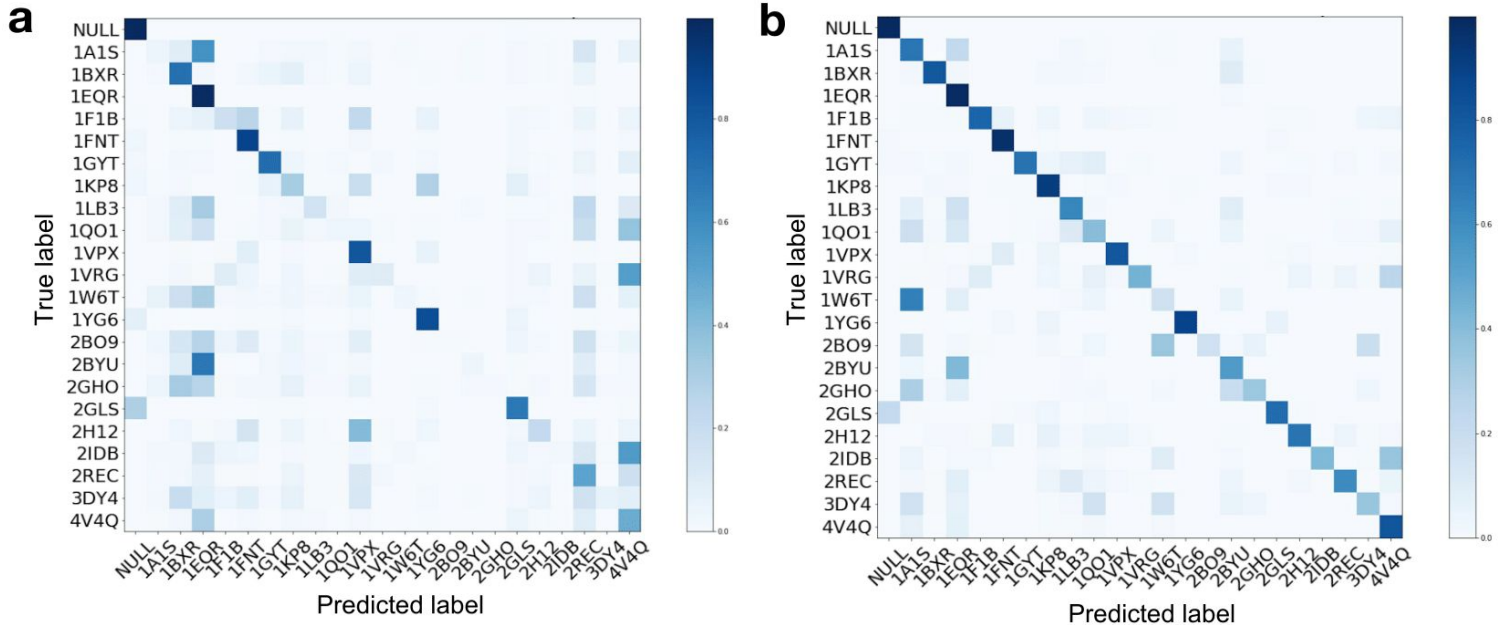
- No DA
- Direct importance estimation
- Structural correspondence learning
- 3D ADA

# Result visualization

TSNE embedding




Confusion matrix




Before ADA

After ADA

# Experimental dataset

1. Purified human 20S Proteasome and *E.coli* Ribosome (Zeev-Ben-Mordehai *et al.* 2016)

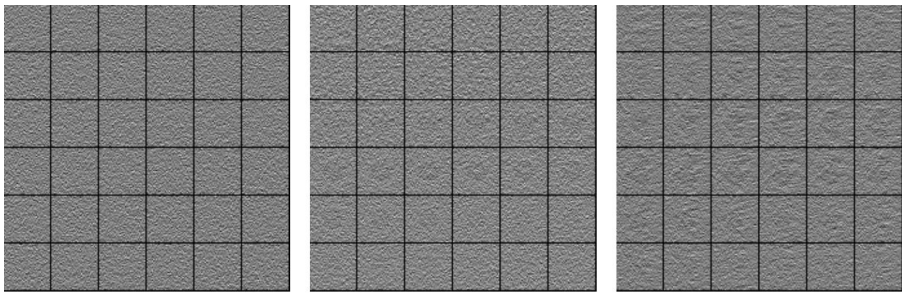
   100 subtomograms in each class (including a None class)

2. Ribosome, TRiC, Proteasome from a rat neuron culture tomogram (Guo *et al.* 2018)

   80 subtomograms in each class

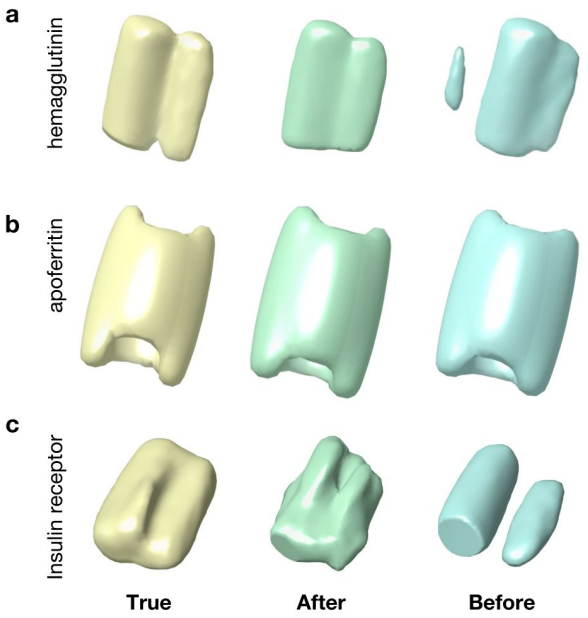
3. Purified Hemagglutinin, Apoferritin, Insulin receptor (Noble *et al.* 2018)

   400 subtomograms in each class

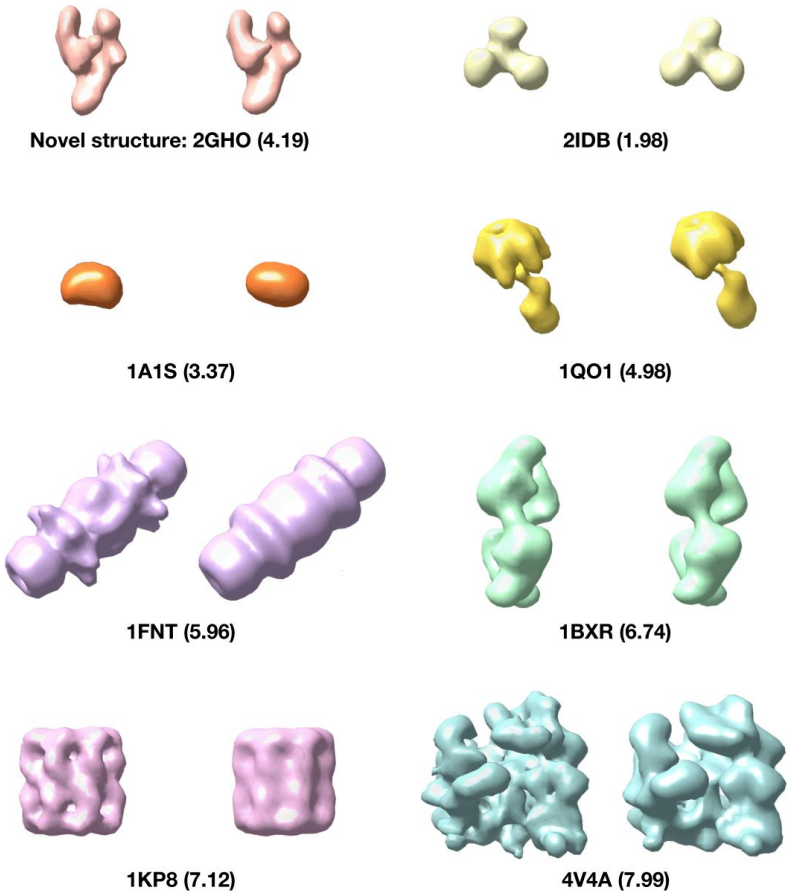


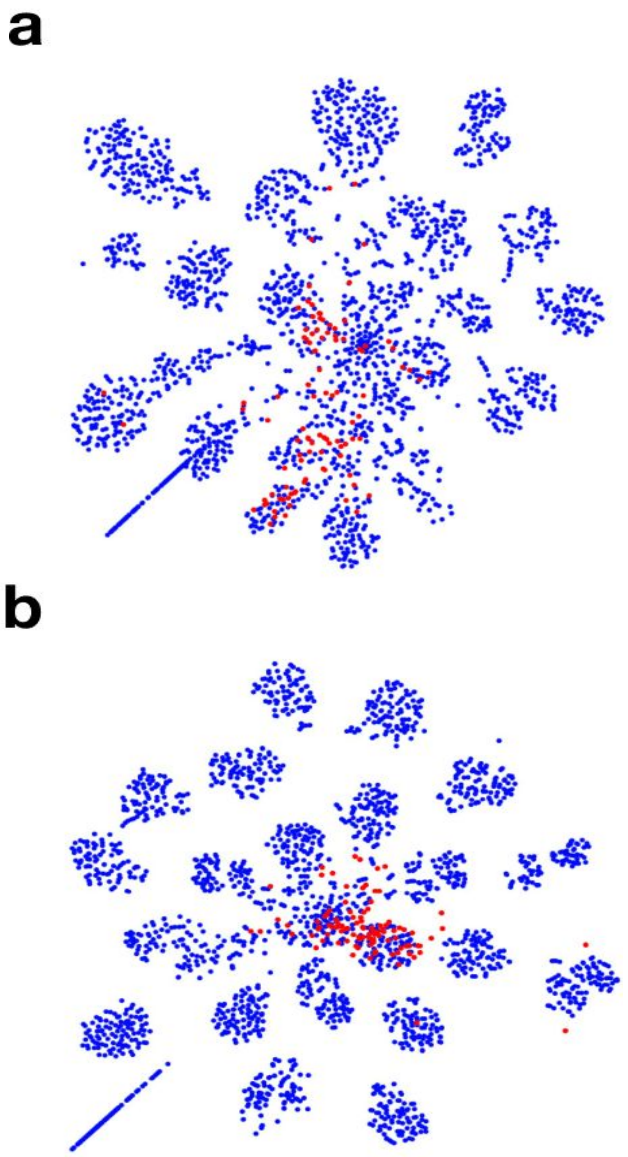
# Cross-domain prediction accuracy

| Accuracy | SNR of Source Domain | | | | |
|---|---|---|---|---|---|
| Dataset: | 1000 | 0.5 | 0.1 | 0.05 | 0.03 |
| $S_{e1}$ | 0.375 | 0.313 | 0.465 | 0.331 | 0.566 |
| | **0.578** | **0.641** | **0.563** | **0.584** | **0.606** |
| $S_{e2}$ | 0.400 | 0.370 | 0.311 | 0.308 | 0.336 |
| | **0.495** | **0.469** | **0.471** | **0.450** | **0.377** |
| $S_{e3}$ | 0.313 | 0.376 | 0.375 | 0.372 | 0.375 |
| | **0.688** | **0.656** | **0.625** | **0.621** | **0.624** |



# Improvement on novel structure recovery



# Conclusion

1. Lack of annotated data is a major bottlenect for deep learning based supervised subtomogram classification
2. Beneficial to have training data from a separate data source where the annotation is readily available or can be performed in a high-throughput fashion.
3. Domain shift is a major bottleneck in cross-domain subtomogram classification, which leads to low prediction accuracy.
4. 3D-ADA stably improves the cross-domain prediction under different imaging conditions.

# Automatic Localization and Identification of Mitochondria in Cellular Electron Cryo-Tomography using Faster-RCNN

R Li*, X Zeng*, et al. *BMC Bioinformatics*

# Automatic Localization and Identification of Mitochondria in Cellular Electron Cryo-Tomography using Faster-RCNN

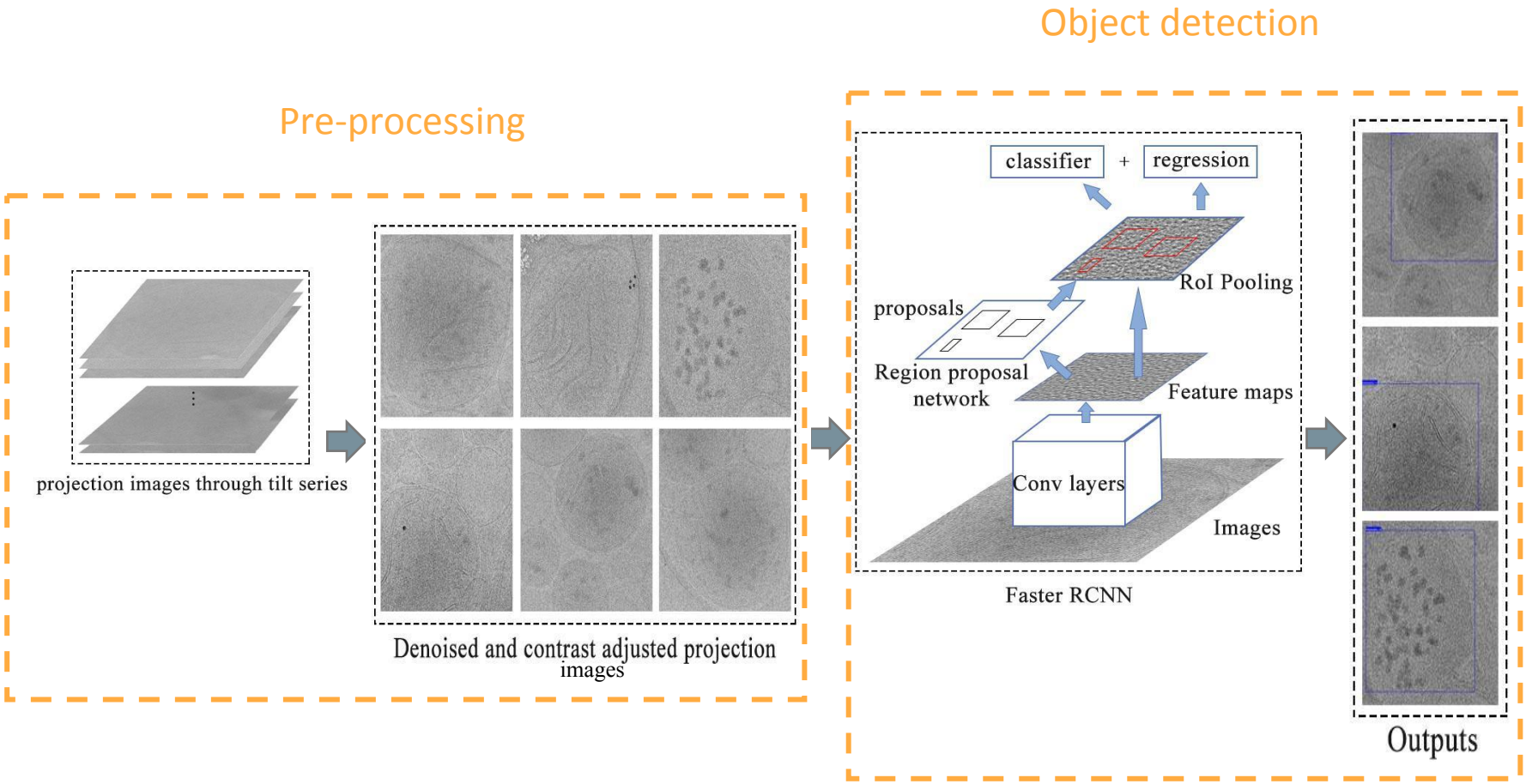
# Background

- Existing methods:
  - Manual segmentation
    - Time- and effort- consuming
  - Automatic segmentation
    - Focus on specific structures
    - Need precise annotations of contours

- Our goal: a simple and generic method of automatic identification and localization of subcellular structures of interest within *in situ* cryo-ET images with weak annotations

# Method

The flowchart of our model



# Method

- Preprocessing
  - Bilateral filtering[1]

$$h(x) = k^{-1}(x)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi)c(\xi,x)s(f(\xi),f(x))d\xi \qquad (1)$$

Where $c(\xi,x)$ is related to the distance between point $x$ and $\xi$, and $s(f(\xi),f(x))$ is related to the difference between the intensity of $x$ and $\xi$.

Choose $c(\xi,x) = e^{-\frac{1}{2}\left(\frac{\|\xi-x\|}{\sigma_d}\right)^2}, s(f(\xi),f(x)) = e^{-\frac{1}{2}\left(\frac{\|f(\xi)-f(x)\|}{\sigma_r}\right)^2}$

Then $\sigma_d$ is the spatial parameter, and $\sigma_r$ is the range parameter.

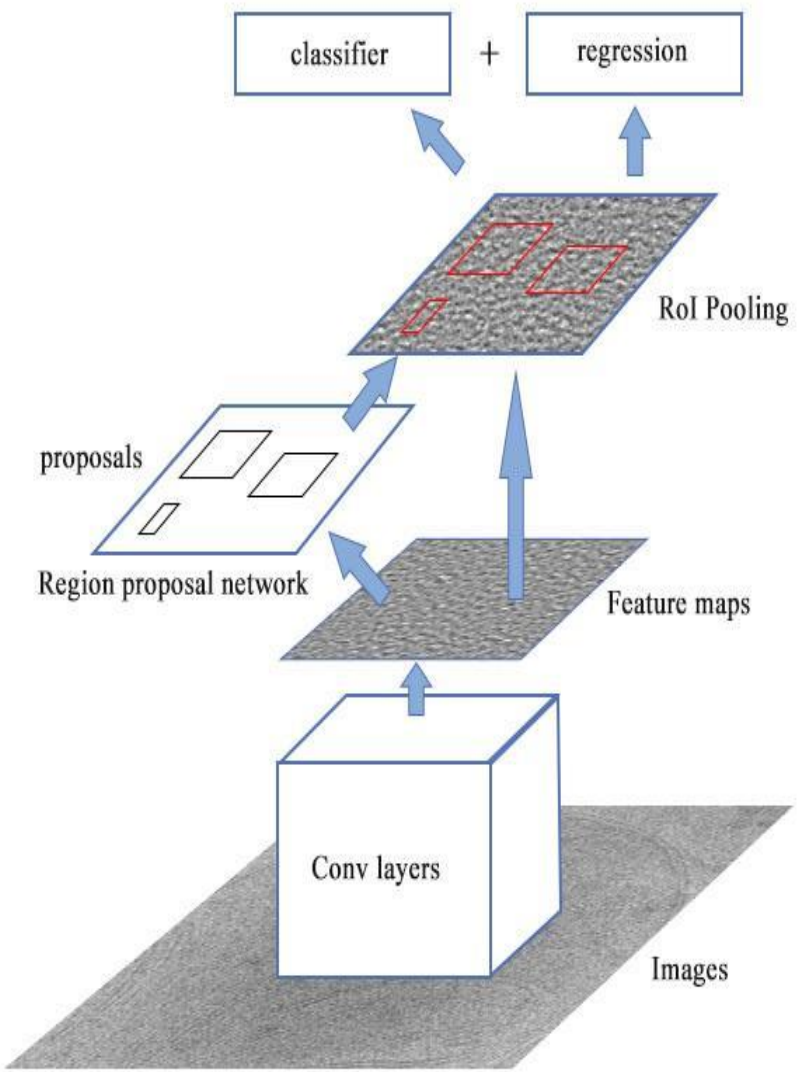
  - Histogram Equalization

Improve local contrast through evenly distributing grayscale in the histogram

[1] Tomasi, C., Manduchi, R. Bilateral filtering for gray and color images[C]. Proceedings of the IEEE International Conference on Computer Vision, 1998: 59-66

# Method: object detection based on Faster RCNN

- Object detection in 2D images
  - Faster RCNN[1]
    - Feature extraction
    - Region proposal generation
    - RoI pooling
    - Classification and regression
- Application in reconstructed tomogram slices




Faster RCNN

[1] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. Conference and Workshop on Neural Information Processing Systems(NIPS). 2015.

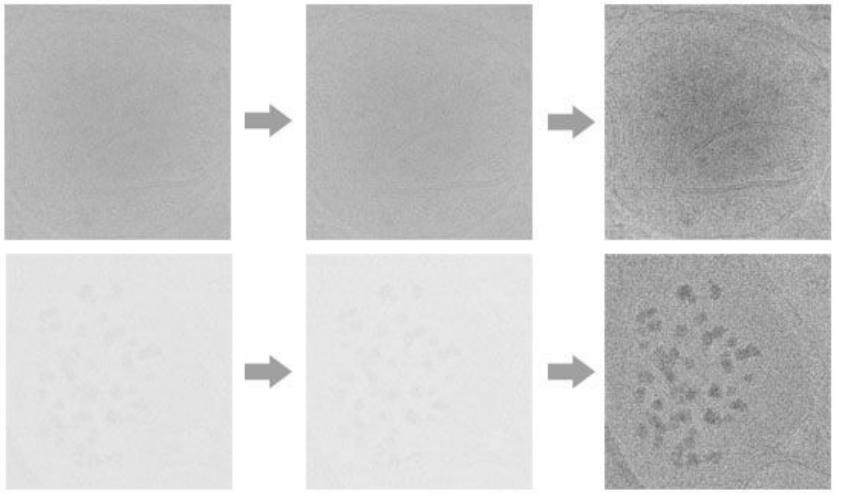
# Results

- Dataset
  - 9 tomograms containing mitochondria
  - 486 2D slices manually annotated through LabelImg
  - Train set:402
  - Test set:80
  - Annotation format:
    - PASCAL VOC
- Metrics: AP, IoU, F1 score

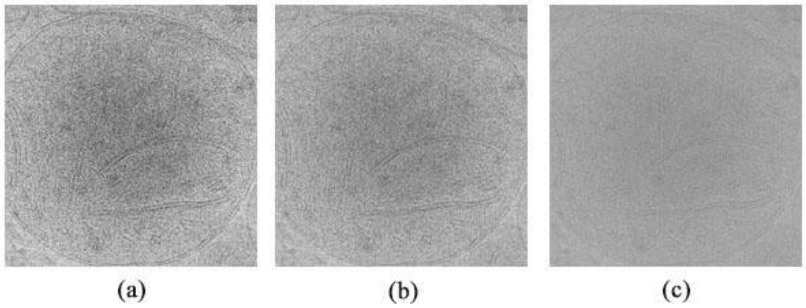
| Tomogram basename | Image size | All slices | Used slices |
|---|---|---|---|
| Unstim_20k_mito1 | 3708×3838 | 101 | 75 |
| Unstim_20k_mito2 | 3708×3838 | 89 | 44 |
| CTL_Fibro_mito1 | 3708×3838 | 82 | 36 |
| M2236_Fibro_mito2 | 3708×3838 | 90 | 46 |
| M2236_truemito3 | 3708×3838 | 86 | 39 |
| CHX + Glucose Stimulation A2 | 3708×3838 | 53 | 51 |
| HighGluc_Mito1 | 3708×3838 | 101 | 71 |
| HighGluc_Mito2 | 3708×3838 | 101 | 69 |
| INS_21_g3_t10 | 3708×3838 | 81 | 51 |
| Total | | 786 | 482 |

# Results

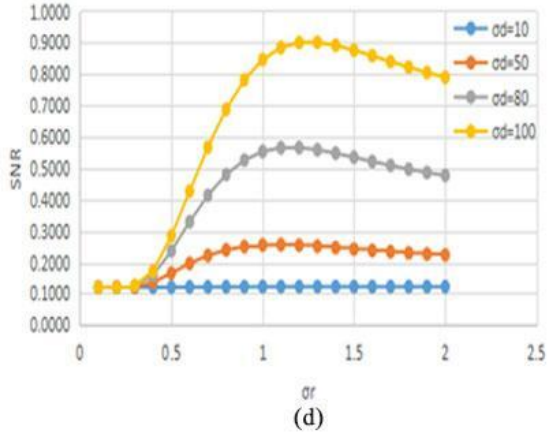
- Data preprocessing: noise reduction and contrast enhancement



(a) Original (b) Noise reduction (c) Contrast adjustment



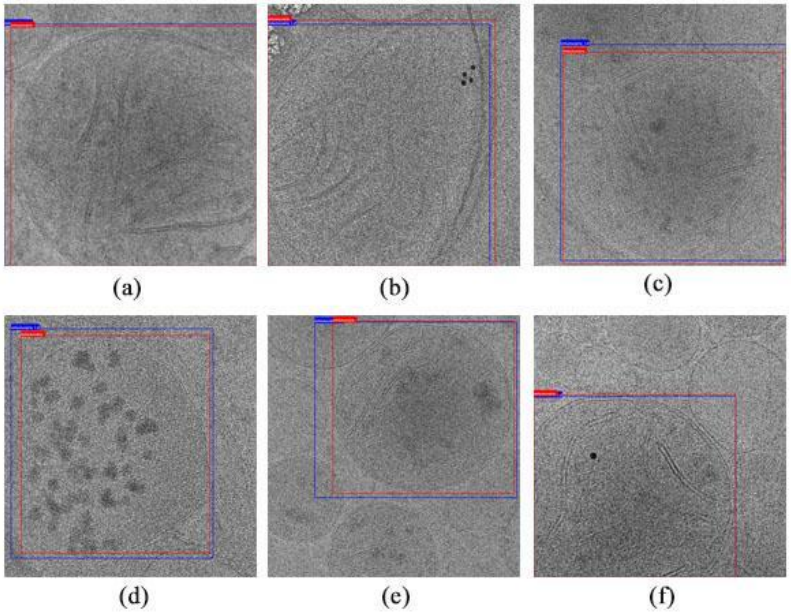
(a) (b) (c)

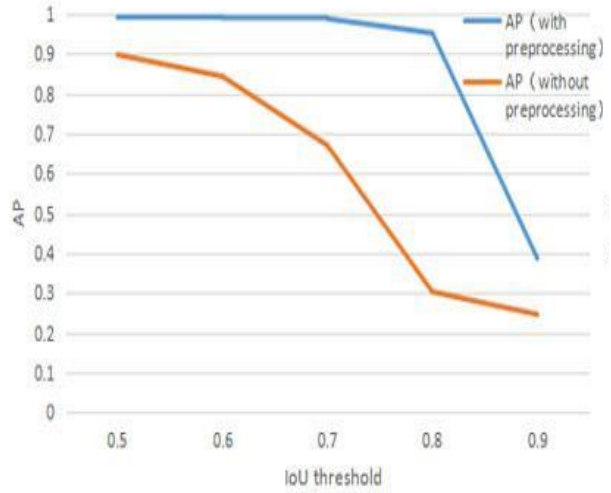



(d)

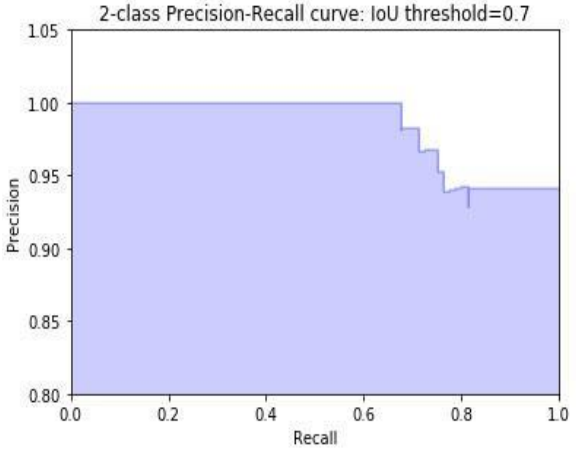
# Results

- Prediction performance




(a)



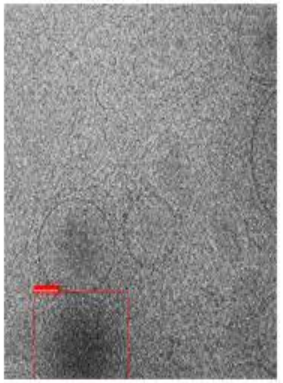

(b)

# Results

- Source of error
  - Too small mitochondria
  - Incomplete structure
  - Quality of the original image

| Tomogram basename | $F_1$ *score* | AP | mIoU | Incomplete mitochondria |
|---|---|---|---|---|
| Unstim_20k_mito1 | 0.91 | 0.98 | 0.826 | YES |
| Unstim_20k_mito2 | 1 | 1 | 0.864 | NO |
| CTL_Fibro_mito1 | 0.97 | 0.99 | 0.843 | NO |
| M2236_Fibro_mito2 | 0.96 | 0.99 | 0.887 | YES |
| M2236_turemito3 | 0.91 | 0.97 | 0.783 | NO |
| CHX + Glucose Stimulation A2 | 0.94 | 1 | 0.75 | YES |
| HighGluc_Mito1 | 0.97 | 0.99 | 0.843 | NO |
| HighGluc_Mito2 | 0.97 | 0.96 | 0.837 | NO |
| INS_21_g3_t10 | 0 | 0 | 0 | YES |

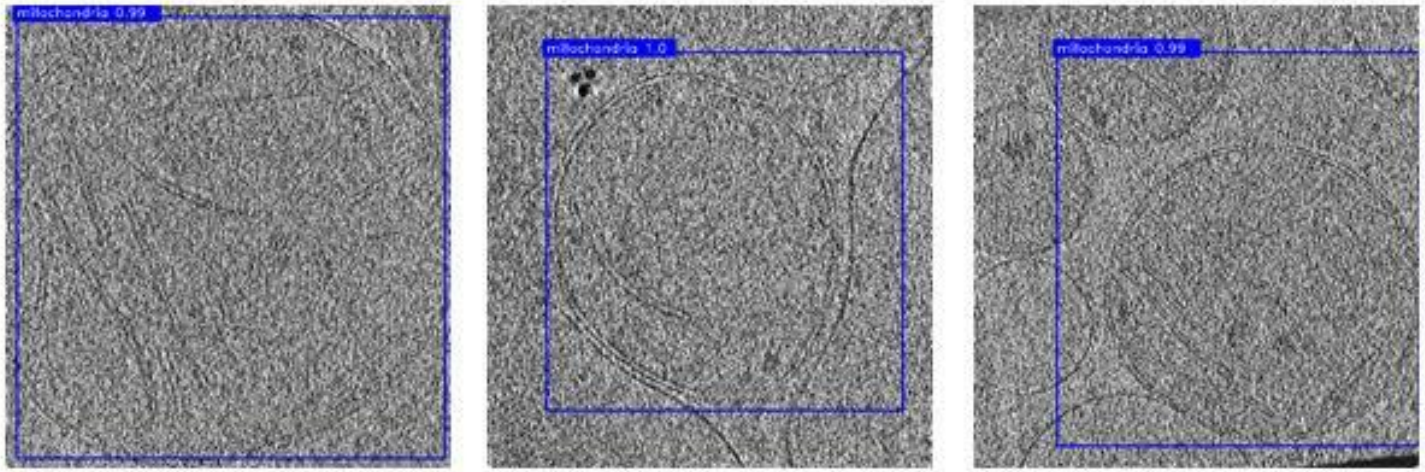
(a)



(b)

# Results

- Prediction on 3D tomogram slices




(a) (b) (c)

# Conclusion

- The first work to apply Faster-RCNN model to Cryo-ET data
- Demonstrated the high accuracy ($AP > 0.95$ and $IoU > 0.7$) and robustness of detection and classification tasks of intracellular mitochondria
- Can be generalized to detect multiple cellular components

- Future work
  - Improving the accuracy of localization
  - Exploring the effects of different network structures

Thank you