Deep learning based structural pattern mining in tomograms -- several exploratory studies

Min Xu
Computational Biology Department
School of Computer Science
Carnegie Mellon University

Systematic detection of macromolecular structures in cellular tomograms

Structural pattern mining / in silico purification: template-free detection of macromolecular structures

Challenges

- Imaging limits
  - Missing data (missing wedge effect)
  - Low signal-to-noise ratio
- High structural content complexity
  - Macromolecule structure highly diverse
  - High molecular crowding level
- Big data
  - Hundreds of tomograms
  - Millions of macromolecules

Deep learning

- Convolutional neural network (CNN) performs regression using large amount of parameters
    - Multiple layers + nonlinearity → exponential increase of flexibility for approximation of arbitrary mapping between input and output
    - Convolution: parameter sharing & local connectivity→ increases efficiency by taking advantage of composition structure in images
    - Stacked convolutional layers → learning of image feature hierarchy
- Linear scalability respect of training sample number → learn from big data
    - Back-propagation training, easy to implement and parallelize on GPU
- Dropout → improved generalization ability

Exploratory projects

1. Macromolecule structure classification and subdivision

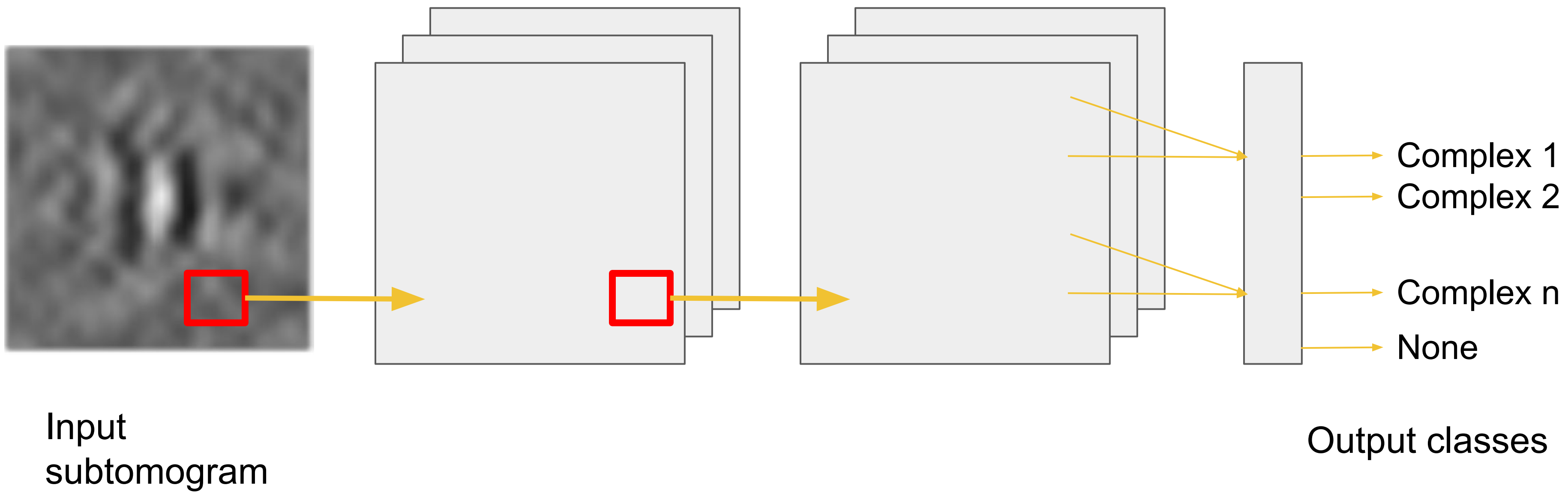
2. Autoencoder based pattern detection

3. Subtomogram segmentation

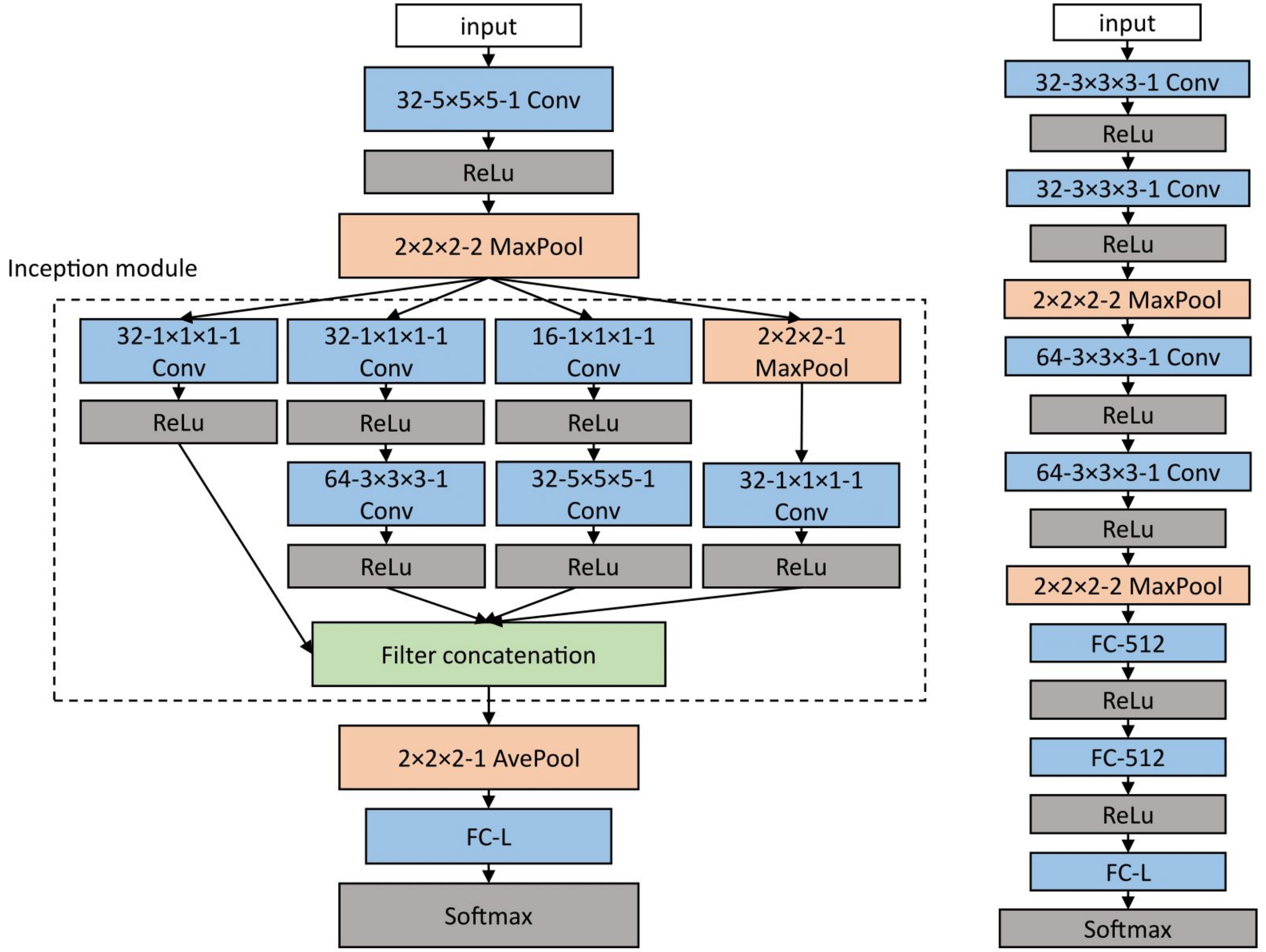
Supervised subtomogram classification

Xu et al. ISMB 2017

Supervised subtomogram classification



CNN classification models



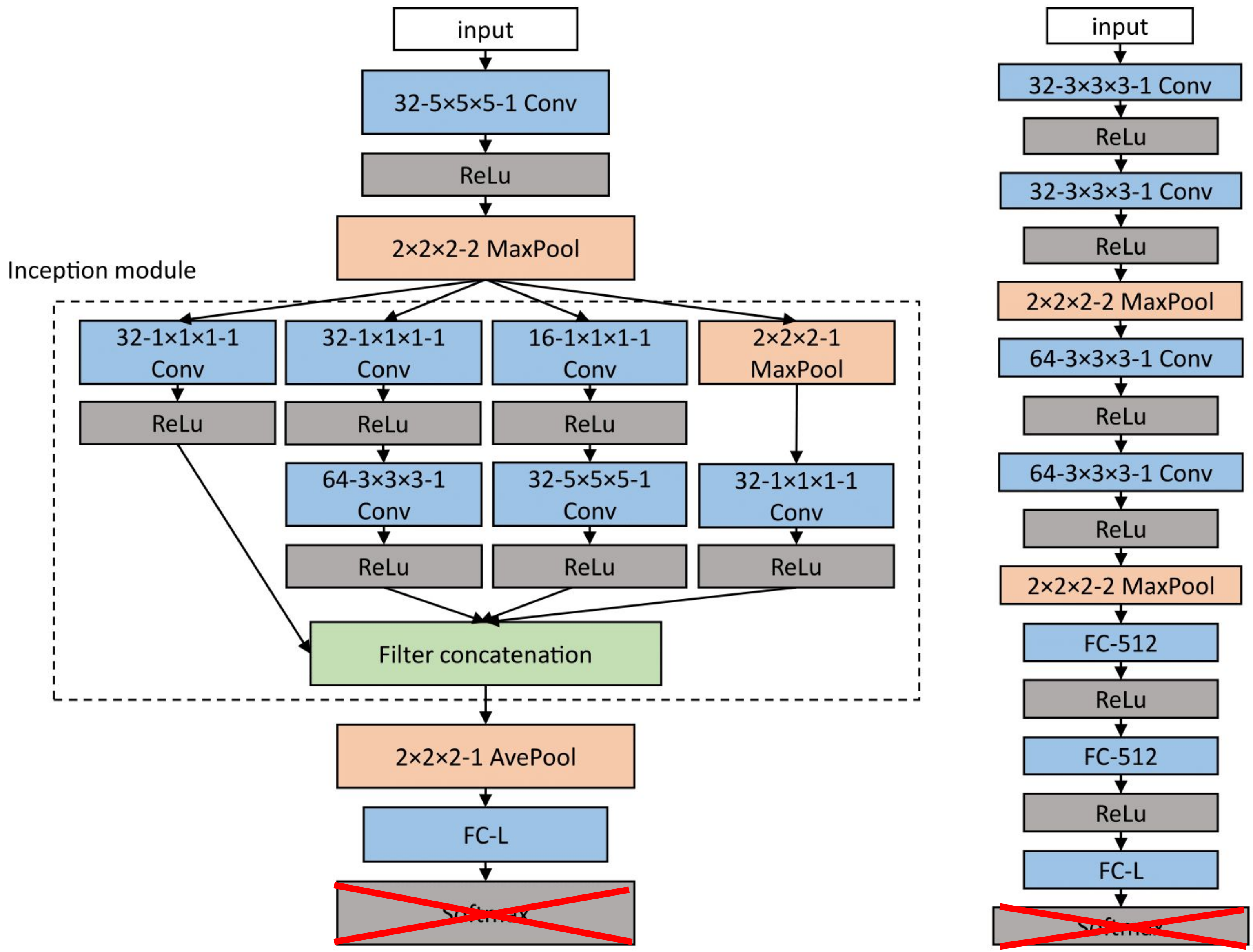
(a) Inception3D network (b) DSRF3D network

Performance

- Classification accuracy significantly better than Rotational Invariant Features + Support Vector Machines

- Once trained, classifying 1M subtomograms take < 2 hours on a single GPU

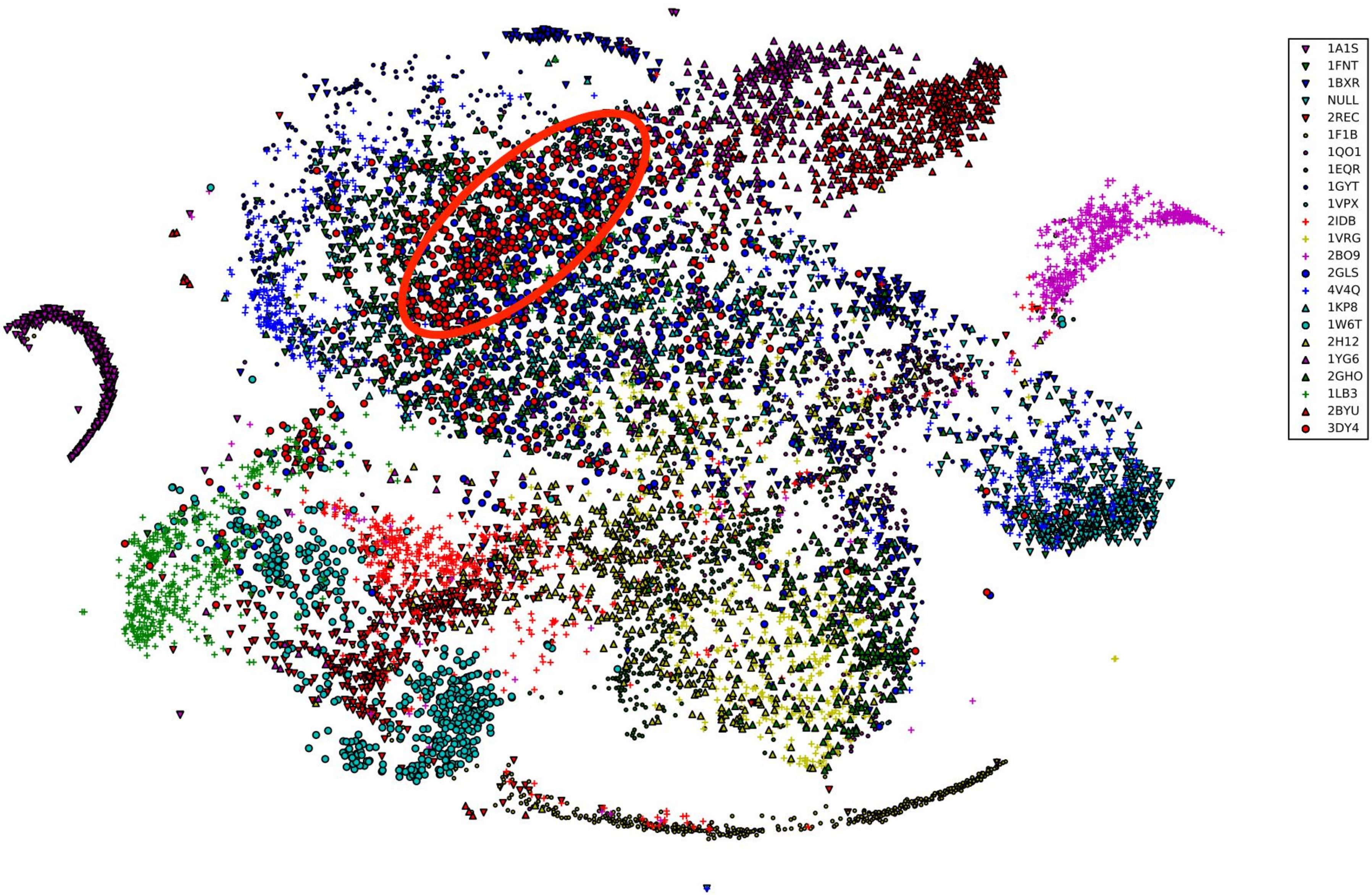
Supervised structural feature extraction



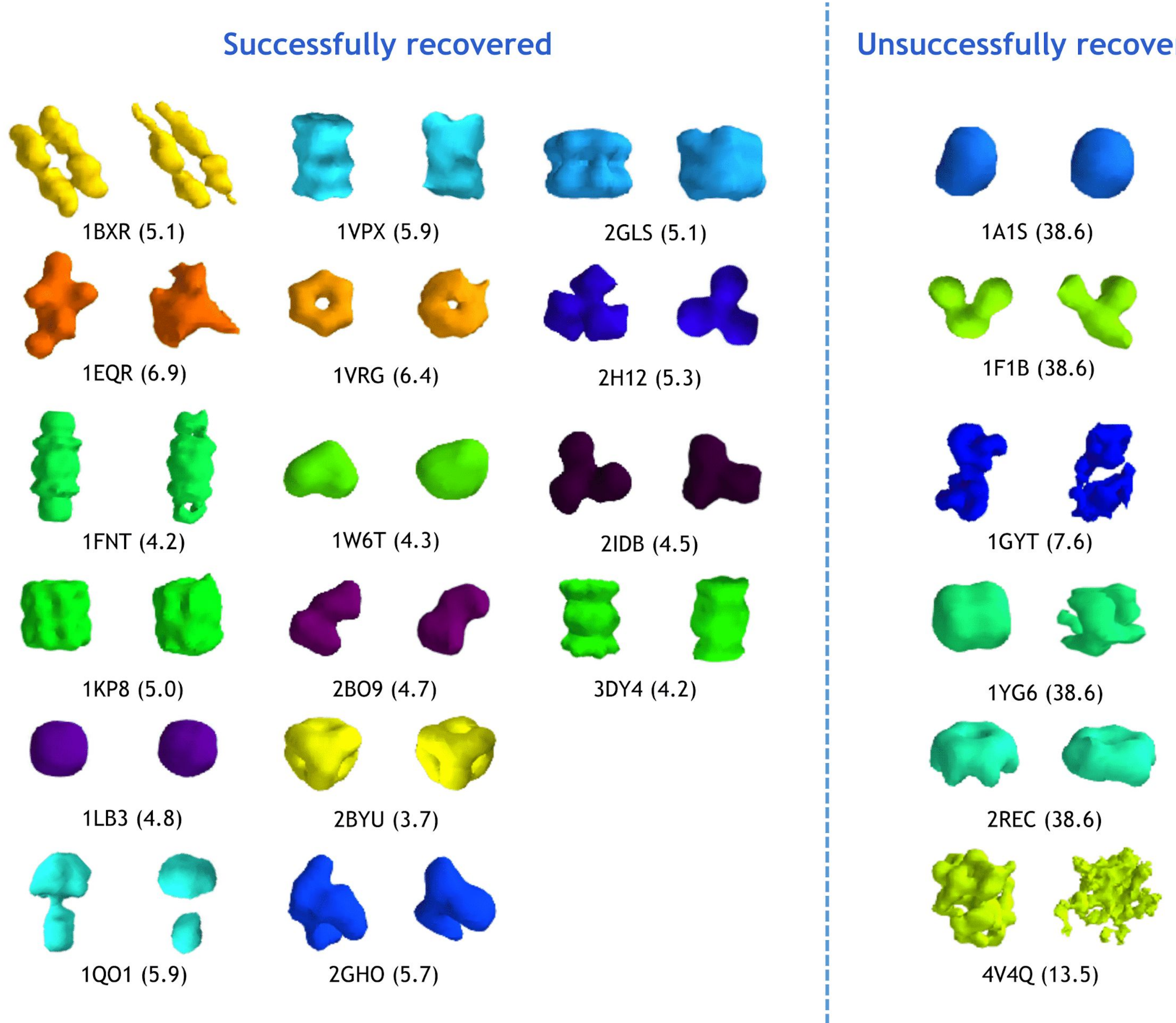
Supervised structural feature extraction

- Continuous representation of the likelihood of the class assignments
- Project the input subtomogram into a low dimensional structural feature space spanned by the training classes
- Invariant to
  - Rigid transformations
  - Missing wedge effects

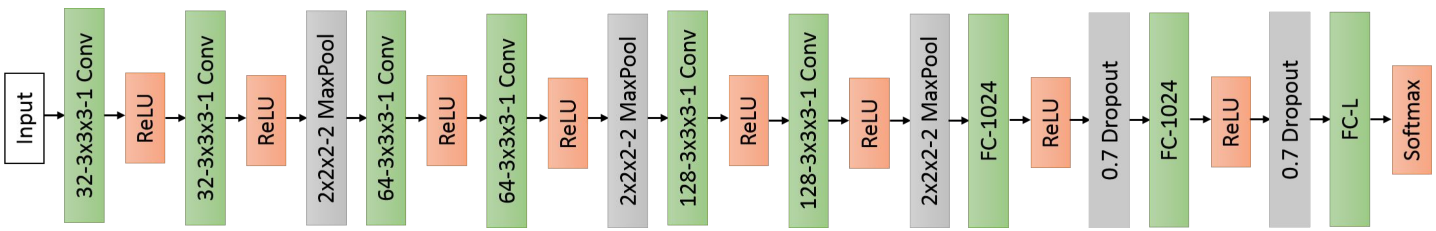
Detection of new structures



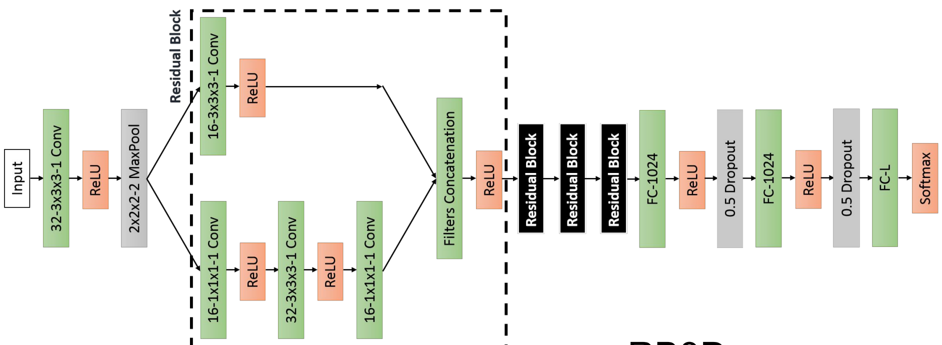
Detection of new structures: leave-one-out test



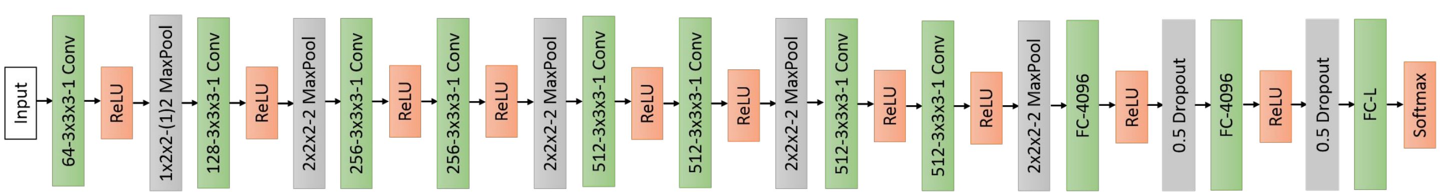
Improvements: deeper models for improved accuracy



DSRF3D-v2



RB3D

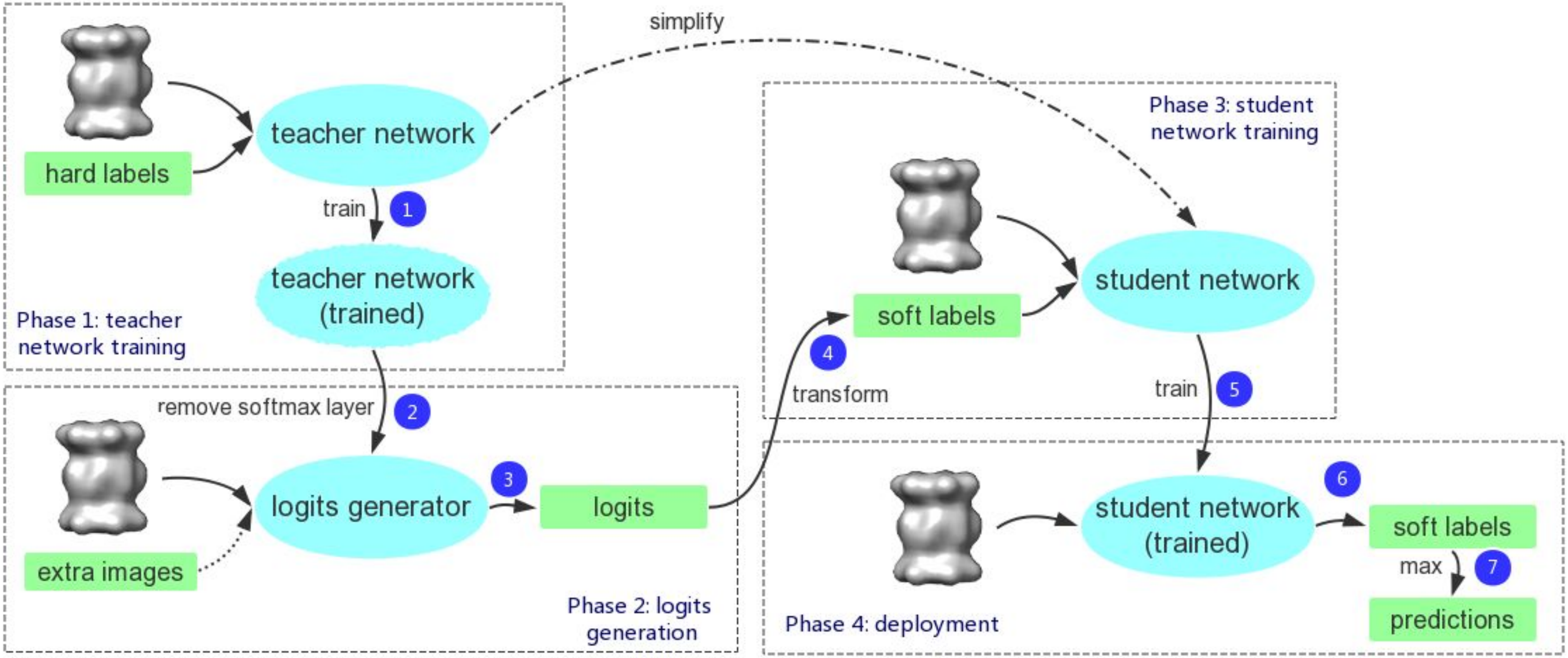


(Best performance)

CB3D

Che et al. arXiv:1707.04885

Improvements: model compression for increased speed

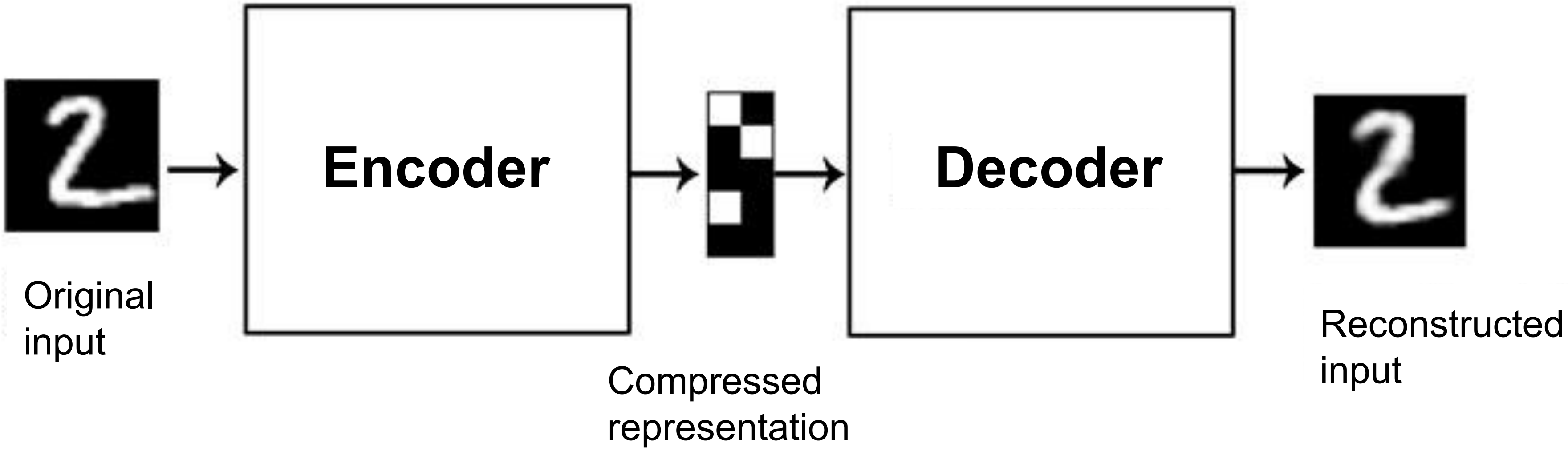


Guo et al. ICIAR 2018

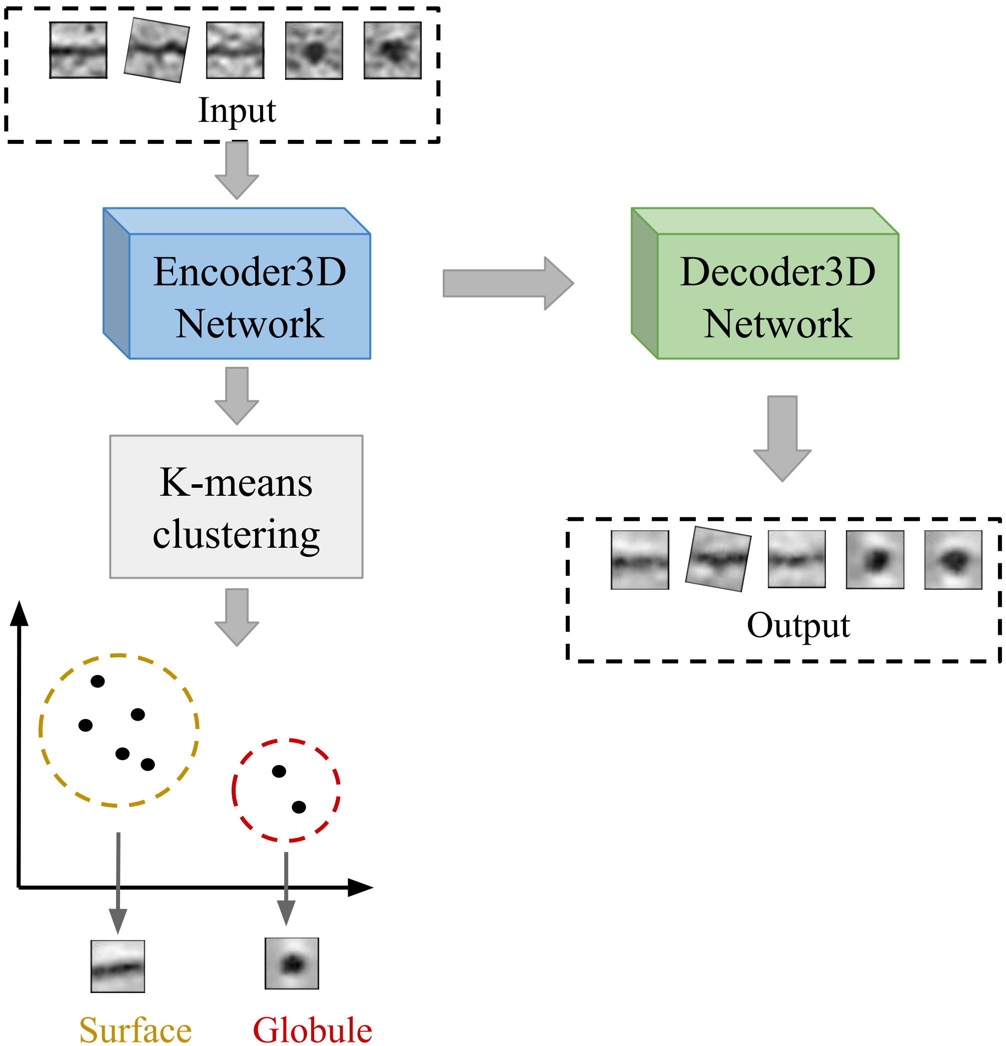
Autoencoder based pattern detection

Zeng et al. JSB 2017

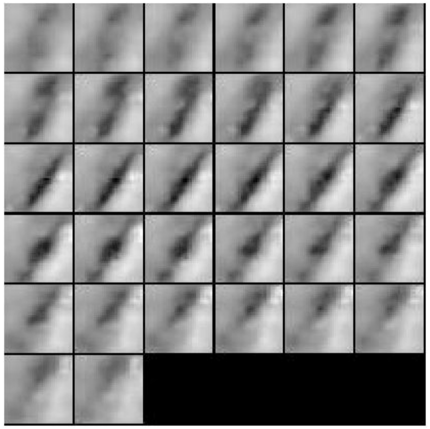
Autoencoder based pattern detection



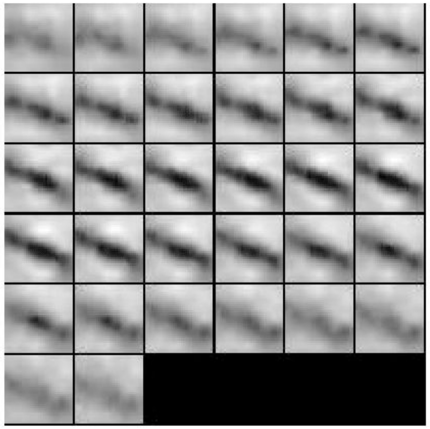
Autoencoder based pattern detection



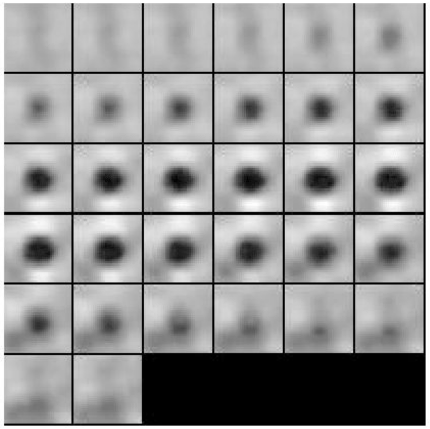
Autoencoder based pattern detection



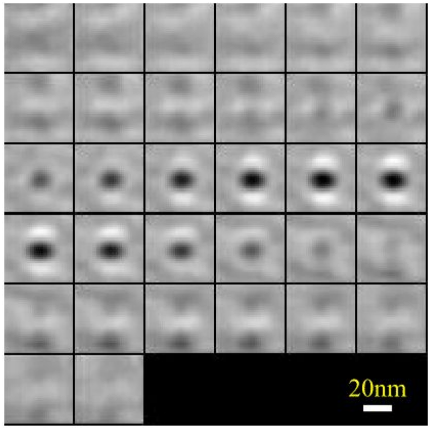
Surface patch



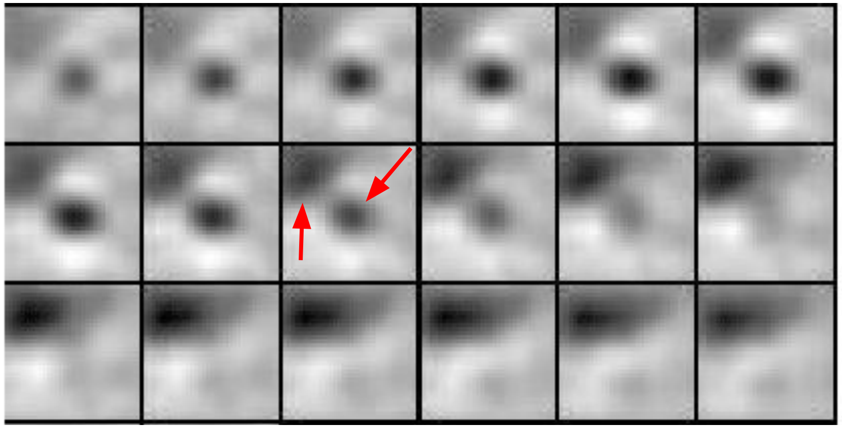
Surface patch



Large globule

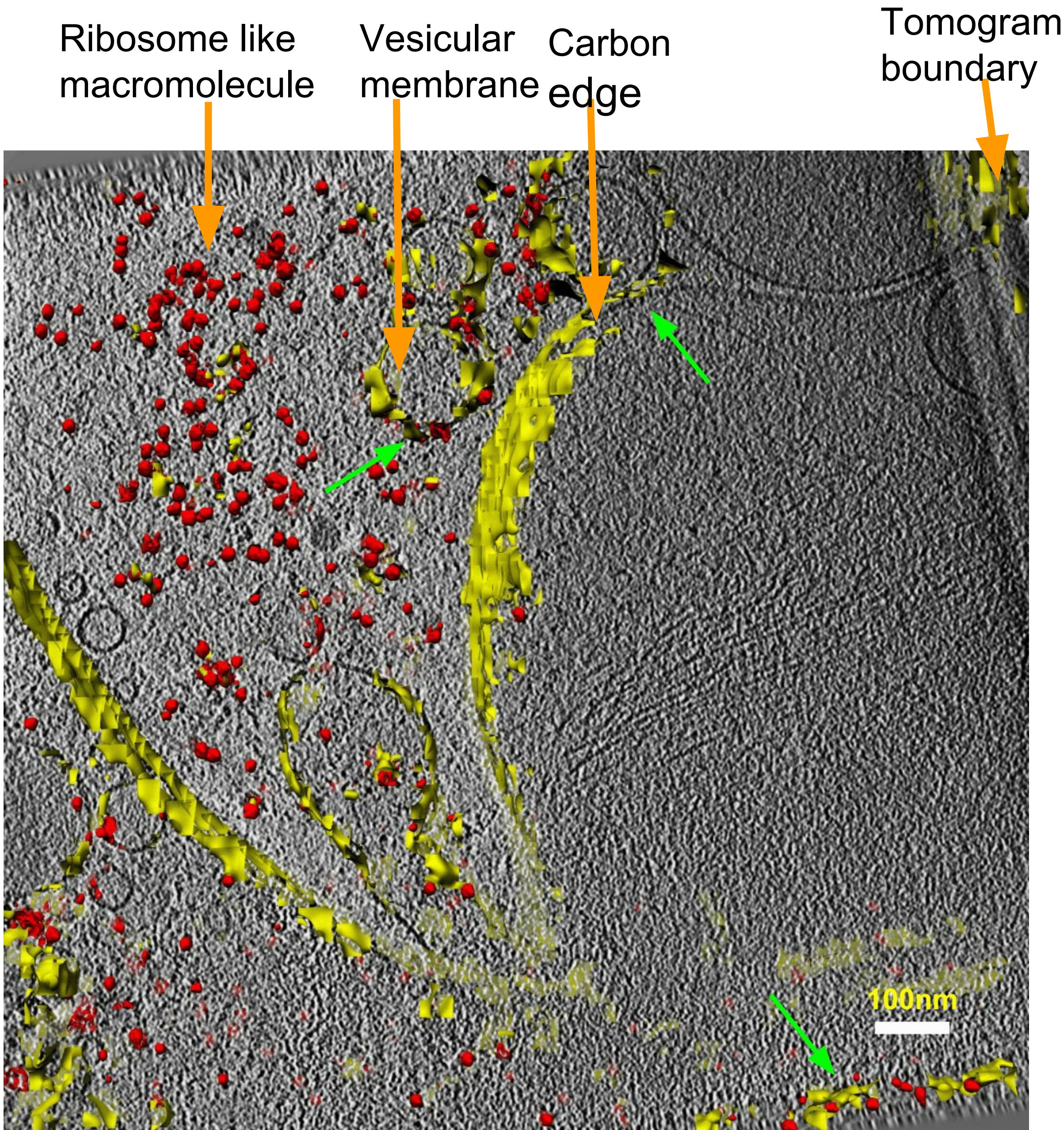


Small globule



Interaction between cellular components

Embedding of detected patterns



Subtomogram segmentation

Motivation: molecular crowding

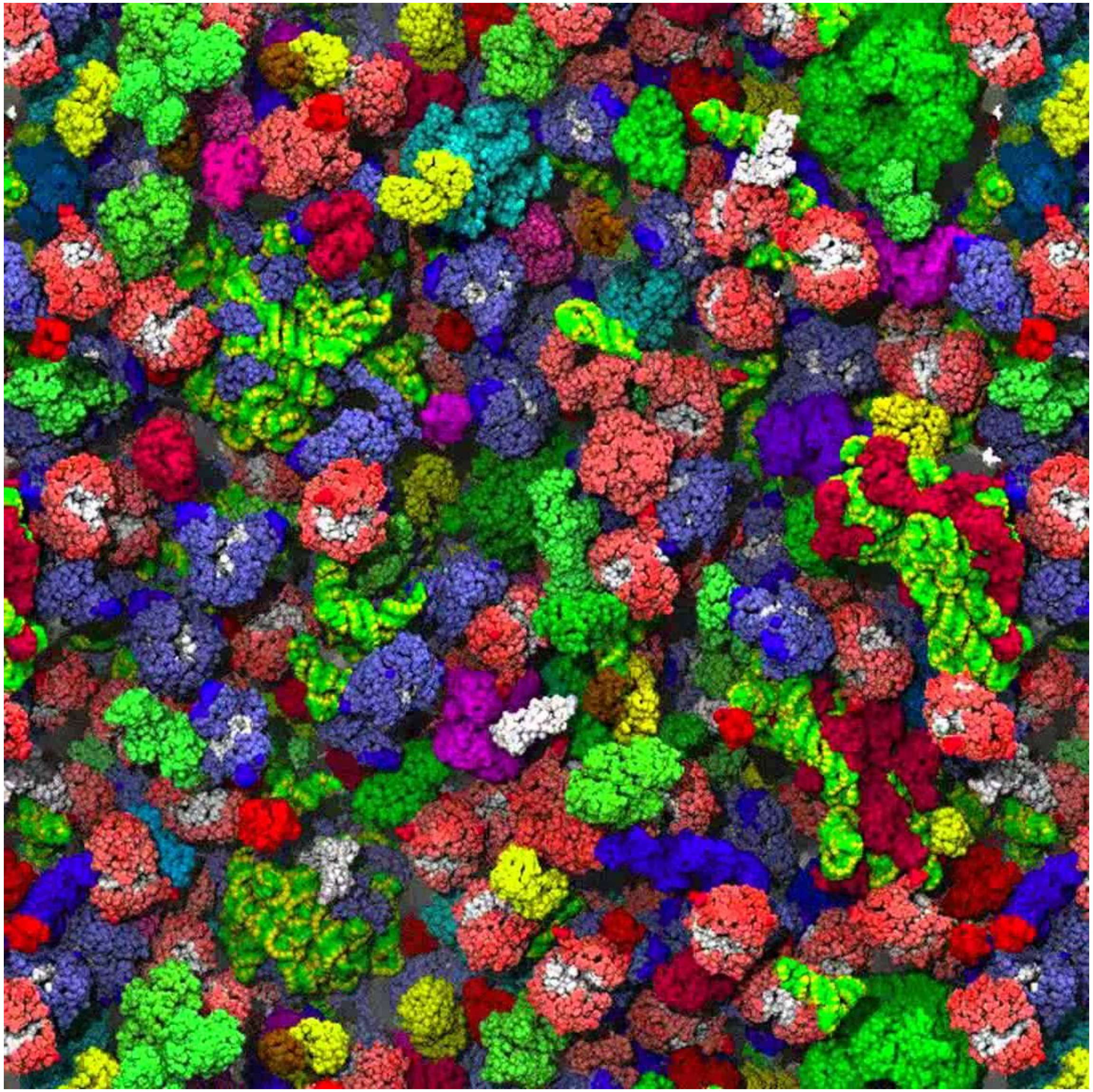
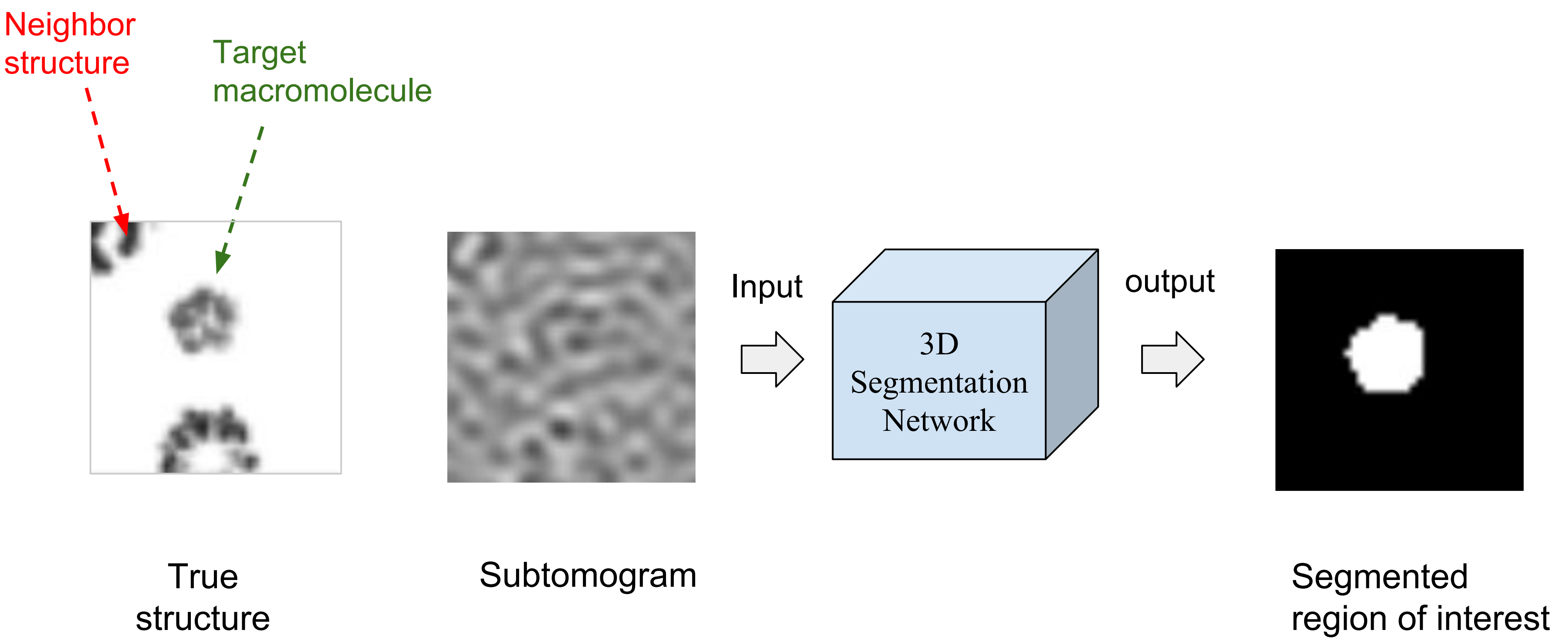


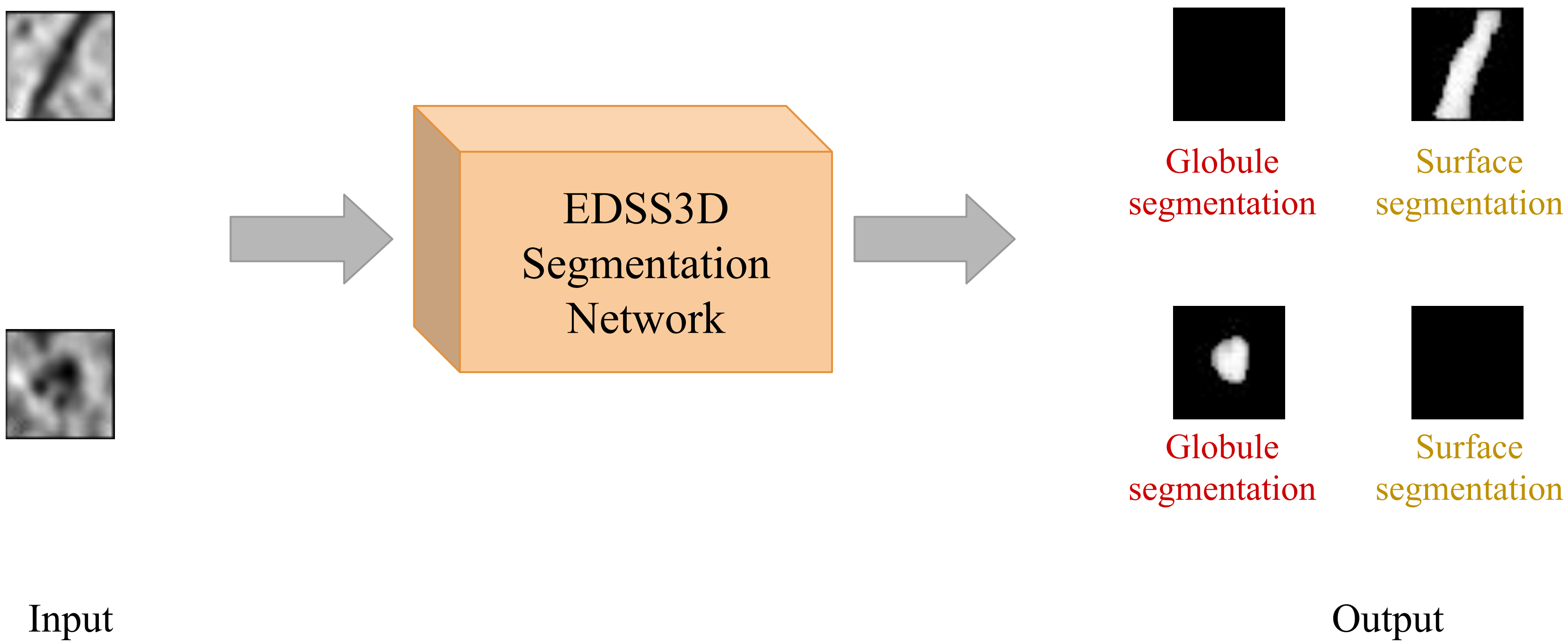
Image of simulated bacterial cytoplasm from McGuffee & Elcock, PLoS Comput Biol

Voxelwise binary classification based segmentation



Liu et al. arXiv:1802.04087

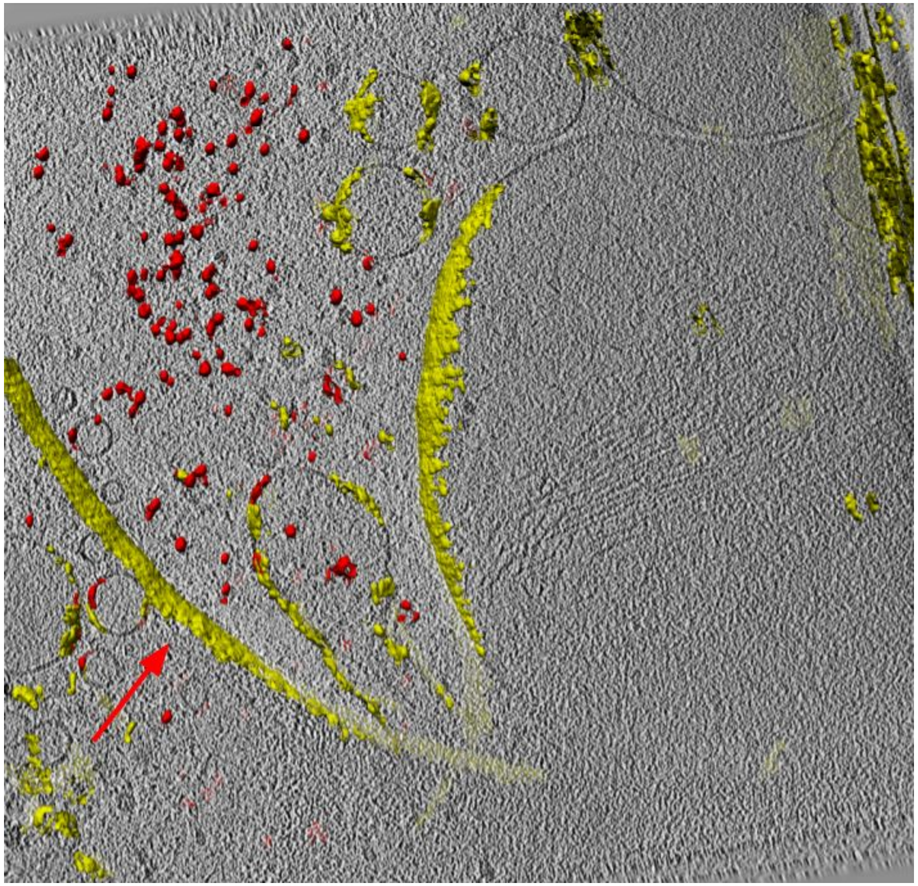
Voxelwise multiclass classification based segmentation



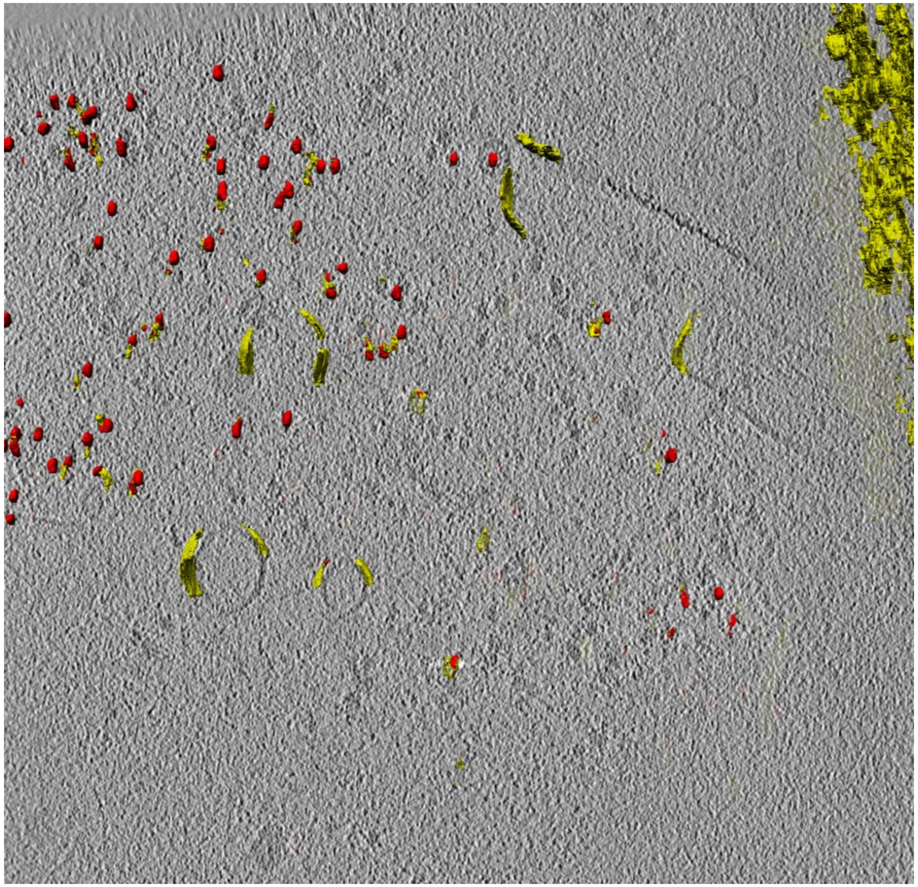
Zeng et al. JSB 2017

Weakly supervised segmentation

Training tomogram



Testing tomogram



Autoencoder training → Segmentor training → Segmenter prediction

Zeng et al. JSB 2017

Summary

- Convolutional neural networks are potentially powerful tools for structural pattern mining
- Substantial further works needed to make supervised deep learning practically useful
  - Construction of good training data
  - Optimization of network models
  - Reduction of supervision

Thank you

Funding support: NIH P41 GM103712, Samuel and Emma Winters Foundation