Image-derived hypothesis testing for CECT template matching

Min Xu, PhD
Computational Biology Department
Carnegie Mellon University

Wang, Kai Wen, et al. "Image-derived generative modeling of pseudo-macromolecular structures-towards the statistical assessment of Electron CryoTomography template matching." *BMVC Newcastle 2018*.

Application: Generating pseudo macromolecular structures

Wang et al. Image-derived generative modeling of pseudo-macromolecular structures - towards statistical assessment of electron cryotomography template matching. BMVC 2018

Outline

- Cellular Electron CryoTomography
- Template Matching
  - Problem: not statistically rigorous
- Shape space modeling
  - Approach 1: Large Deformable Diffeomorphic Metric Mapping (LDDMM)
  - Approach 2: 3D Generative Adversarial Nets
- Hypothesis test for template matching
  - Sampling away from macromolecular complex
  - Our paper’s results

Outline

- Cellular Electron CryoTomography
- Template Matching
  - Problem: not statistically rigorous
- Shape space modeling
  - Approach 1: Large Deformable Diffeomorphic Metric Mapping (LDDMM)
  - Approach 2: 3D Generative Adversarial Nets
- Hypothesis test for template matching
  - Sampling away from macromolecular complex
  - Our paper’s results

Cellular Electron CryoTomography

- Great precision for 3D imaging of macromolecules
  - Submolecular resolution
  - Minimal disturbance

- Important applications in biomedical sciences
  - Viewing and studying structures of macromolecules

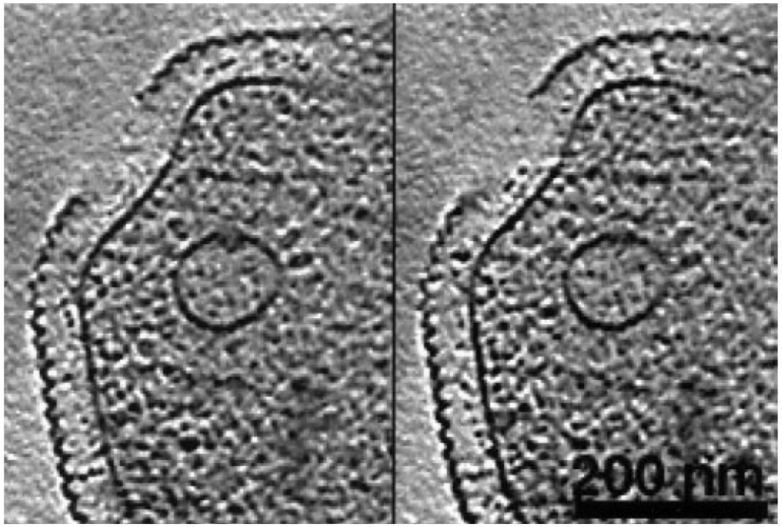


Fig 1. CECT image of Pyrodictium abyssi cell. Vesicle is clearly visible. [1]

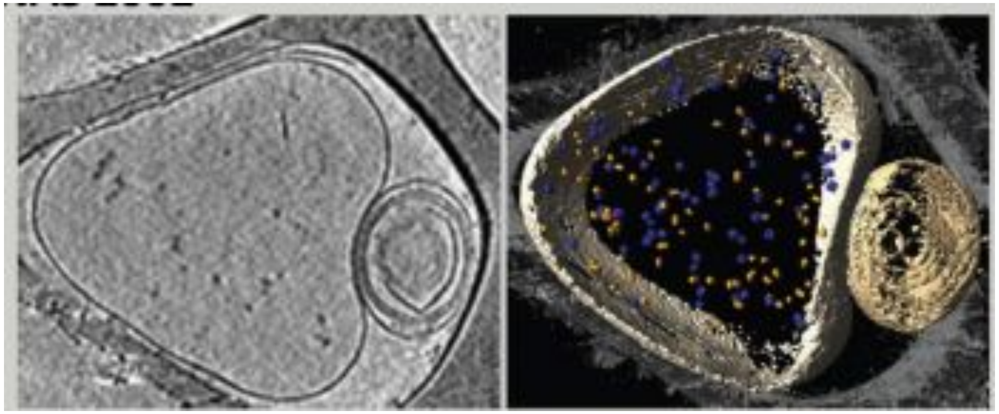
[1] Jochen Böhm, Achilleas S Frangakis, Reiner Hegerl, Stephan Nickell, Dieter Typke, and Wolfgang Baumeister. Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proceedings of the National Academy of Sciences*, 97(26):14245–14250, 2000.

Template matching

- *De facto* method for locating known macromolecules in tomograms
  - Low signal-to-noise (SNR), missing wedge. Visual inspection impractical!
  - Like a very hard “Where’s Waldo” for macromolecular structures!

- A subtomogram is a subvolume of a tomogram containing a macromolecule

- Calculates *relative* similarity between a subtomogram and a template
  - Rotate subtomogram to be most aligned with template
  - Calculate Pearson correlation (compensate for missing wedge effect from CECT!)



Mahamid, Julia, and Wolfgang Baumeister. "Cryo-electron tomography: the realization of a vision." *Microscopy and Analysis* 26.6 (2012): 45-48.

Problem with template matching

- Calculates *relative!* similarity between a subtomogram and a template
    - Using a hard threshold is not statistically rigorous

- Hypothesis testing can provide statistical credibility
    - Instead of thresholding, calculate an empirical p-value
    - We can be confident if p-value is small

- Where can we get random macromolecules for the hypothesis test?

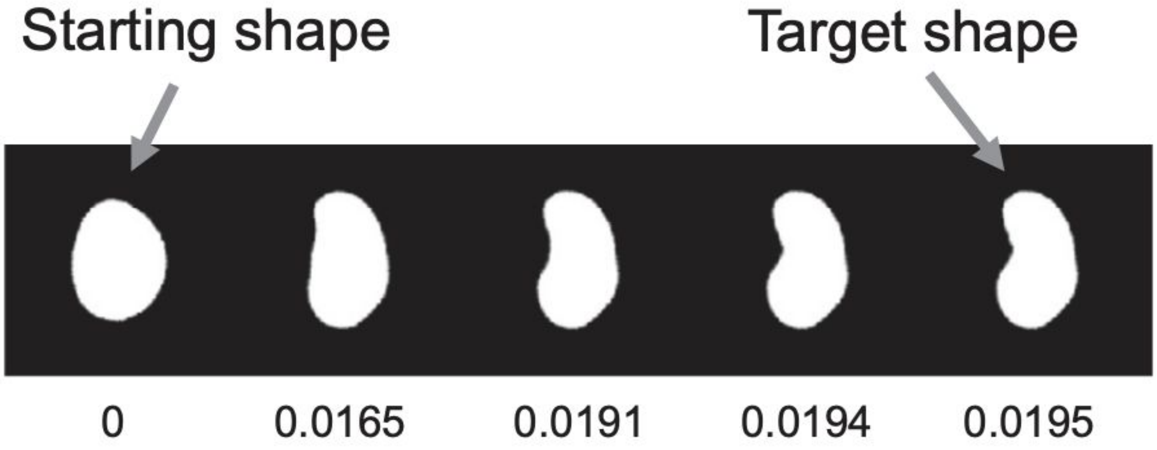
Outline

- Cellular Electron CryoTomography
- Template Matching
    - Problem: not statistically rigorous
- Shape space modeling
    - Approach 1: Large Deformable Diffeomorphic Metric Mapping (LDDMM)
    - Approach 2: 3D Generative Adversarial Nets
- Hypothesis test for template matching
    - Sampling away from macromolecular complex
    - Our paper’s results

Approach 1: Large Deformation Diffeomorphic Metric Mapping (LDDMM)

- Calculates “geodesic distance” between two shapes by:
  - Gradually morphing one into the other
  - Measuring the “change” needed along the way

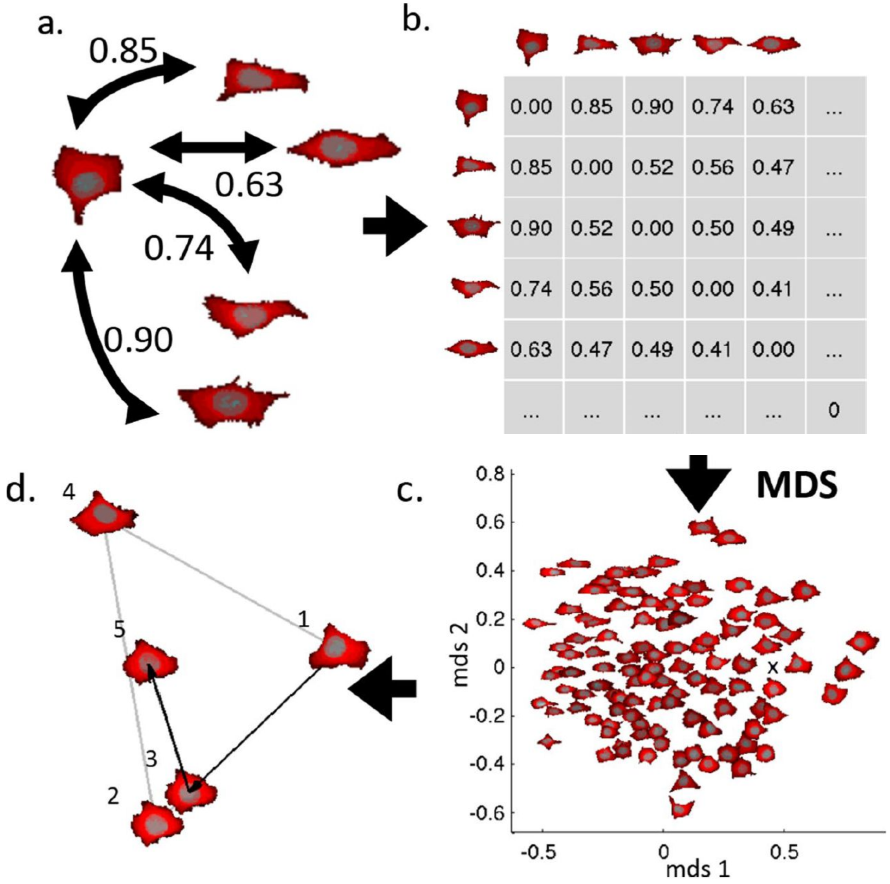
- Once one shape is completely morphed into the other, the sum of changes represents the distance



[3] Robert F Murphy. Building cell models and simulations from microscope images. *Methods*, 96:33–39, 2016.

LDDMM shape space modeling

- From given structures, construct distance matrix
- Multidimensional scaling (MDS) projects onto 2D plane
- Interpolate new shape by morphing from three shapes



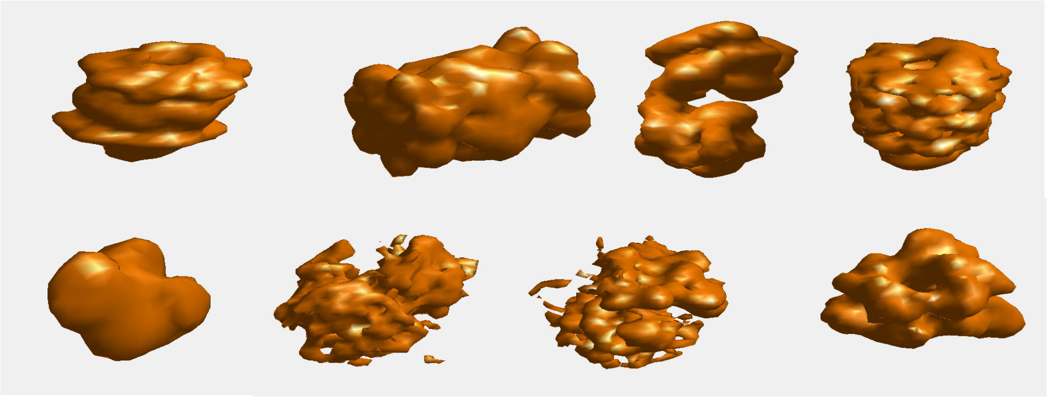
0.00	0.85	0.90	0.74	0.63	...
0.85	0.00	0.52	0.56	0.47	...
0.90	0.52	0.00	0.50	0.49	...
0.74	0.56	0.50	0.00	0.41	...
0.63	0.47	0.49	0.41	0.00	...
...	...	...	...	...	0

[3] Robert F Murphy. Building cell models and simulations from microscope images. *Methods*, 96:33–39, 2016.

LDDMM Results

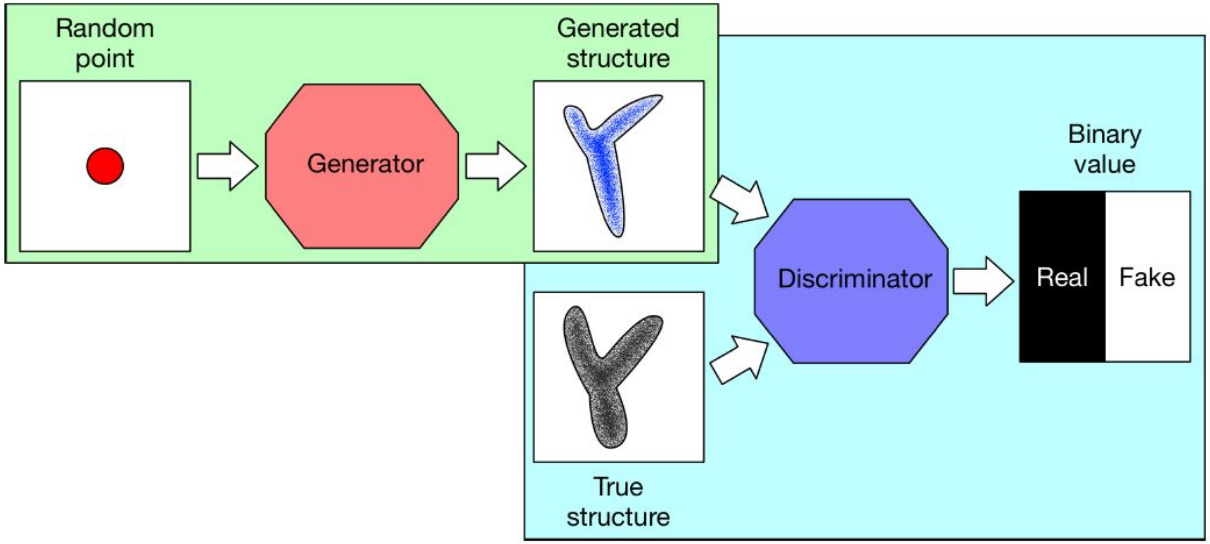
- Two problems...

- Shapes are often similar to originals
- Too computationally expensive!
  - Must iterative deform twice for each new shape
  - Not optimized for GPU!
  - O(n^2) full deformations for distance matrix



Approach 2: Generative Adversarial Networks (GAN)

- Deep unsupervised learning approach for generating images and shapes
  - Minimax game between two neural networks: generator and discriminator
  - Discriminator want to correctly classify whether an image is fake
    - minimize binary cross entropy loss of classification
  - Generator wants to “fool” discriminator

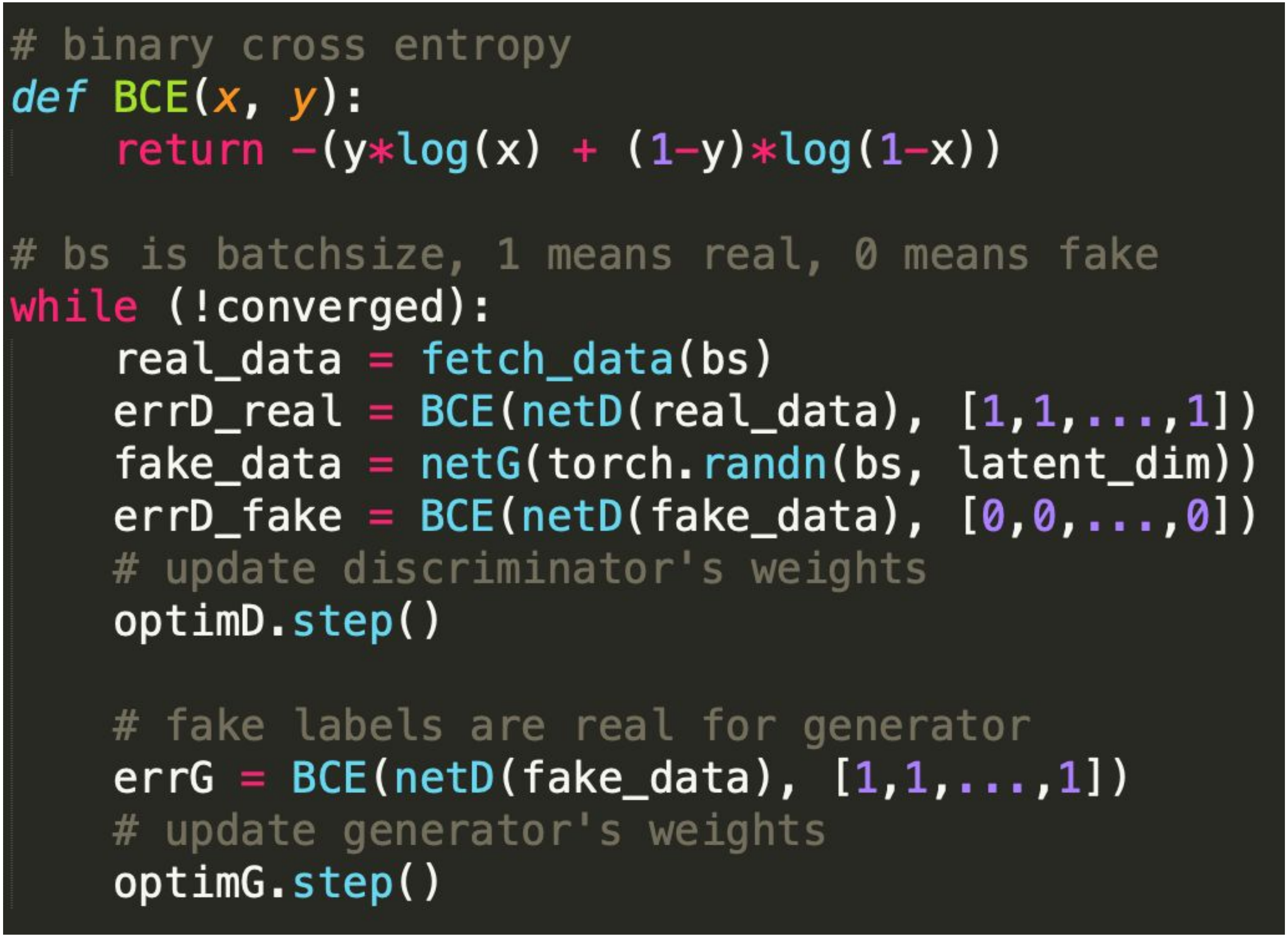


$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

Wang, Kai Wen, et al. "Image-derived generative modeling of pseudo-macromolecular structures-towards the statistical assessment of Electron CryoTomography template matching." *BMVC Newcastle 2018*.

Training loop of the GAN

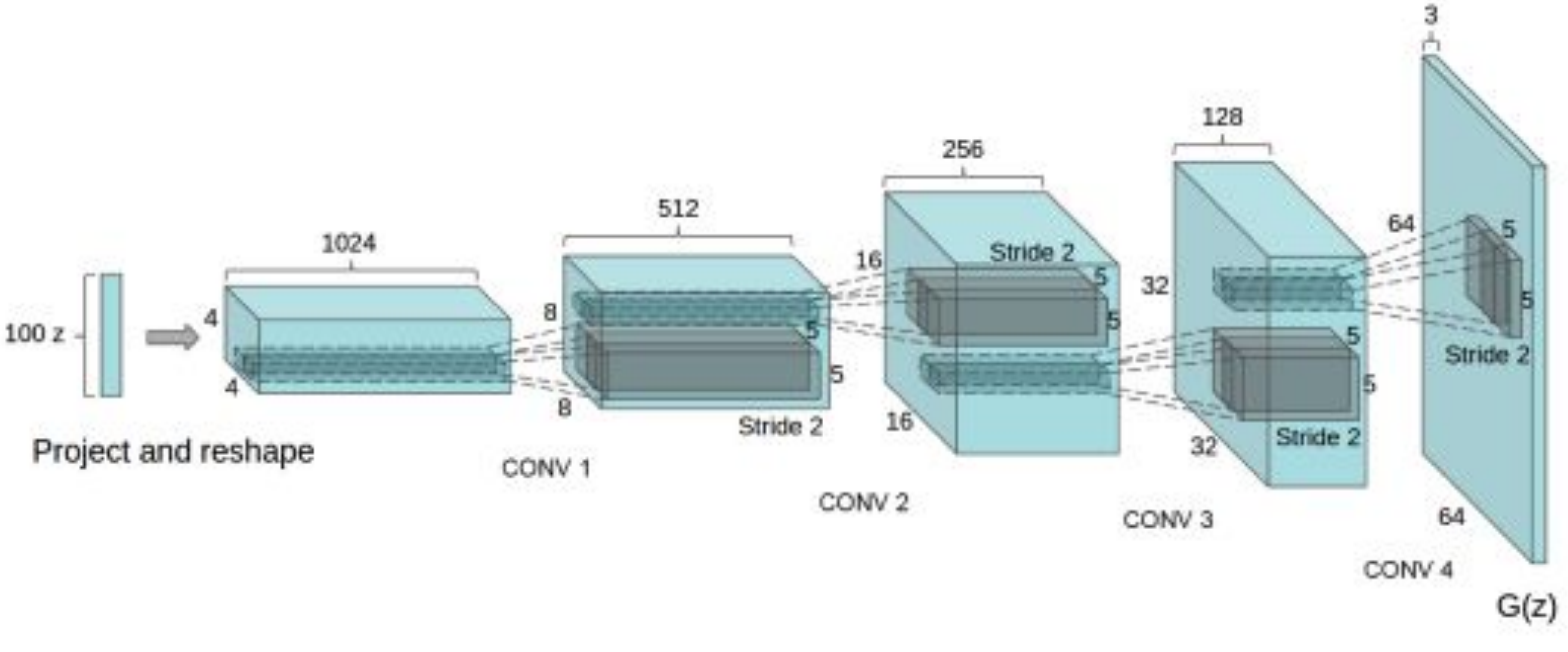


```
# binary cross entropy
def BCE(x, y):
    return -(y*log(x) + (1-y)*log(1-x))

# bs is batchsize, 1 means real, 0 means fake
while (!converged):
    real_data = fetch_data(bs)
    errD_real = BCE(netD(real_data), [1,1,...,1])
    fake_data = netG(torch.randn(bs, latent_dim))
    errD_fake = BCE(netD(fake_data), [0,0,...,0])
    # update discriminator's weights
    optimD.step()

    # fake labels are real for generator
    errG = BCE(netD(fake_data), [1,1,...,1])
    # update generator's weights
    optimG.step()
```

Neural network architectures

- Both generator and discriminator are convolutional neural networks

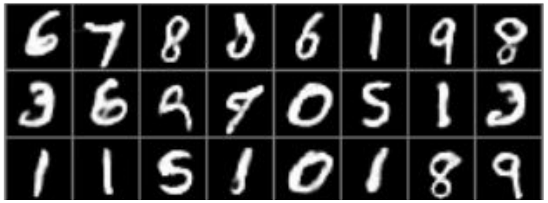
- Generator and discriminator are “mirror images”



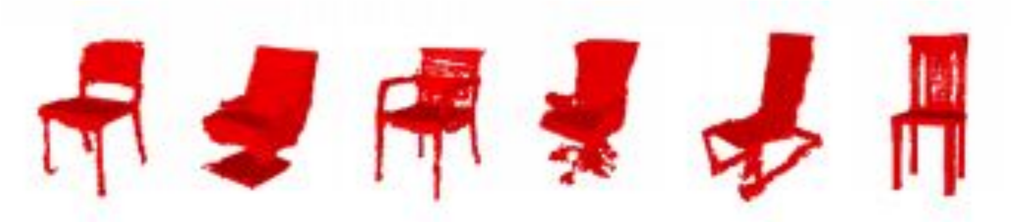
Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).

GAN results

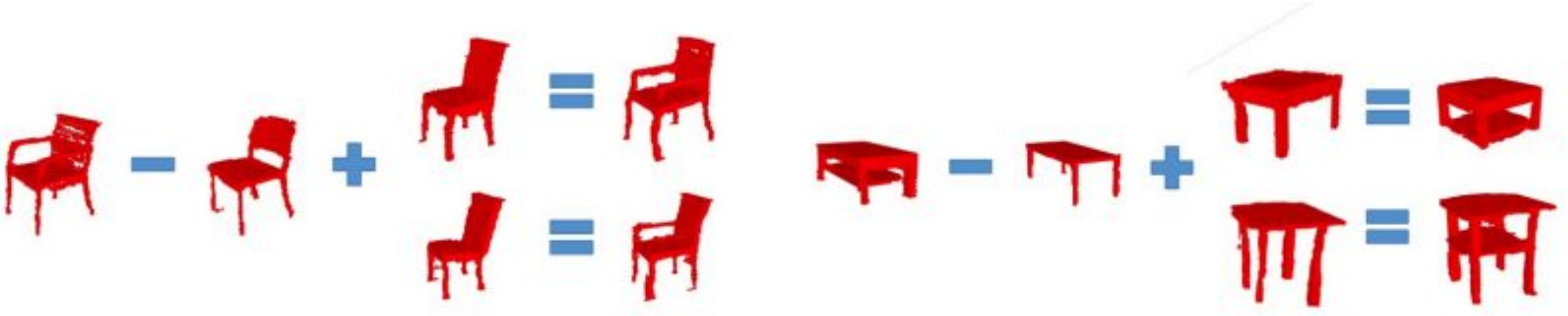
- Trained on MNIST dataset



- Extending GAN to 3D shapes

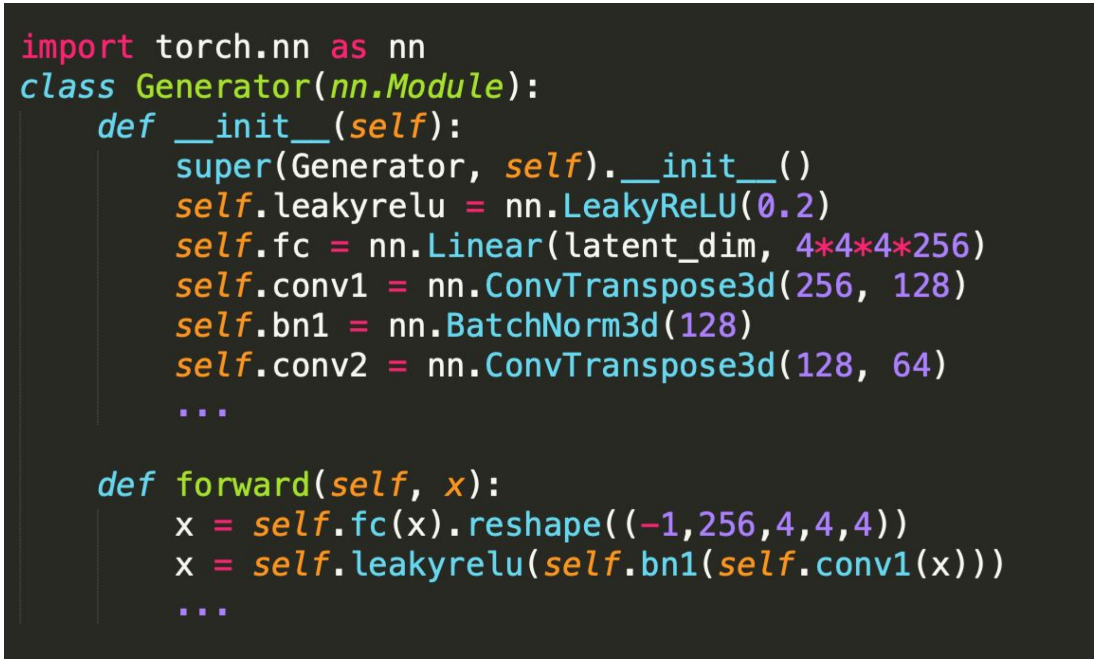


- Latent space arithmetic

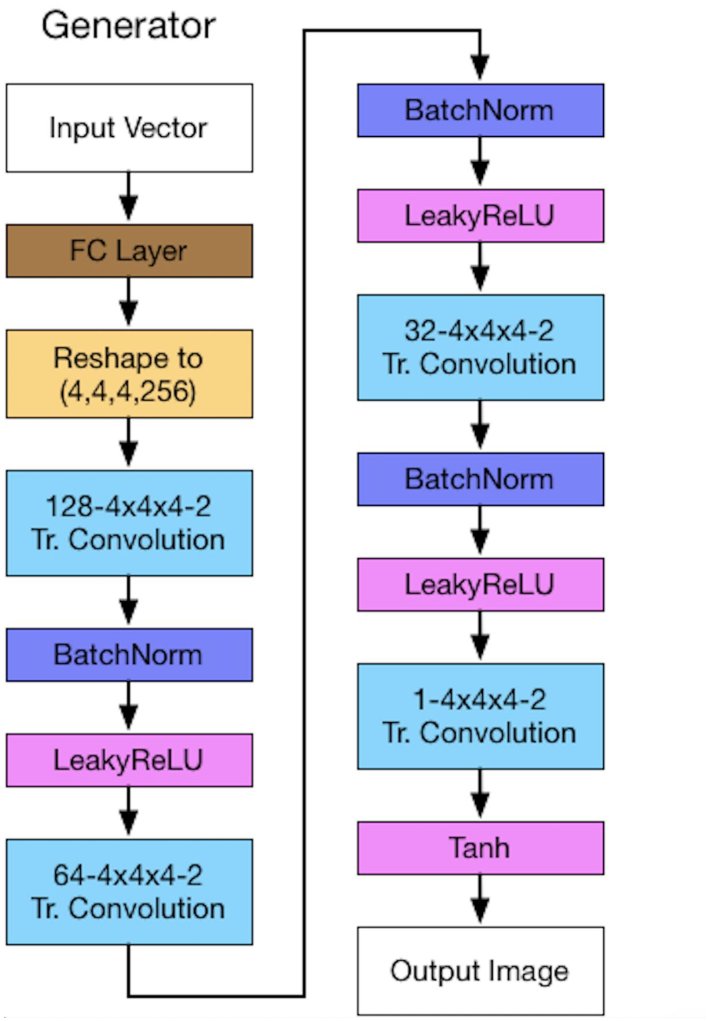


Wu, Jiajun, et al. "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling." *Advances in neural information processing systems*. 2016.

Our model architecture



```python
import torch.nn as nn
class Generator(nn.Module):
    def __init__(self):
        super(Generator, self).__init__()
        self.leakyrelu = nn.LeakyReLU(0.2)
        self.fc = nn.Linear(latent_dim, 4*4*4*256)
        self.conv1 = nn.ConvTranspose3d(256, 128)
        self.bn1 = nn.BatchNorm3d(128)
        self.conv2 = nn.ConvTranspose3d(128, 64)
        ...

    def forward(self, x):
        x = self.fc(x).reshape((-1,256,4,4,4))
        x = self.leakyrelu(self.bn1(self.conv1(x)))
        ...
```



Generator

Input Vector

FC Layer

Reshape to
(4,4,4,256)

128-4x4x4-2
Tr. Convolution

BatchNorm

LeakyReLU

64-4x4x4-2
Tr. Convolution

BatchNorm

LeakyReLU

32-4x4x4-2
Tr. Convolution

BatchNorm

LeakyReLU

1-4x4x4-2
Tr. Convolution

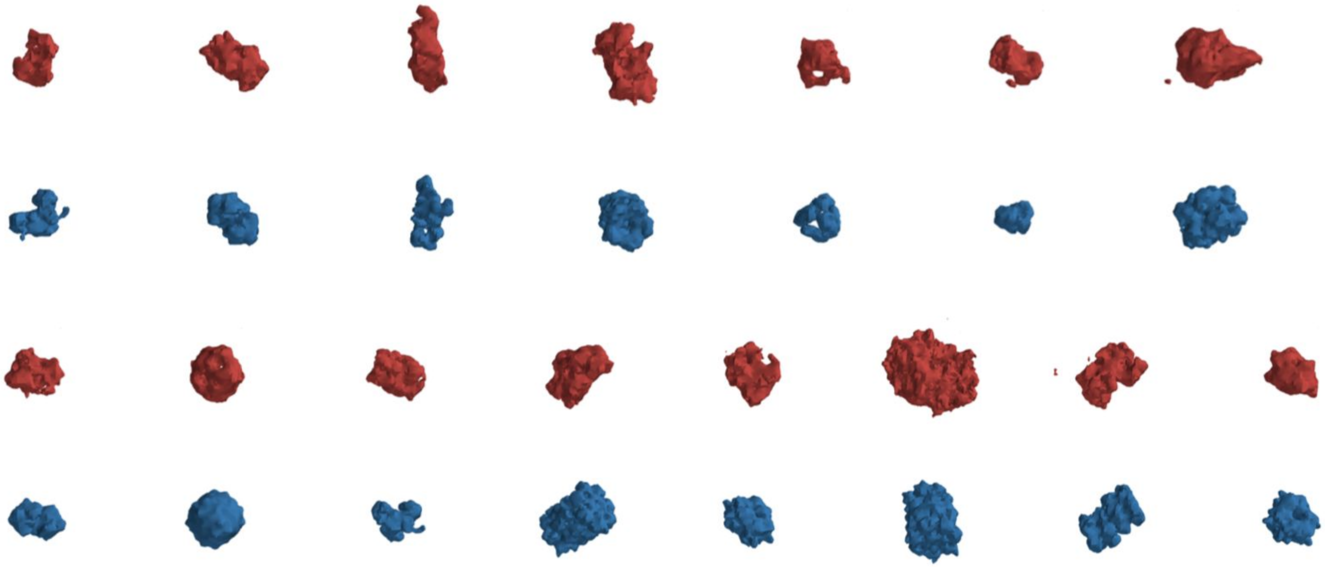
Tanh

Output Image

GAN for macromolecular complexes

- Our paper* uses the 3DGAN with a few modifications
    - Training set of 15 experimental macromolecular complexes (64^3)
    - Rotated 600 times for total training set of 9000 structures.

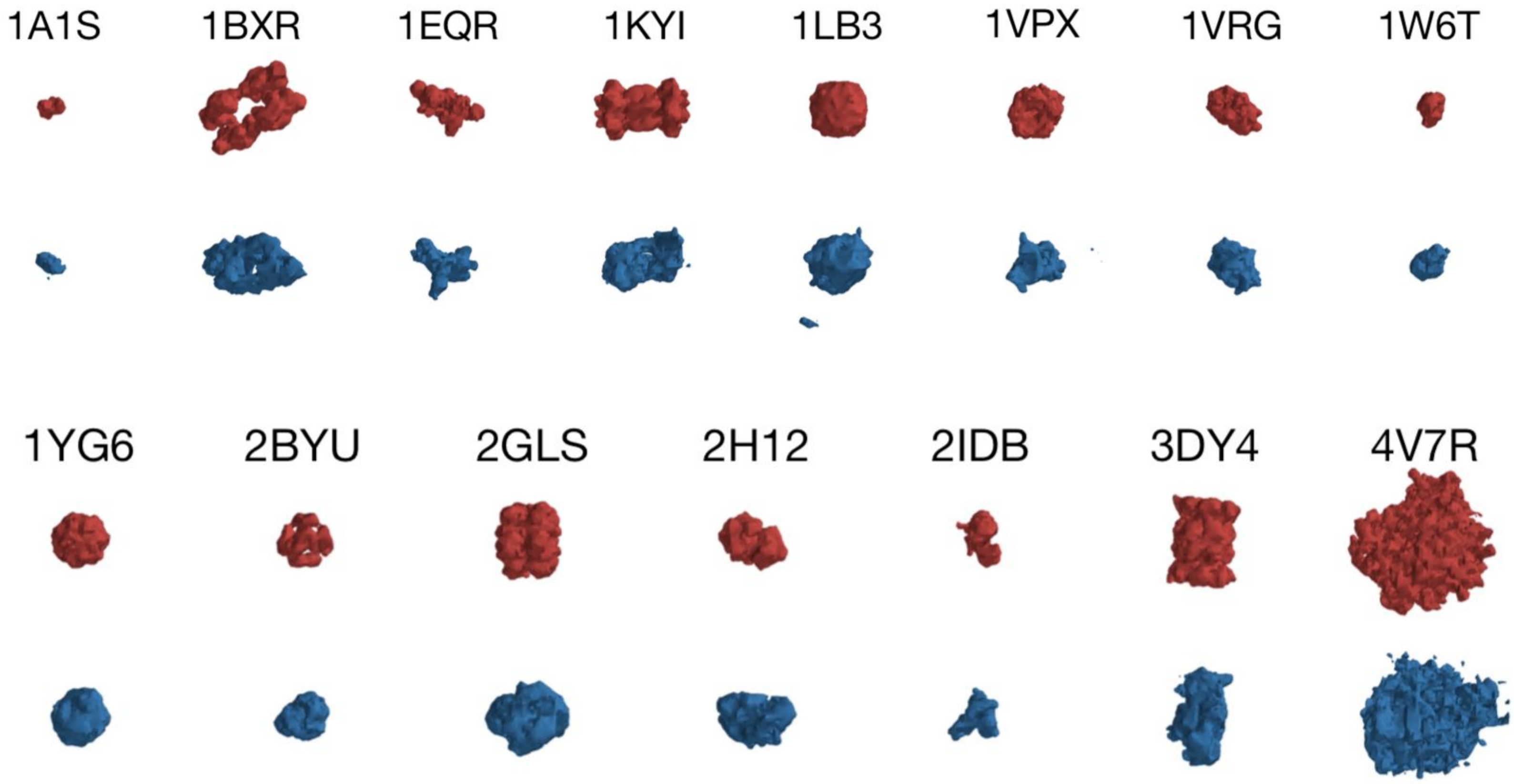
- The generated shapes are reasonably similar to training shapes
    - Red: generated, blue: ground truth



*Wang, Kai Wen, et al. "Image-derived generative modeling of pseudo-macromolecular structures-towards the statistical assessment of Electron CryoTomography template matching." *BMVC Newcastle 2018*.

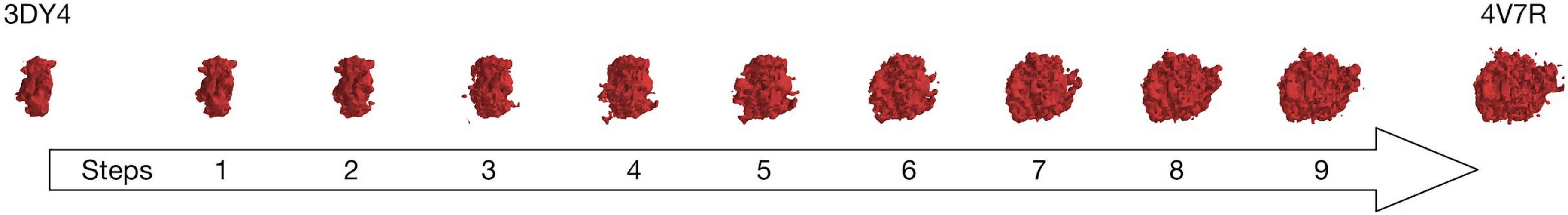
Results (cont.)

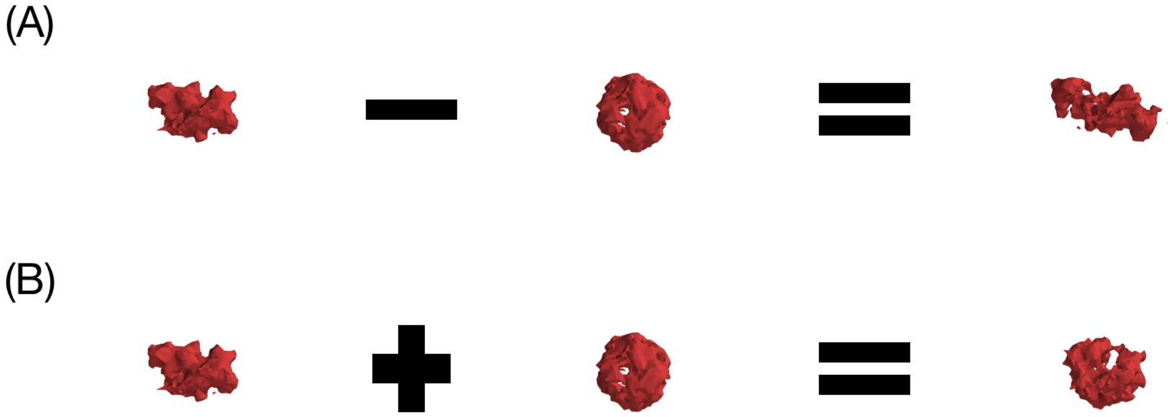
- Find closest generated shape to each training structure
  - Red: ground truth, blue: generated



Shape space manifold

- Linear steps in the latent 100 dimensional shape space
- Much more faster than LDDMM since we can use GPUs!





Outline

- Cellular Electron CryoTomography
- Template Matching
  - Problem: not statistically rigorous
- Shape space modeling
  - Approach 1: Large Deformable Diffeomorphic Metric Mapping (LDDMM)
  - Approach 2: 3D Generative Adversarial Nets
- Hypothesis test for template matching
  - Sampling away from macromolecular complex
  - Our paper’s results

Back to template matching

- Hypothesis test for subtomogram P containing complex C
  - C is the complex with the largest cross-correlation amongst the 15 known complexes

- Setup
  - Test statistic: cross correlation scores for a fix template and random subtomograms
  - Null hypothesis: P doesn’t contain C
  - Alternative hypothesis: P contains C

$$H_0 : P \neq T(C_I) \qquad\qquad H_A : P = T(C_I)$$

- Goal: derive Monte Carlo empirical distribution of test statistic under the null
  - By law of large numbers, empirical estimate converges to true p-value as number of samples grows!

Sampling away from C

- Under the null hypothesis (P doesn’t contain C), C should not be sampled from the distribution of random complexes
  - Otherwise, they could be viewed as copies of C in the hypothesis test

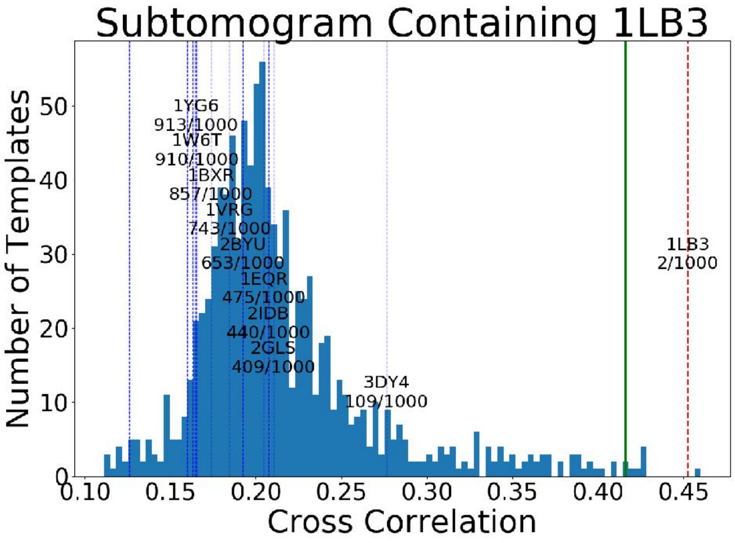
- We skew the learned shape space to avoid C
  - Learn the latent distribution E of complex C
  - Construct rejection region to reject the following points:

$$P \in \mathbb{R}^{100}, N(P) < \pi \cdot \mathcal{E}(P)$$

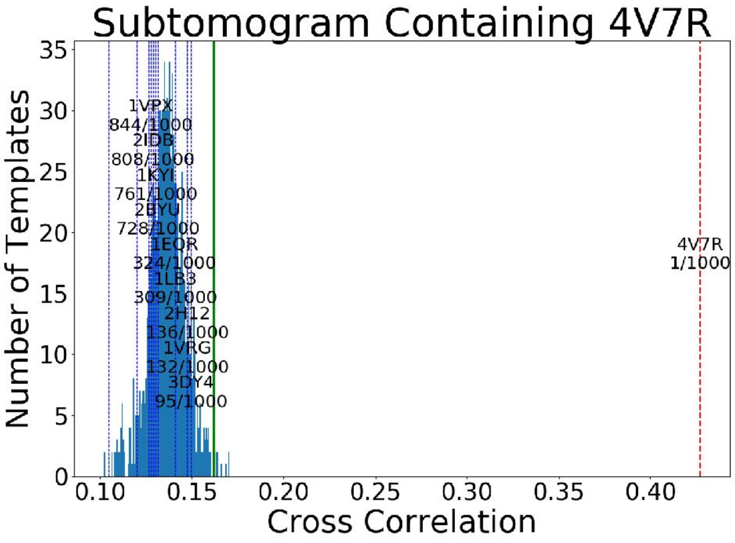
where N is 100-dimensional multivariate Gaussian,
pi is the prior (1 / num classes)

Hypothesis tests

- Results for complexes 1LB3 and 4V7R
  - Indeed, the correlation is higher than the scores from the other subtomograms!



Subtomogram Containing 1LB3



Subtomogram Containing 4V7R

- The hypothesis test eliminated 40% of false positives

Summary

- Cryo-ET is powerful but not perfect.
- Image analysis must be automated, but previous attempts with template matching were not statistically rigorous.
    - Cross-correlation is a relative measure of structural similarity
    - Used hardcoded thresholds
- Hypothesis testing can fix this.
    - But where to get all the random molecules?
- Model shape space from a small set of known complexes
    - Approach 1: Large Deformable Diffeomorphic Metric Mapping (LDDMM)
        - Too computationally expensive!
    - Approach 2: 3D Generative Adversarial Nets
        - Can be sped up with GPUs.