Carnegie Mellon University

Image-derived generative modeling of pseudo-macromolecular structures — towards the statistical assessment of Electron CryoTomography template matching

Kai Wen Wang[1], Xiangrui Zeng[1], Xiaodan Liang[1], Zhiguang Huo[2], Eric P. Xing[1], Min Xu[1]

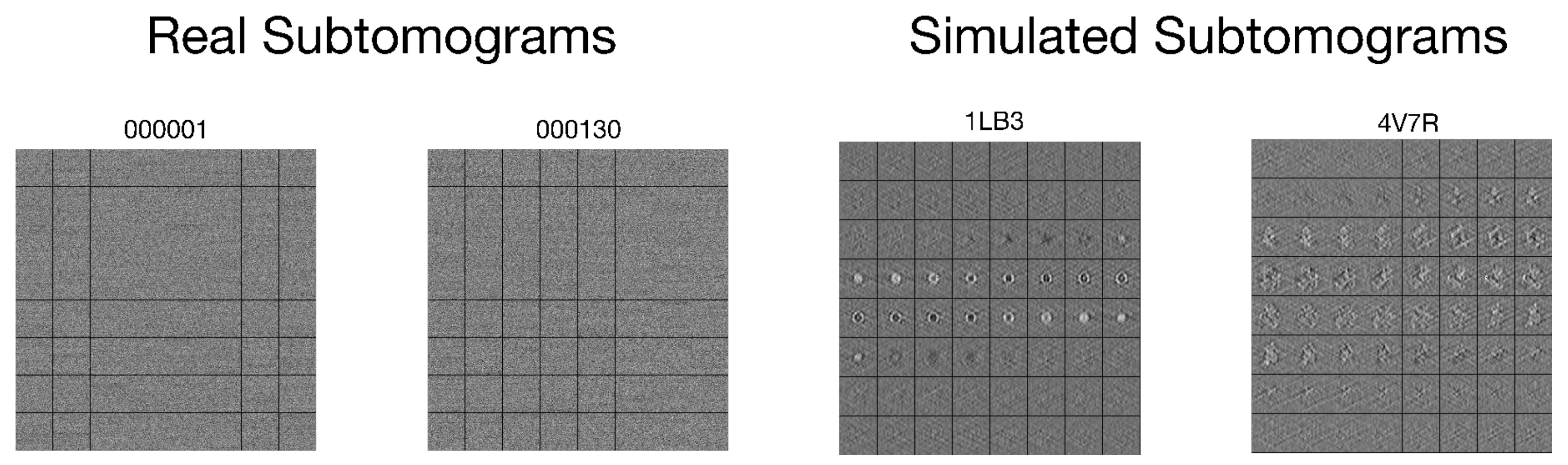
[1]Carnegie Mellon University, [2]University of Florida

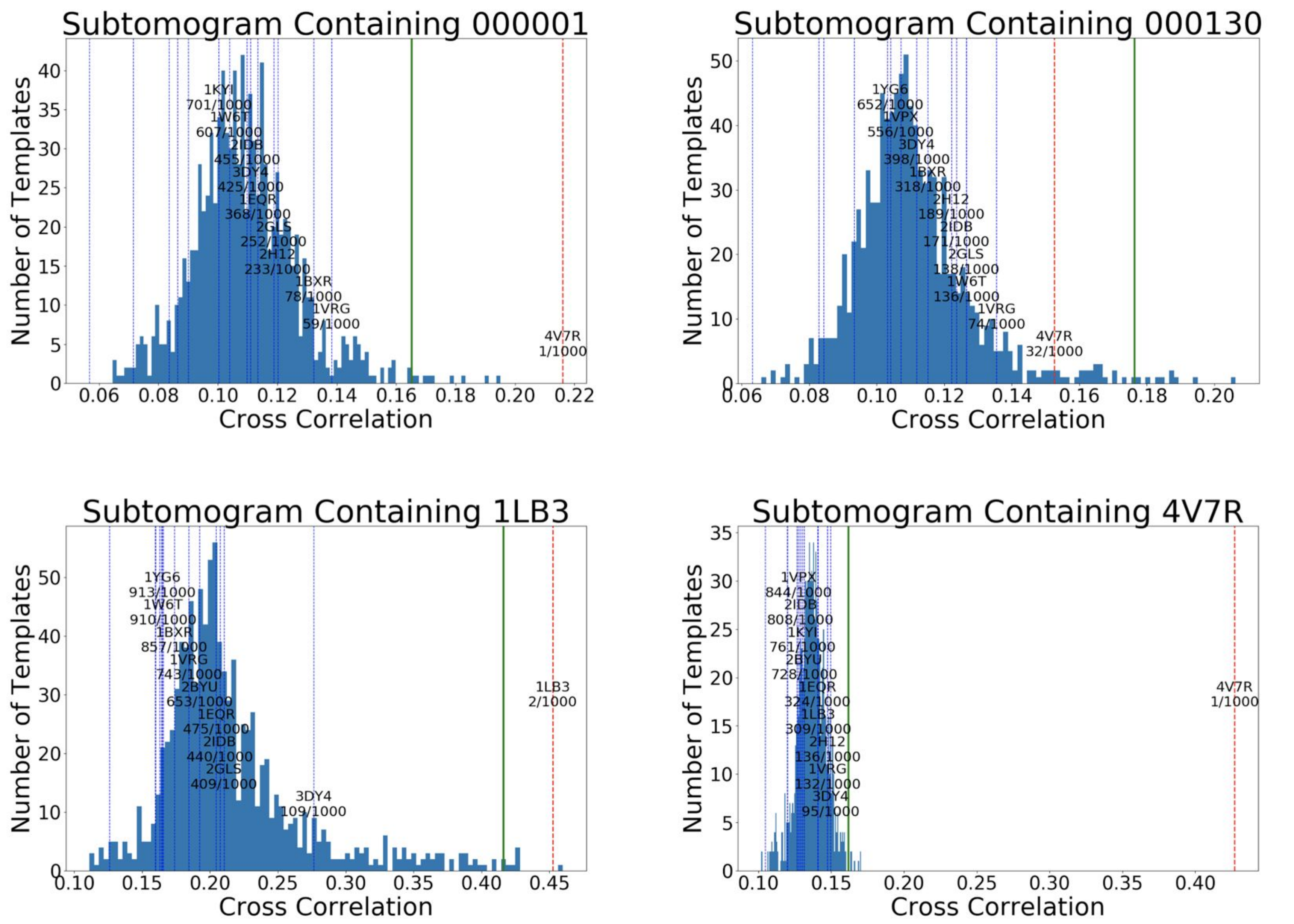
INTRODUCTION

- Cellular Electron CryoTomography (CECT) is a 3D imaging technique for macromolecular complexes.
- Low signal-to-noise ratio and missing wedge makes analysis very difficult.
- A subtomogram is a cubic sub-volume of an image captured by CECT containing a macromolecular complex.
- Given subtomogram P and template T, template matching with Pearson correlation c(P,T) is often used to locate macromolecules in a CECT image, but is insufficient since c(P,T) only measures relative structural similarity.
- Our research introduces and validates a novel, Monte Carlo approach for statistically assessing template matching through hypothesis testing to calculate empirical p-values.

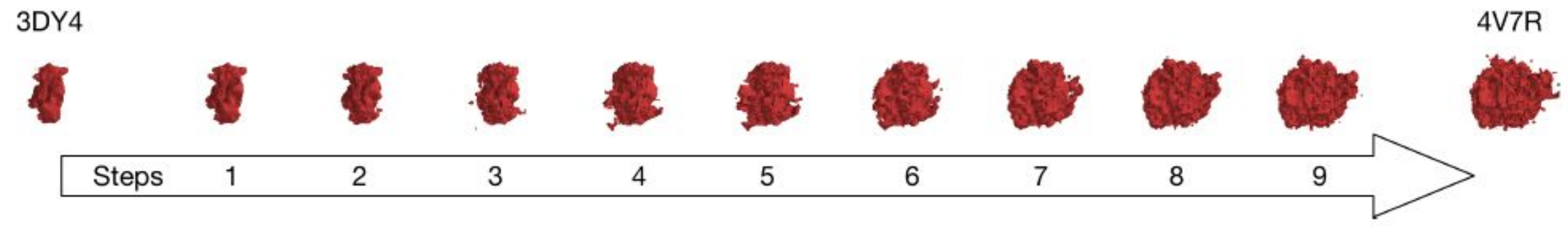
EXPERIMENTS AND RESULTS

- 3D-WGAN trained on 15 unique complexes * 600 rotations per complex for a total of 9000 3D gray-scale images.
- 376 real subtomograms containing ribosome with 70.21% success.
- 15 simulated subtomograms (for each unique complex) with 80% success.
- The blue histograms model the distribution of cross correlation scores of pseudo-complexes with the given subtomogram P.





- 3D-WGAN also learned manifold of macromolecules:



- 3D-WGAN could serve as relatively efficient smooth deformable registration, as opposed to existing methods like Large Deformation Diffeomorphic Metric Mapping, which is computationally expensive.

3-STEP HYPOTHESIS TESTING PROCEDURE

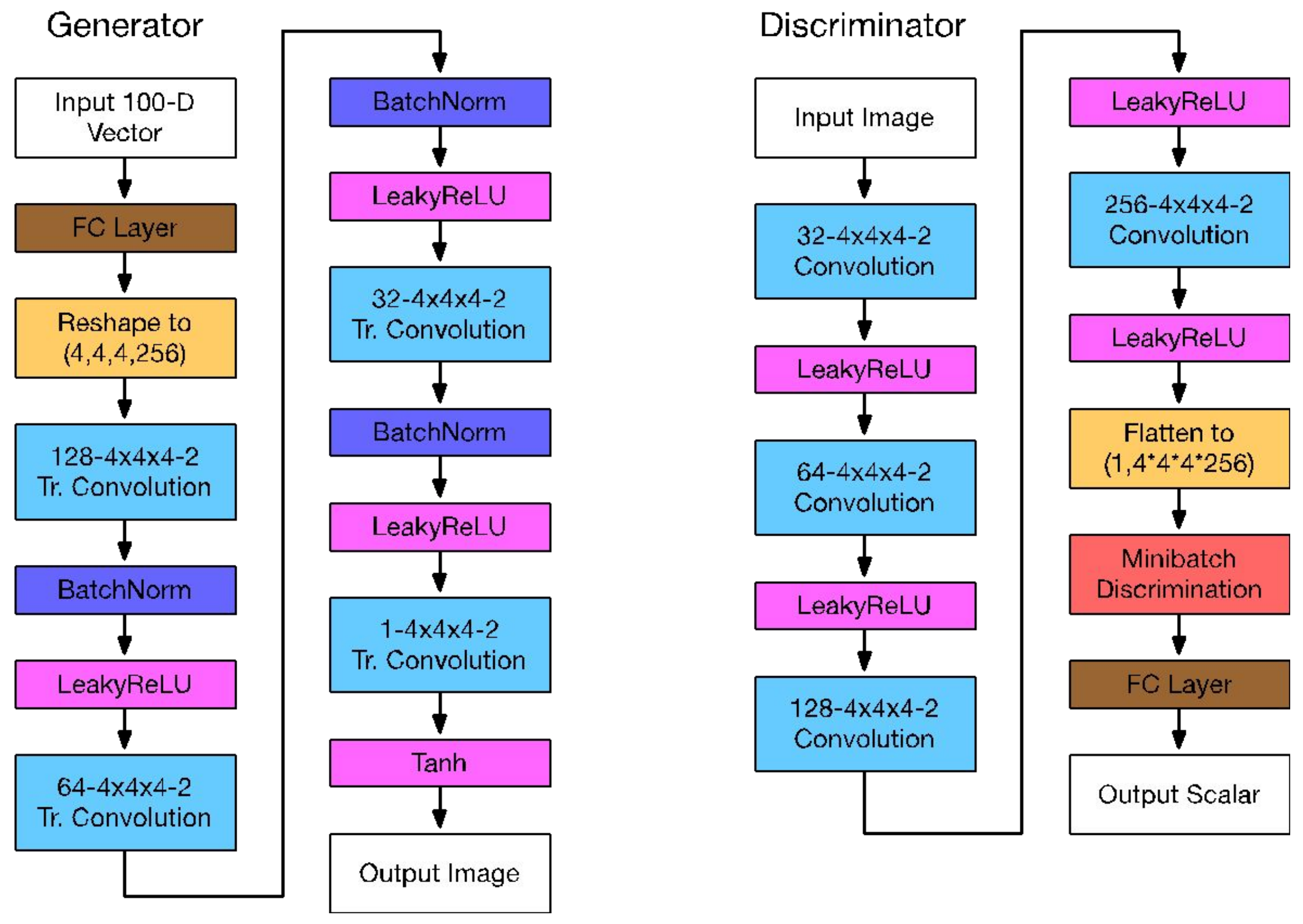
1. Train a 3D-WGAN to learn the structural distribution of macromolecules.
2. Determine macromolecule of interest $C_{interest}$ as the known macromolecule with the largest correlation score with a subtomogram P.
3. Using pseudo-complexes sampled from 3D-WGAN and far away from $C_{interest}$, perform hypothesis test with null and alternative hypothesis:

$$H_0 : P \text{ does not contain a macromolecule identical to } C_{Interest}$$

$$H_A : P \text{ contains a macromolecule identical to } C_{Interest}$$

APPROACH

3D-WGAN Architecture:



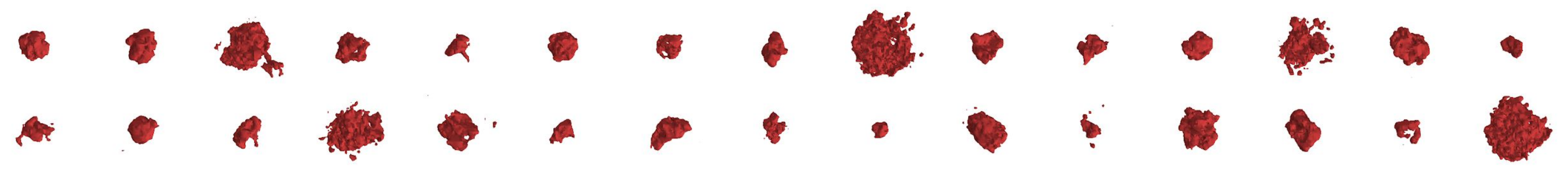
Generator

Input 100-D 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

Discriminator

Input Image → 32-4x4x4-2 Convolution → LeakyReLU → 64-4x4x4-2 Convolution → LeakyReLU → 128-4x4x4-2 Convolution → LeakyReLU → 256-4x4x4-2 Convolution → LeakyReLU → Flatten to (1,4*4*4*256) → Minibatch Discrimination → FC Layer → Output Scalar

Examples of pseudo-complexes generated using 3D-WGAN:



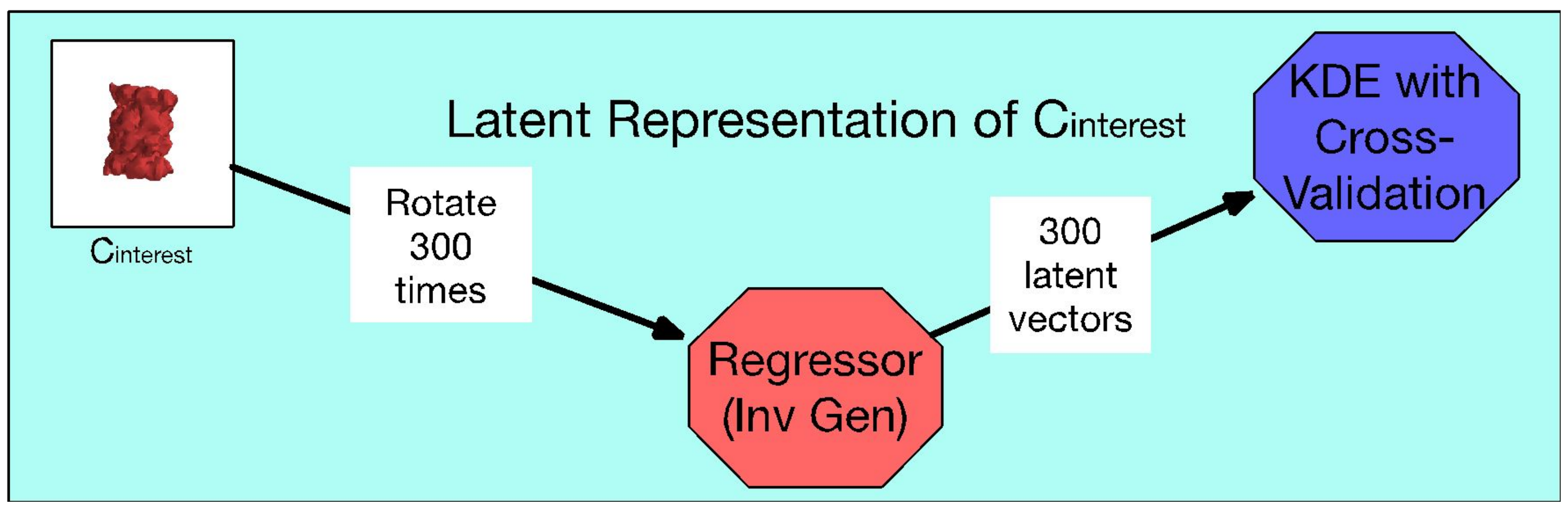
Sampling away from $C_{interest}$:

1. Determine latent representation of $C_{interest}$ as the distribution $\mathcal{E}$
2. Define rejection region for Bayes Classifier:

$$\mathcal{R} = \{G(v) : v \in \mathbb{R}^{100} \text{ such that } \mathcal{N}(v) < \pi \cdot \mathcal{E}(v)\}$$

Where $\mathcal{N}$ is the normal distribution, π is the prior.

Only pseudo-complexes outside of $\mathscr{K}$ is used for hypothesis test.



Latent Representation of $C_{interest}$

$C_{interest}$ → Rotate 300 times → Regressor (Inv Gen) → 300 latent vectors → KDE with Cross-Validation

Calculate empirical p-value that approaches true p-value almost surely

$$\begin{aligned} p &= \mathbb{E}_{H_0;C_0 \sim f_{structure}}[\mathbb{I}\{c(P,T(C_{Interest})) \leq c(P,T(C_0))\} | C_0 \notin \mathcal{R}] \\ &= \Pr(c(P,T(C_{Interest})) \leq c(P,T(C_0)) | C_0 \notin \mathcal{R}) \end{aligned}$$

$$\hat{p} = B^{-1} \sum_{b=1}^{B} \mathbb{I}[c(P,T(C_{Interest})) \leq c(P,T(C_0^{(b)}))] \xrightarrow[B \to \infty]{a.s.} p$$