2017 International Conference on Tools with Artificial Intelligence (ICTAI)

Protein Loop Modeling Using Deep Generative Adversarial Network (GAN)

Zhaoyu Li, Son Nguyen, Dong Xu, Yi Shang

Department of Electrical Engineering and Computer Science (EECS)
University of Missouri - Columbia

Contents

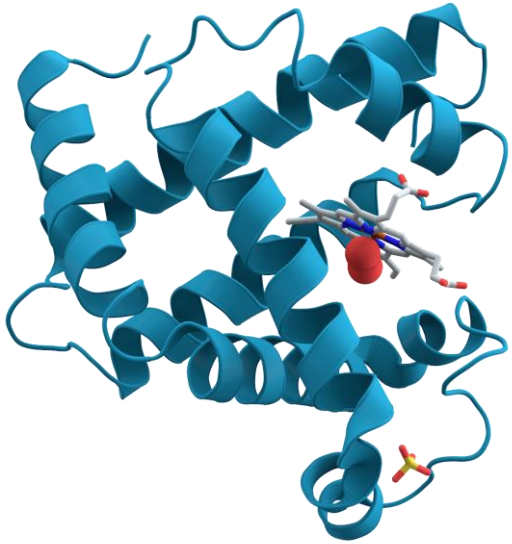
Introduction

Invitation to the world of Bioinformatics

- Proteins are most common molecules in cells
- Important role in medicine and life science

- 20 amino acid types
- Amino acid sequence determines the 3D structure

- During this talk, your body will produce 10,000,000,000,000,000,000 protein molecules

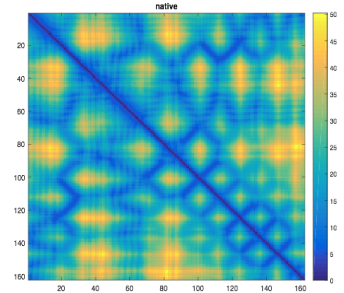


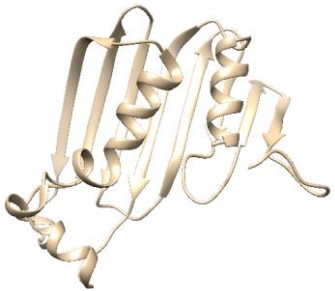
A visualization of the 3D structure of a protein

Background

Distance Map and Multidimensional Scaling (MDS)

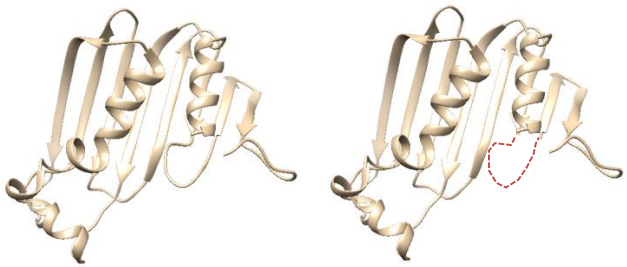
- Each amino acid has a coordinate
- Distance map: Euclidean distance of $C\alpha$ atoms of all amino acid pairs
  - From 3D to 2D
  - Can be treated as an image
  - Orientation independent
- MDS can restore 3D space while preserving the distance constraints in the 2D space





Protein Loop Modeling

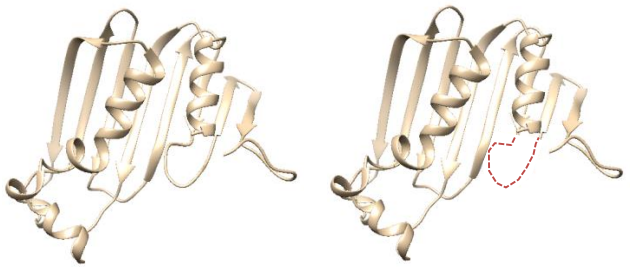
- Small parts in a protein may be missing
- Loop modeling is to predict those missing regions

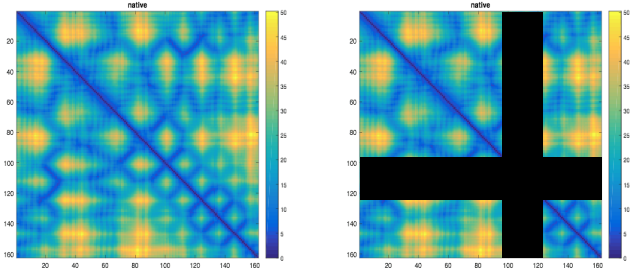


Protein Loop Modeling

- Small parts in a protein may be missing
- Loop modeling is to predict those missing regions

- We solve this problem in 2D space using distance map



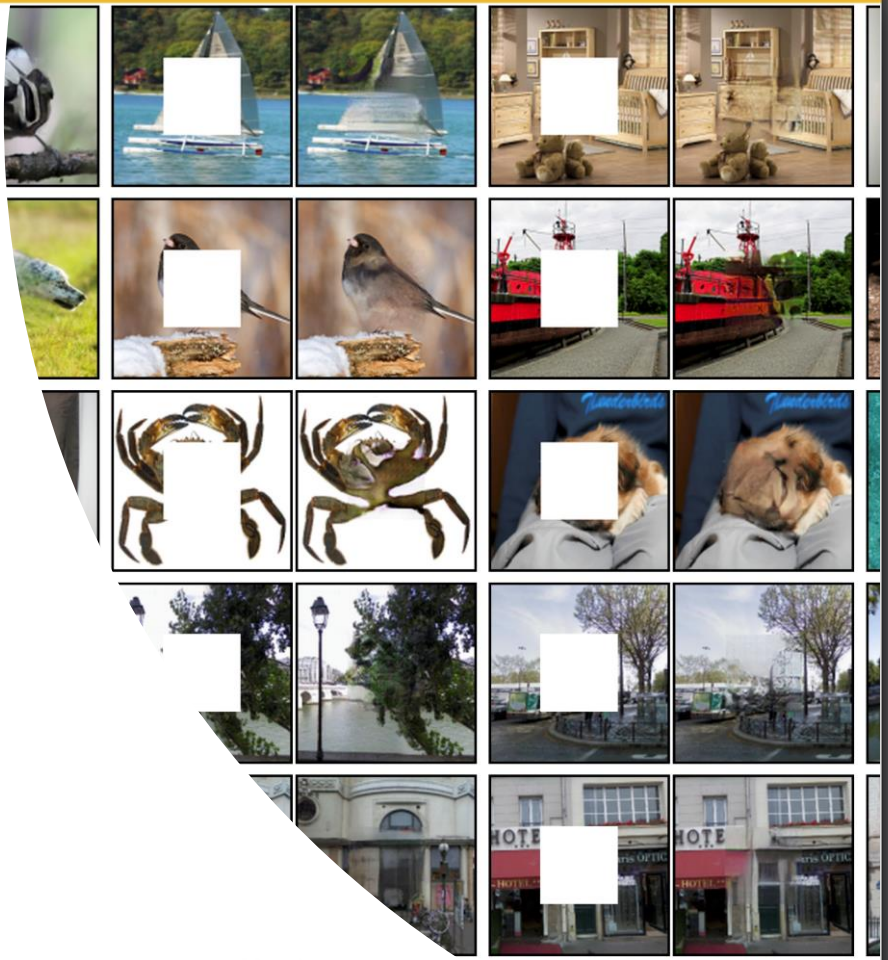


Complete protein

A missing region in the protein

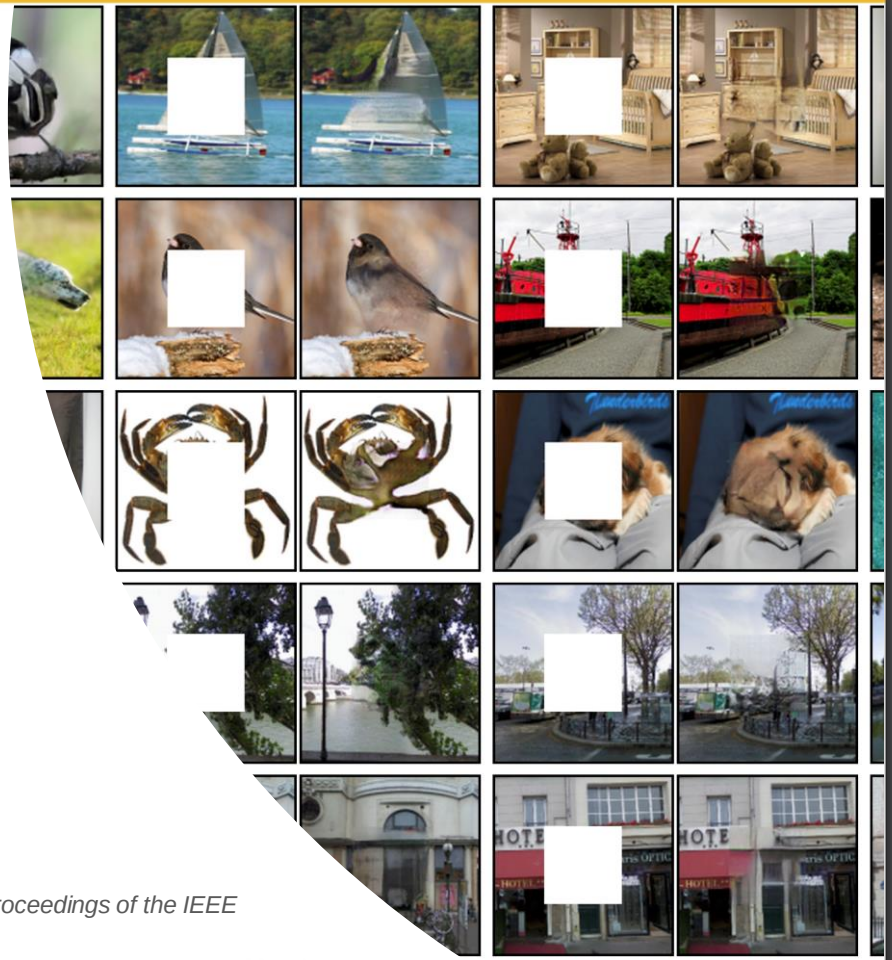
Motivation

- Image completion (inpainting): complete the missing region based on the image context



Motivation

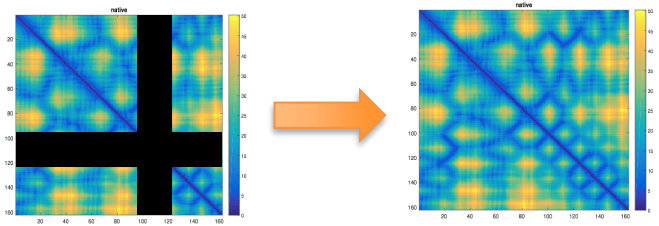
- Image completion (inpainting): complete the missing region based on the image context
- Generative Adversarial Network (GAN) outperformed previous methods [1]

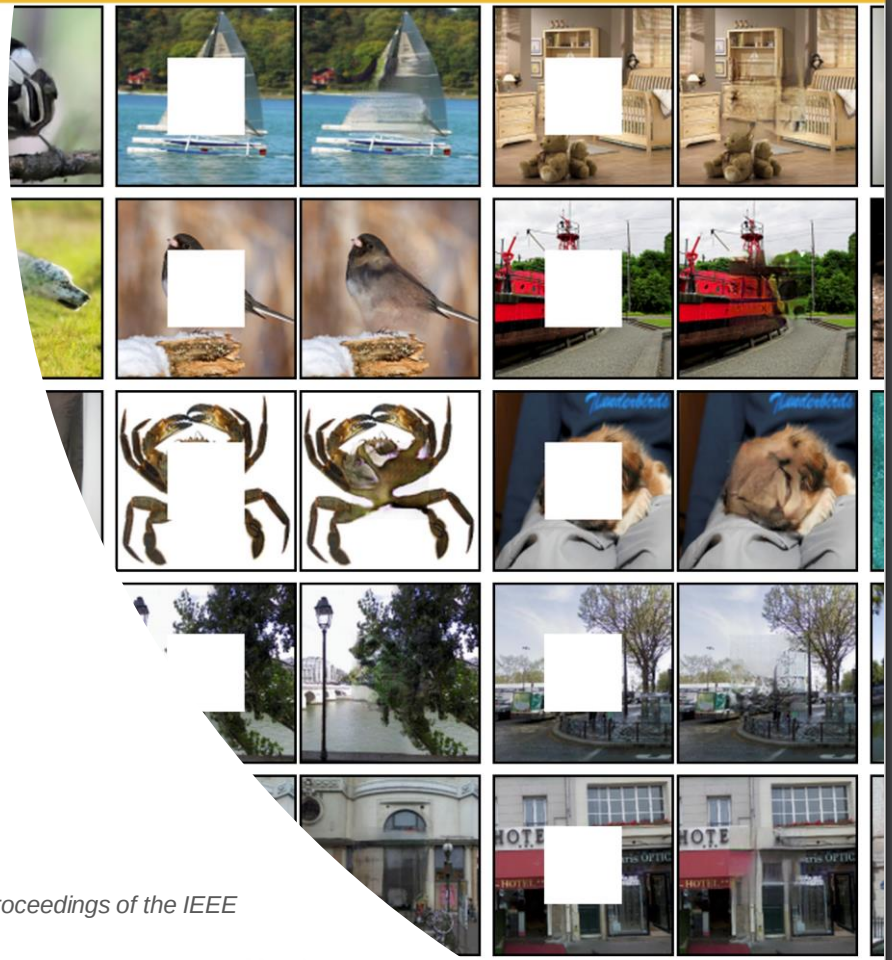


[1] Pathak, Deepak, et al. "Context encoders: Feature learning by inpainting." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Motivation

- Image completion (inpainting): complete the missing region based on the image context
- Generative Adversarial Network (GAN) outperformed previous methods [1]





[1] Pathak, Deepak, et al. "Context encoders: Feature learning by inpainting." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Prediction Flow of Our New Method

Target:
Length 50 with missing region in the middle

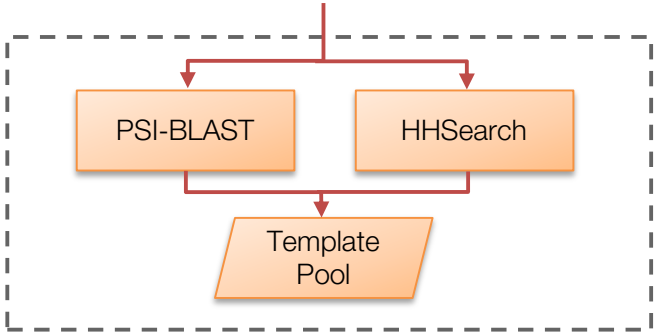
CDWEEISVKGPNGESSVIHDR**KSGKKFSI**EEALQSGRLTPAHYDRYVNKD

Prediction Flow of Our New Method

Target:
Length 50 with missing region in the middle

Alignment Searching:
Templates that can cover the whole query target

CDWEEISVKGPNGESSVIHDR**KSGKKFSI**EEALQSGRLTPAHYDRYVNKD



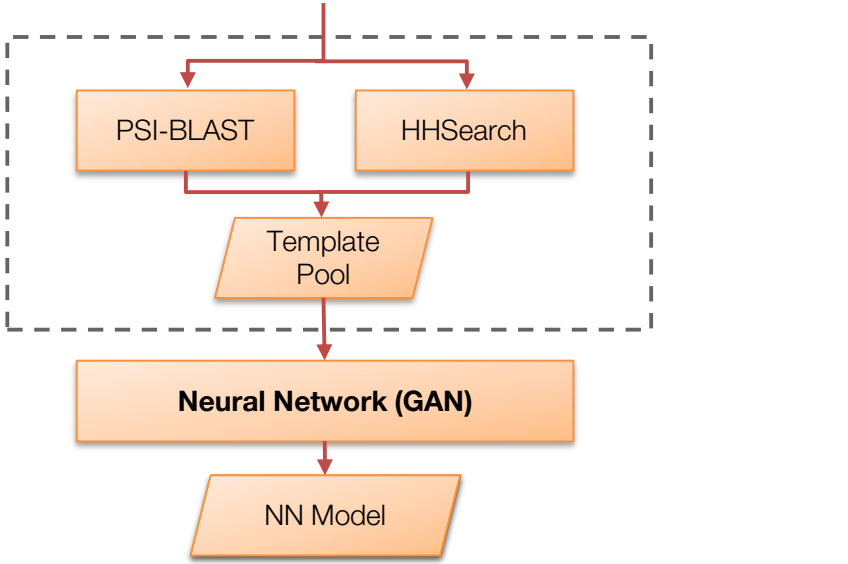
Prediction Flow of Our New Method

Target:
Length 50 with missing region in the middle

CDWEEISVKGPNGESSVIHDR**KSGKKFSI**EEALQSGRLTPAHYDRYVNKD

Alignment Searching:
Templates that can cover the whole query target

Neural Network Training:



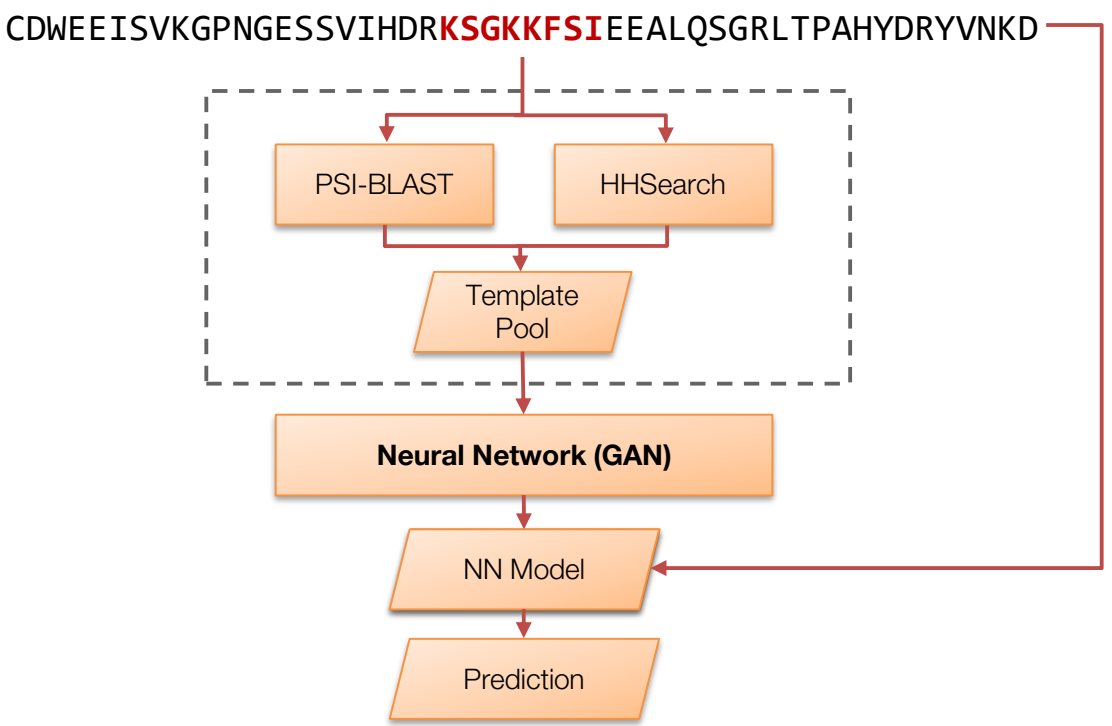
Prediction Flow of Our New Method

Target:
Length 50 with missing region in the middle

Alignment Searching:
Templates that can cover the whole query target

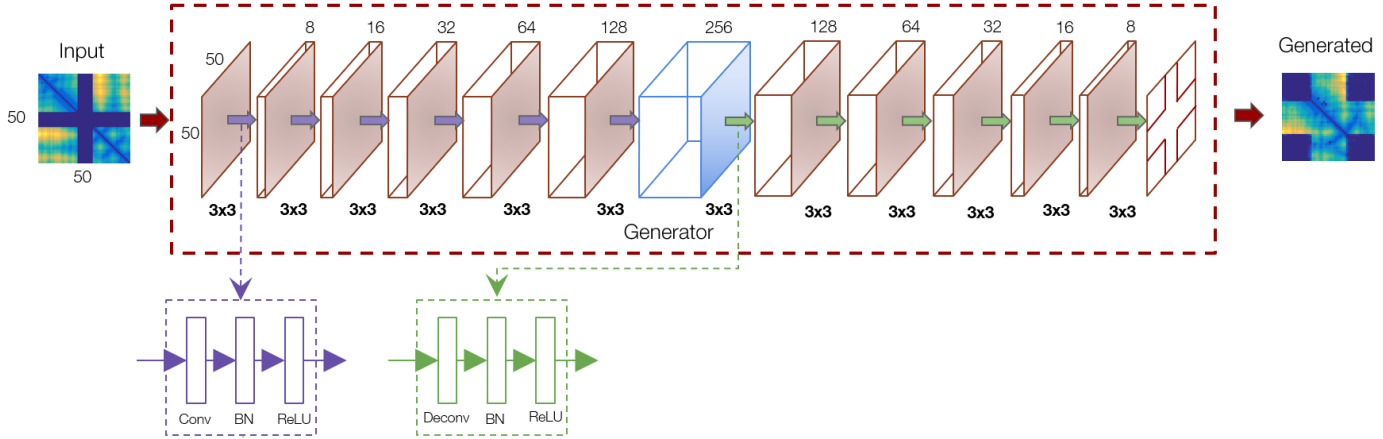
Neural Network Training:

Prediction:

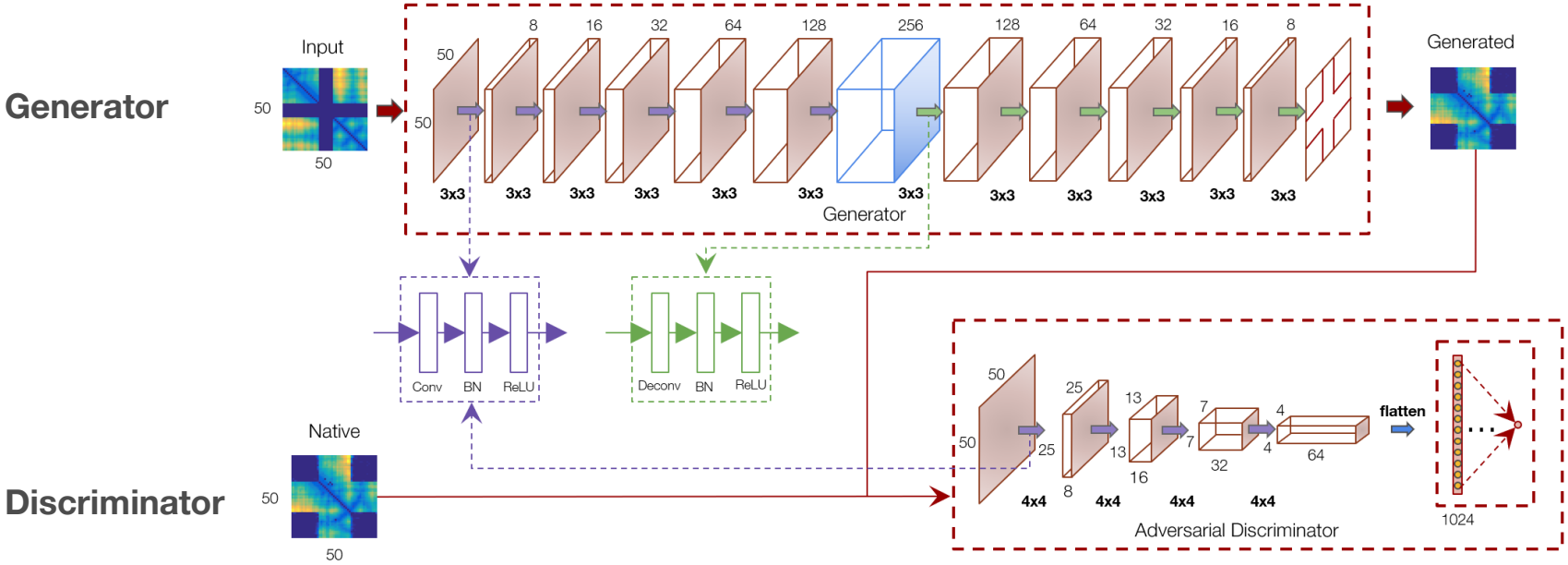


Our GAN Network Structure

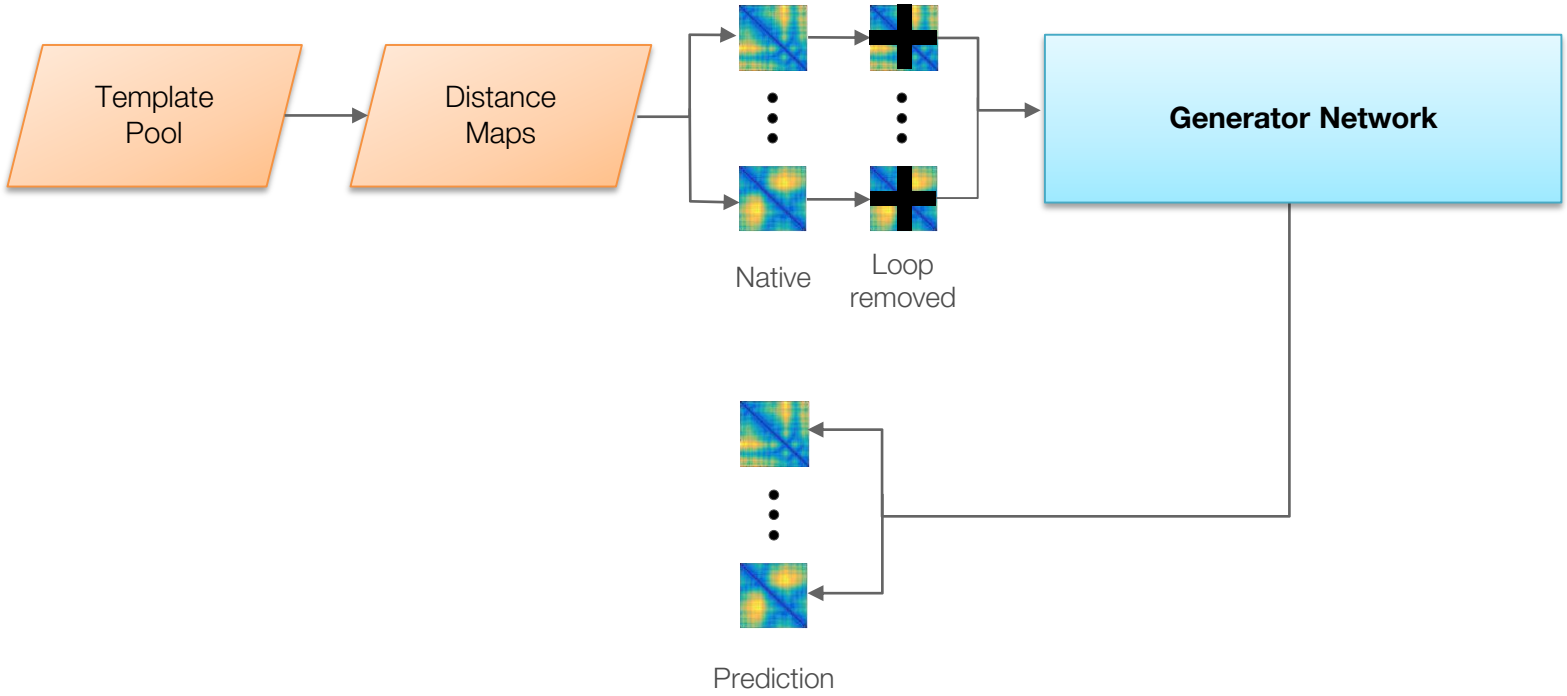
Generator



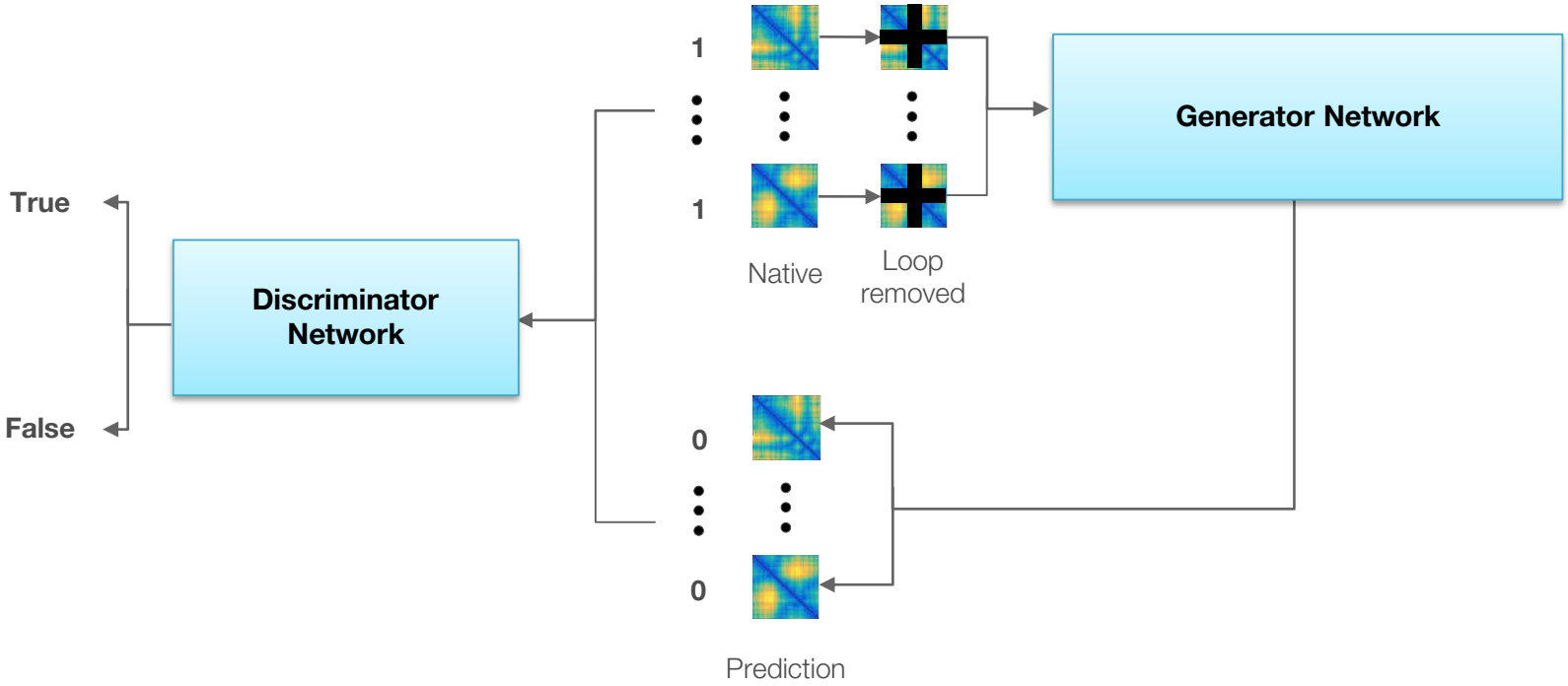
Our GAN Network Structure



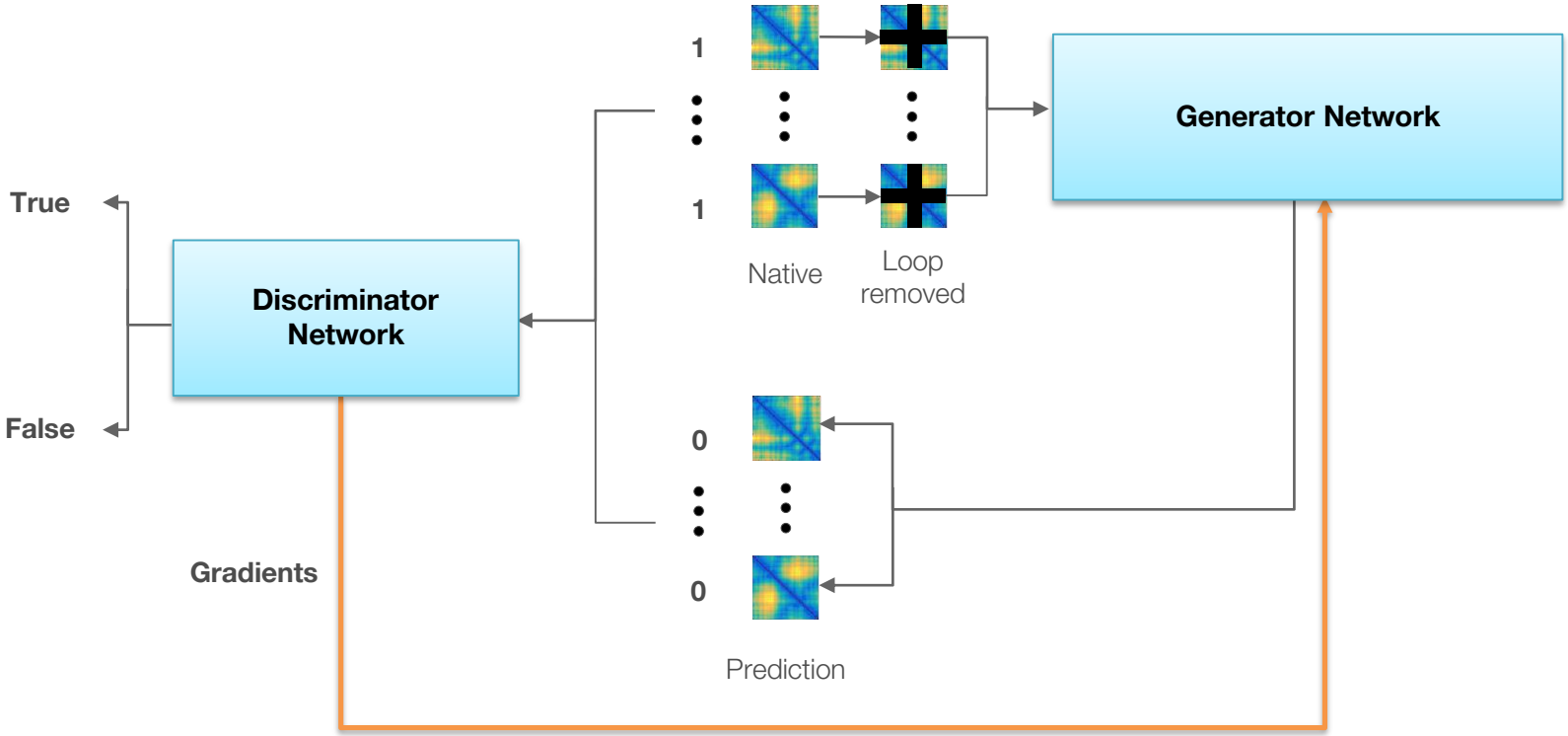
Training of the Network



Training of the Network



Training of the Network

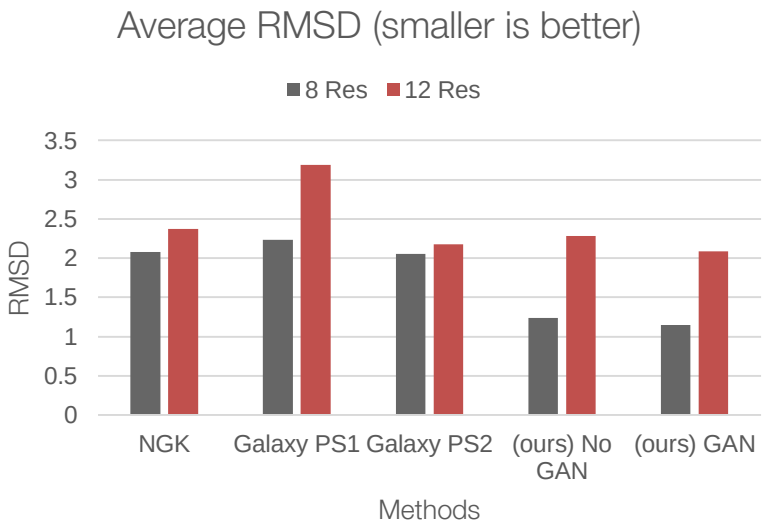


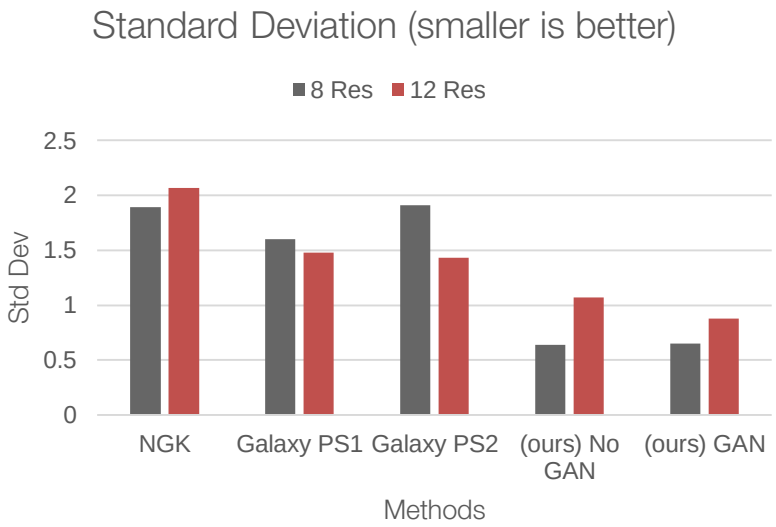
Implementation Details

- Platform
  - TensorFlow 1.0
- Training
  - Generator and Discriminator are trained alternatively
  - Train the Generator every step and train Discriminator every 10 steps
- Configuration
  - Adam optimizer with the learning rate 0.0001
  - Early stopping

Experimental Results

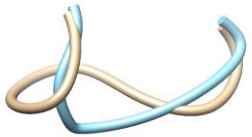
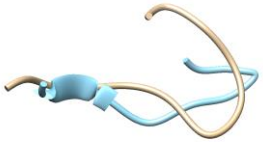
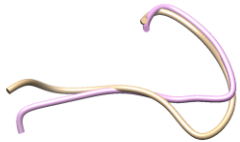
- *Without GAN (Generator)* and *with GAN (Generator + Discriminator)*
- Benchmark datasets from Park et al. including 20 targets with 8-length loop (8 Res), 20 with 12-length loop (12 Res)





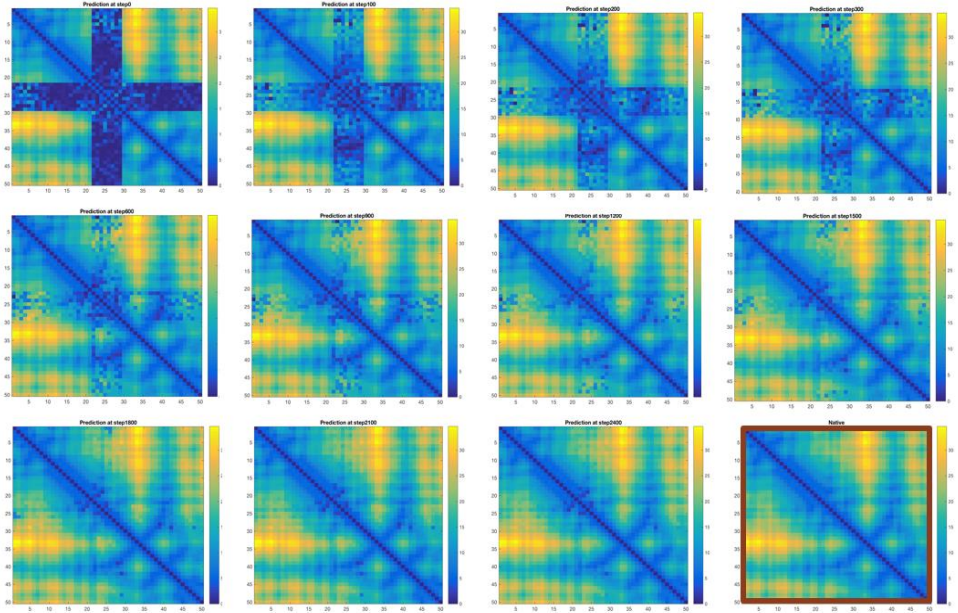
Experimental Results

- Visualization of two structures *without GAN* and *with GAN* (yellow one is the native)

	Without GAN	With GAN
1CLC (8 Res)	 RMSD: 1.6	 RMSD: 0.6
1BN8 (12 Res)	 RMSD: 3.0	 RMSD: 1.7

Experimental Results

- Visualization of *predictions* during the training process



Conclusion

- **First successful application of GAN in bioinformatics**
- GAN makes the results more realistic and stable

- Distance map representation of protein allows us to treat it as an image
- Easy to apply deep neural networks on images
- Future work could be done to enable multiple loop modeling

Thank you for listening

Speaker: Zhaoyu Li
zlht3@mail.missouri.edu
https://zhaoyu.li
Department of Electrical Engineering and Computer Science (EECS)
University of Missouri - Columbia