Deep Learning and Big Data Analytics for Protein Structure Prediction and Bioinformatics

Chao Fang, Zhaoyu Li, Wenbo Wang, Junlin Wang, et al., and Dr. Yi Shang in collaboration with Dr. Dong Xu

Distributed and Intelligent Computing Lab *(DICL, http://dslsrv1.rnet.missouri.edu/)*

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

DICL Lab Overview

- Extensive experiences in machine learning, deep learning, big data analytics, mobile computing, etc.
- 8 ongoing research projects supported by NIH, NSF, MU, the state, and the industry
- 20 graduate and undergraduate researchers
- Developed many state-of-the-art methods and tools for protein structure prediction and bioinformatics problems, including
  - Protein secondary structure & psi/phi angle prediction
  - Protein model quality assessment
  - Protein loop modeling
  - Protein 3D structure prediction and contact prediction
  - Multiple longest common subsequence problems

(1) Secondary Structure Prediction

- Each amino acid in a protein can be classified into one type of secondary structure.
- Accurate secondary structure prediction is useful for protein 3D structure prediction and functional analysis.

Methods

Developed several novel deep neural network architectures, such as

- DeepNRN: Deep Neighbor Residual Network
- Deep3I: Deep Inception Inside Inception Network

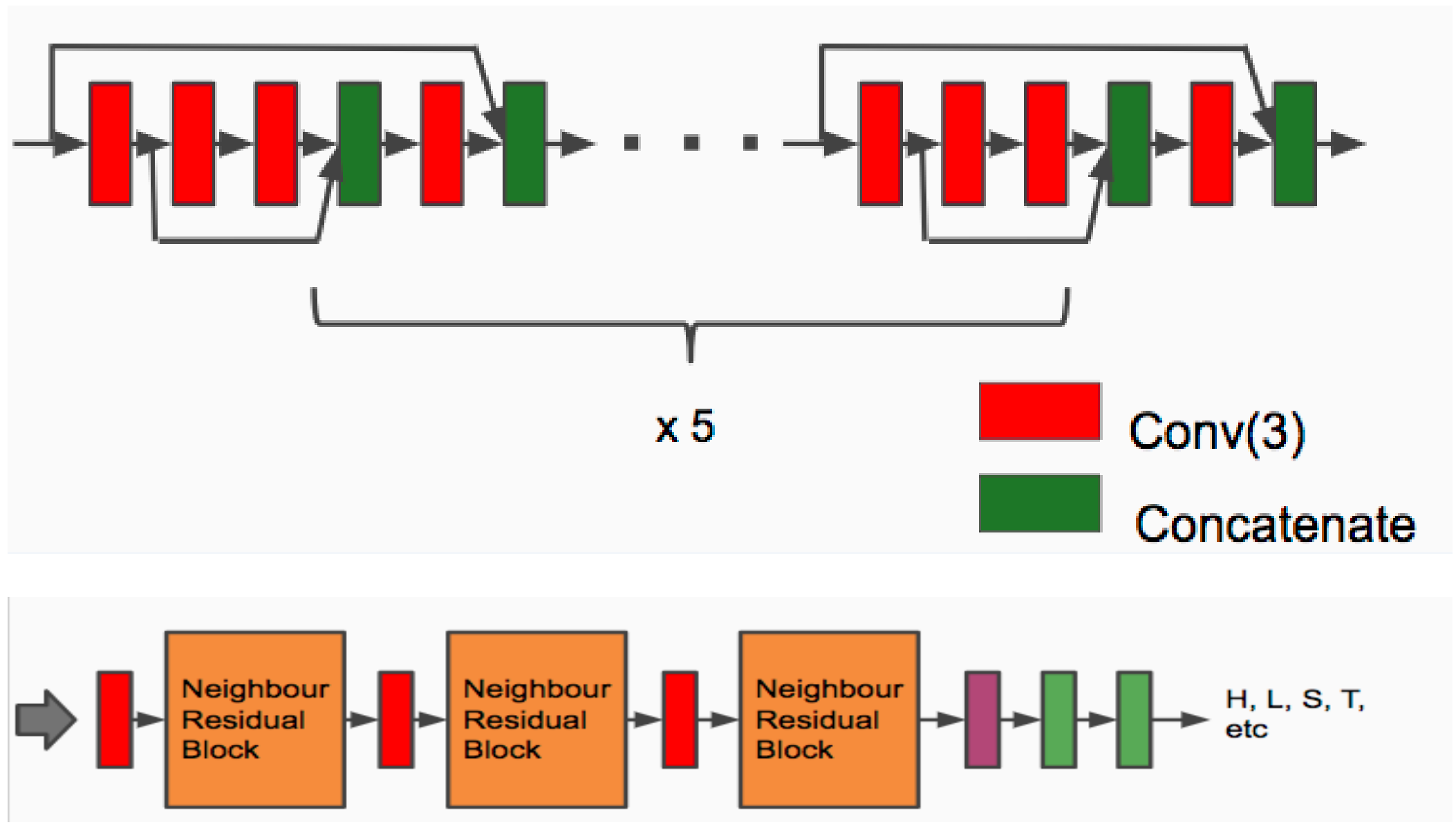

Figure 1. Neighbor-Residual Block and DeepNRN Architecture

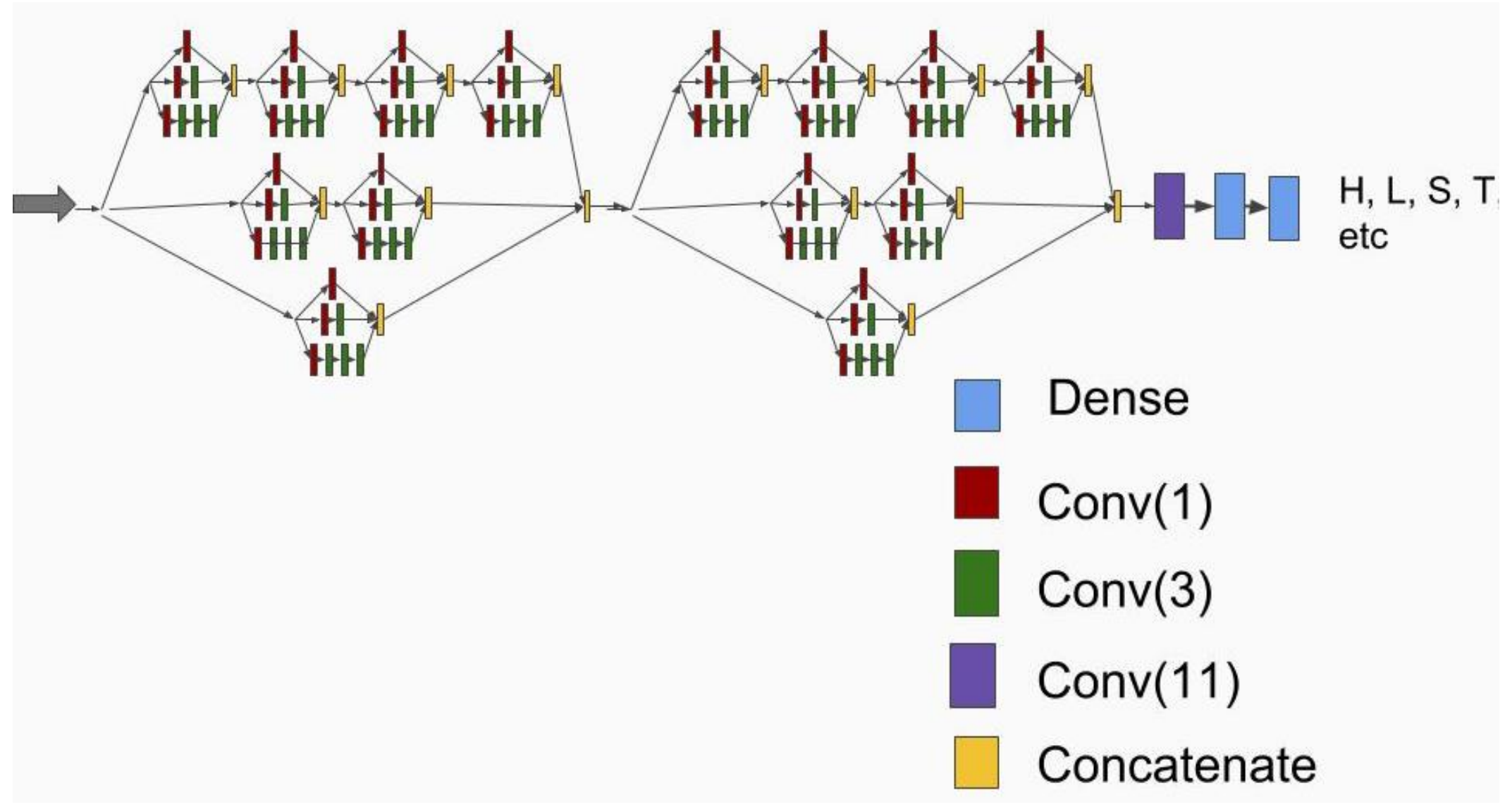

Figure 2. Deep Inception-Inside-Inception (Deep3I) Architecture

Results

- The new deep learning methods outperformed existing methods and obtained the best results on many benchmark datasets in the field.

(2) Quality Assessment

Machine Learning for QA

- Problem: How to select good models from a large set of candidate protein 3D models?
- Approach: machine learning and big data analytics

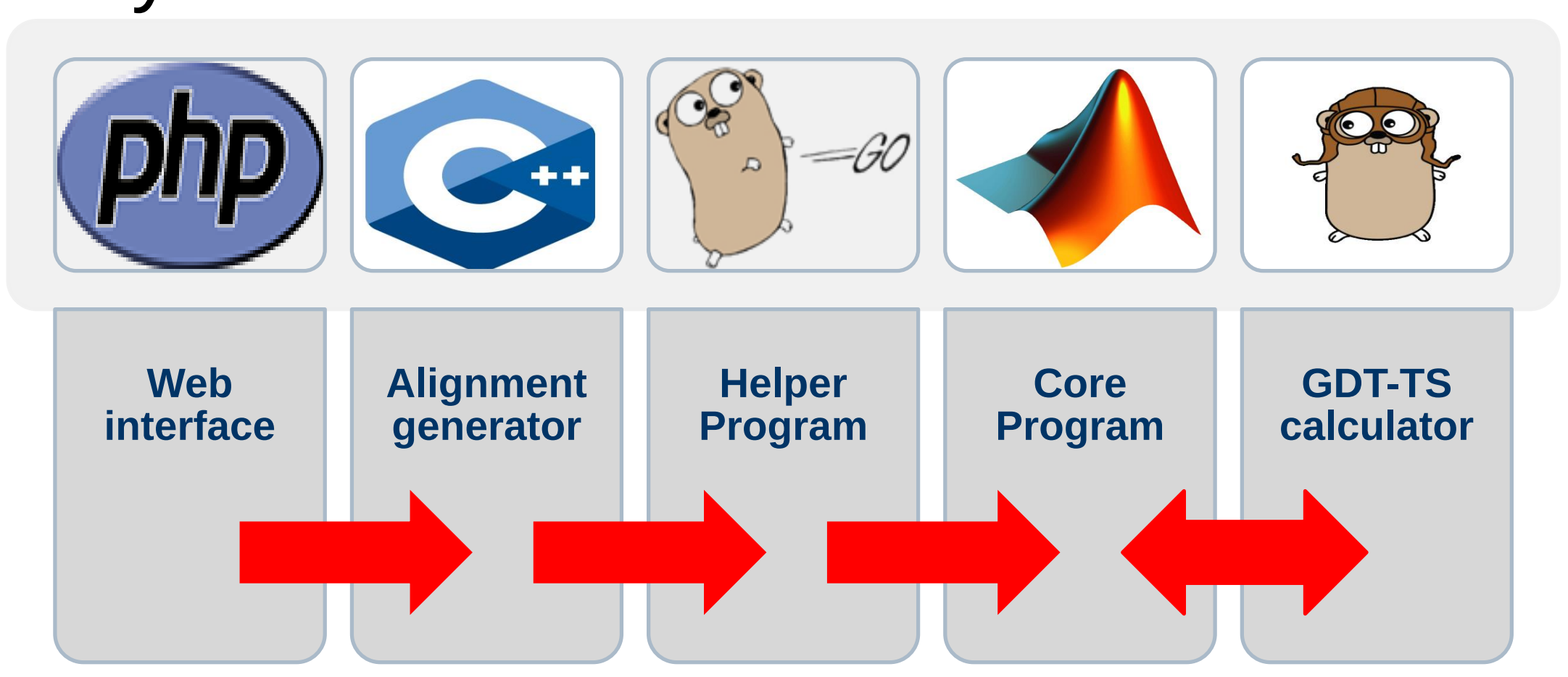

Figure 3. System Architecture For Automatic Protein Model QA

Results

- In the world-wide CASP12 protein structure prediction competition held in 2016, two of our new QA methods, MUfoldQA_S & MUfoldQA_C, ranked No. 1 in their respective categories.

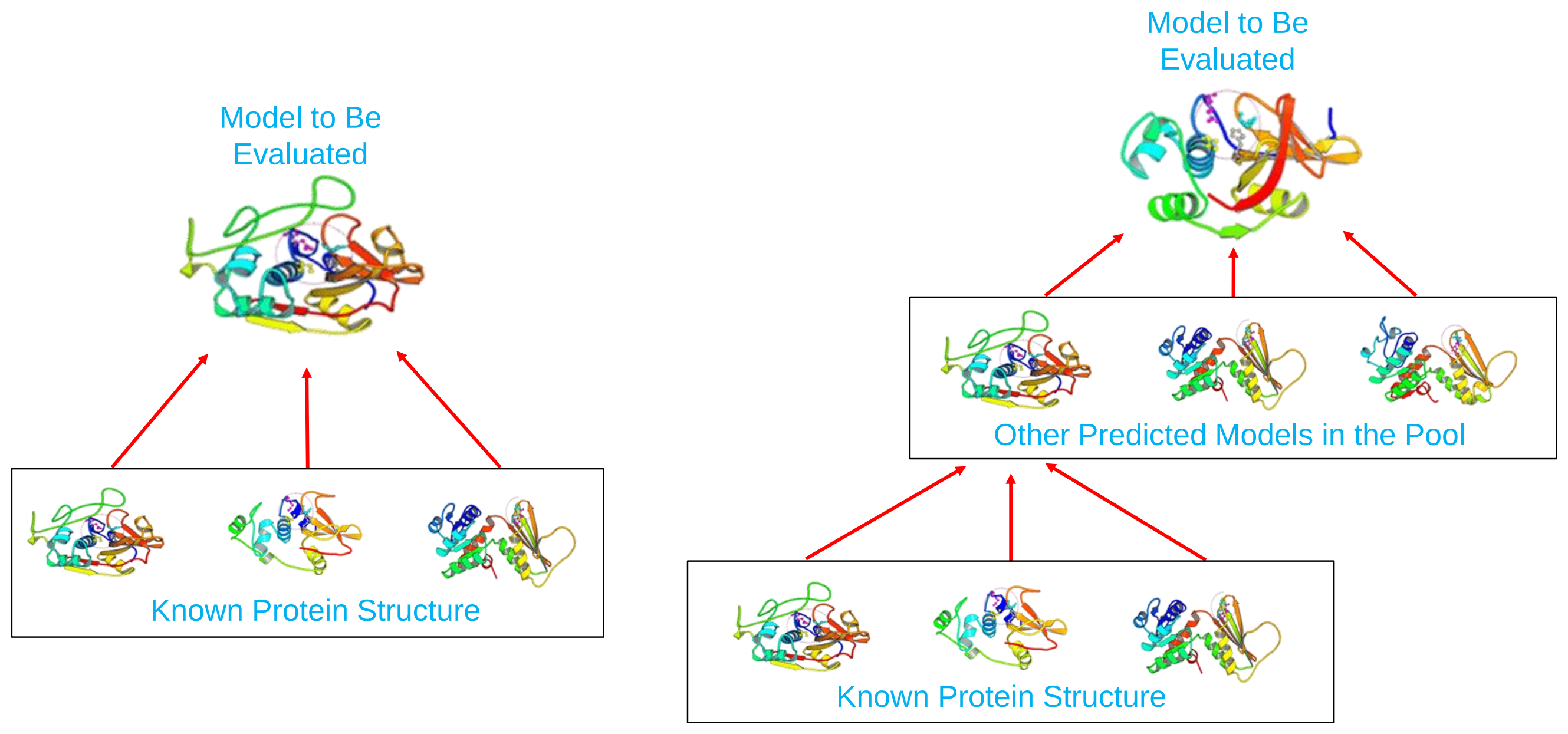

Figure 4. Illustration of MUfoldQA_S (left) and MUfoldQA_C (right)

- MUfoldQA_S is a new single-model QA method, effectively utilizing structural information from known protein fragments.
- MUfoldQA_C is a new multi-model QA methods, effectively utilizing information from the known structures and other predicted models.

Deep Learning for QA

- Several new deep neural networks have been developed for single-model QA, to predict the quality score of a model.

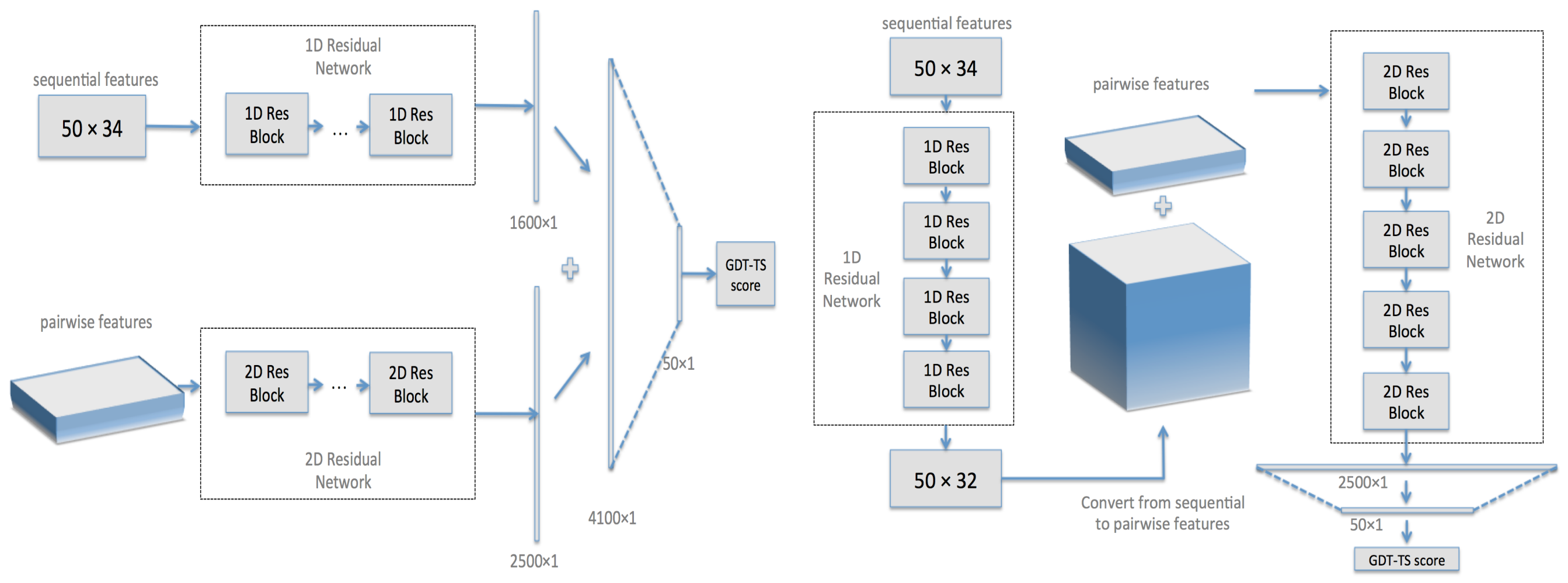

Figure 5. Deep Network Structures for QA

- Our new deep learning QA methods obtained promising results on multiple benchmark datasets including CASP datasets, comparable with the best existing state-of-the-art single-model QA methods.

(3) Loop Modeling

- Problem: filling the missing regions in a known protein structure.
- Approach: image completion techniques based on distance matrices and deep neural networks.
- The first successful application of Generative Adversarial Networks (GAN) architecture, an advanced deep learning method, to protein structure modeling and bioinformatics.

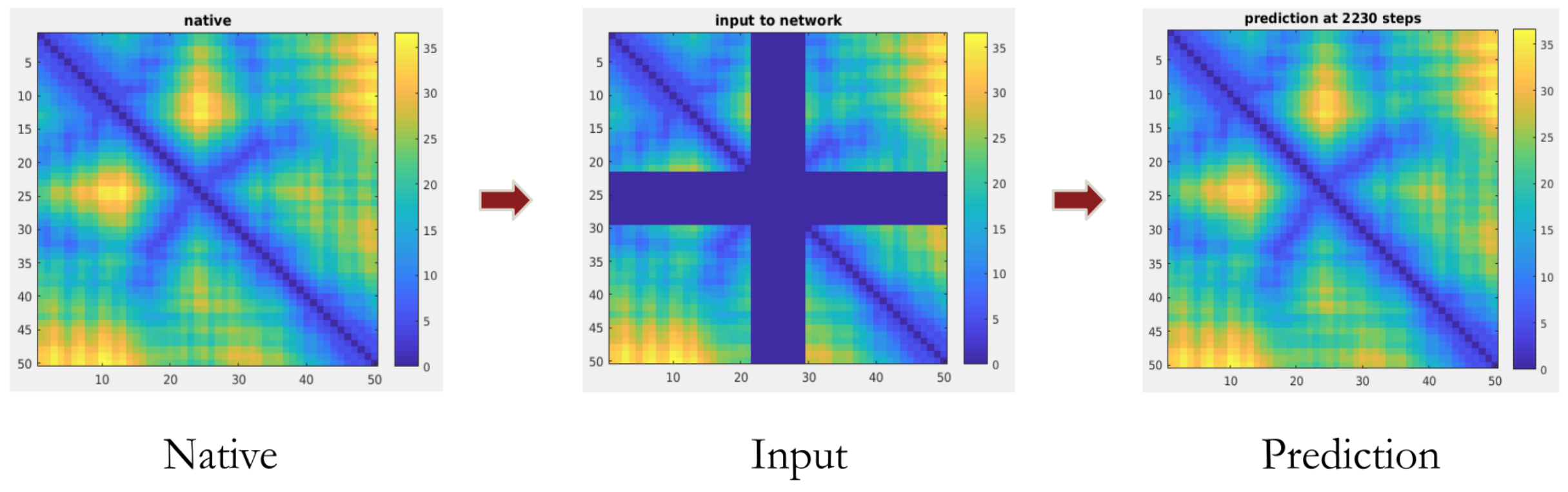

Figure 6. Loop Modeling as Image Completion

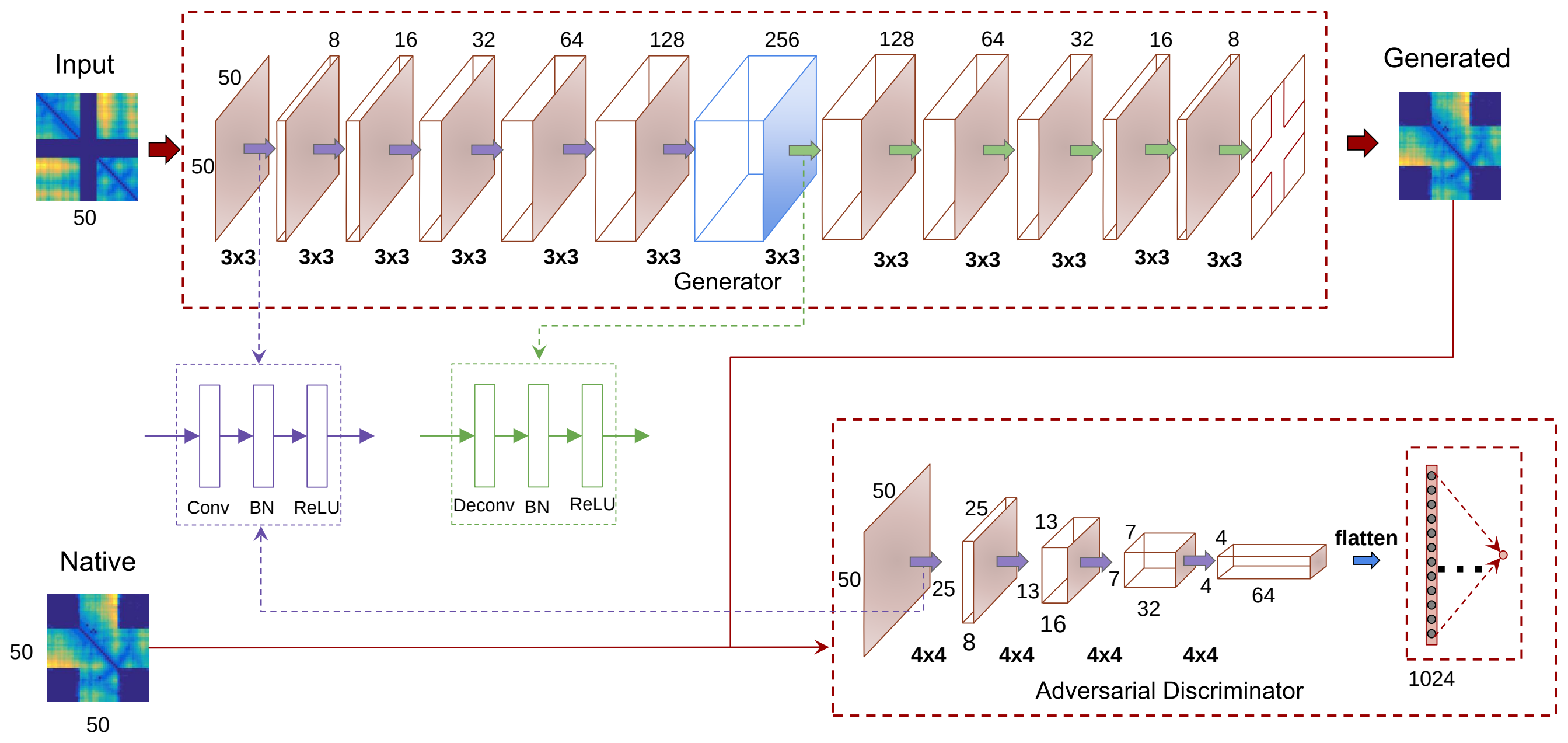

Figure 7. Deep GAN Network Architecture for Loop Modeling

Results

- The new method outperformed existing methods and obtained the best results on multiple benchmark datasets.
- Demonstrated the superior performance of GAN techniques in protein structure prediction.

Additional Areas

Protein Structure Psi-Phi Angle Prediction

- Developed several new deep neural networks for this task and obtained excellent results.
- A new deep inception residual network architecture (DeepIRN) outperformed the best existing method, SPIDERS, significantly.
- A web-based tool is being developed for public use.

Protein Structure Beta Turn Prediction

- Developed several new deep neural networks for this task and obtained excellent results.
- A new deep dense inception network architecture (DeepDIN) outperformed the best existing method, BetaTPred3, significantly.
- A web-based tool is being developed for public use.