Supervised Convolutional GSN for Protein Secondary Structure Prediction

Jian Zhou
Olga Troyanskaya

Princeton University

What’s In this talk..

- Problem: Predict protein secondary structure

- Iterative prediction with multi-layer hierarchical representation
  - Supervised GSN
  - Convolutional architecture for GSN
  - A trick for improving convergence and performance

- Performance evaluations

Protein secondary structure prediction

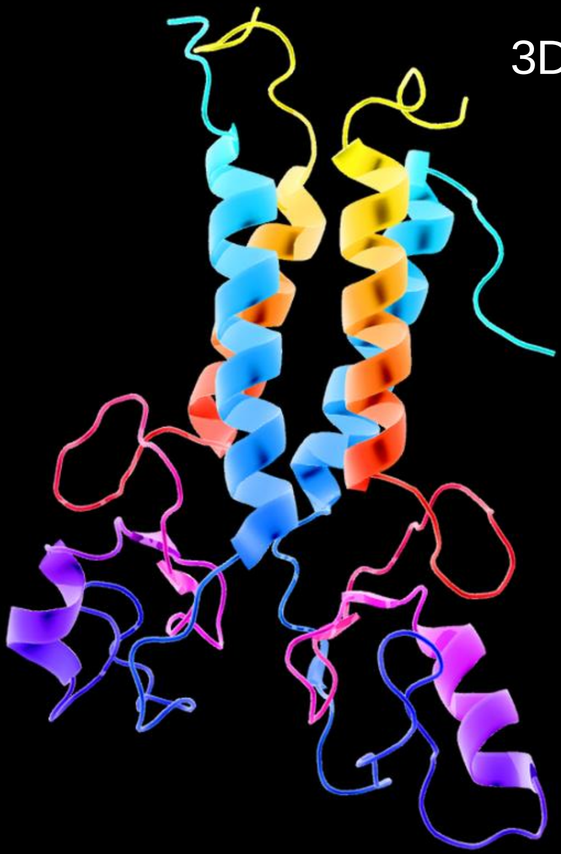
Protein sequence *20 types of amino acids*

MDLSALRVEEVQNVINAMQKILECP
ICLELIKEPVSTKCDHIFCKFCMLKL
LNQKKGPSQCPLCKNDITKRSLQE
STRFSQLVEELLKIICAFQLDTGLEY
ANSYNFAKKGK

Predict

Secondary structure *8 classes*

CCGGGSSHHHHHHHHHHHHHHTS
CSSSCCCCSSCCBCTTSCCCCSH
HHHHHHHSSSSSCCCTTTSCCCC
TTTCBCCCSSSHHHHHHHHHHHH
HHHHTCCCCCC



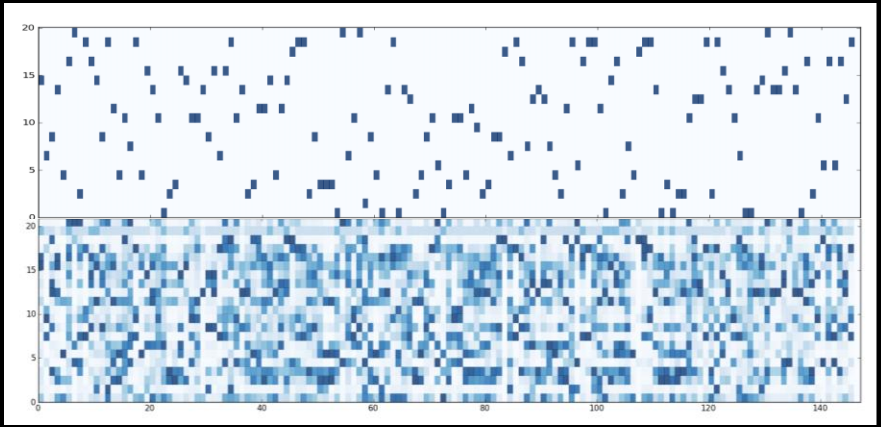
3D structure

Previous Approaches: neural network from 1988 (Qian & Sejnowski); bidirectioal recurrent neural network (Baldi et al., 1999); conditional neural fields (Peng et al., 2009); many more…

Image credit:
Wikimedia common

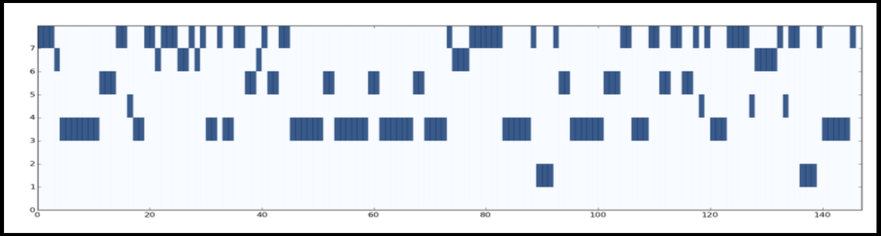
Protein Sequence -> Secondary Structure

Protein sequence *20 types of amino acids*

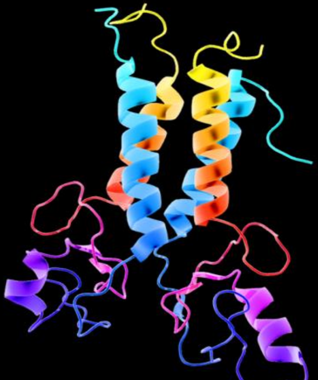


Evolutionary neighborhood

Predict



Secondary structure *8 classes*
label sequence

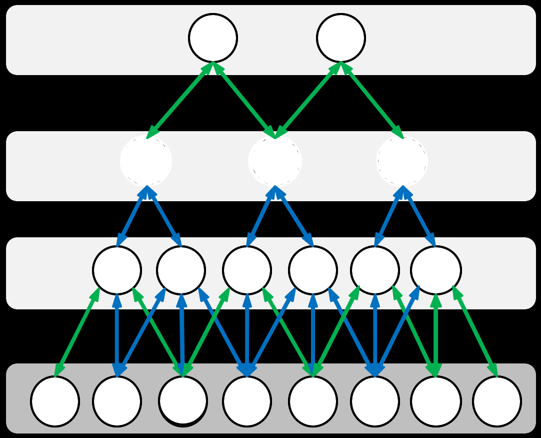


3D structure

Motivation

- Challenge: Prediction with both local and long-range dependencies

- Plan:
  - Multi-layer hierarchical representation
  - Both ‘upward’ and ‘downward’ connections
  - Supervised GSN formulation



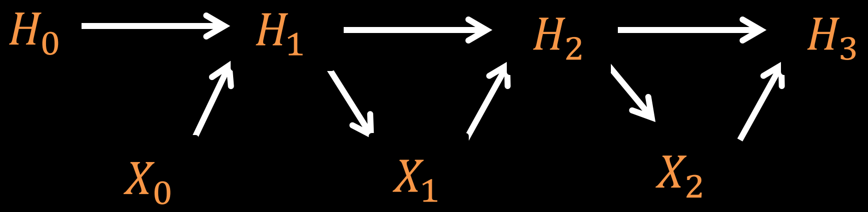
Model

- Generative Stochastic Network

  Bengio, Y., Thibodeau-Laufer, É., Alain, G., and Yosinski, J.
  Deep Generative Stochastic Networks Trainable by Backprop

Learning the transition operators of a Markov chain whose stationary distribution estimates the data distribution $P\ (X)$.

$$H_{t+1} \sim P_{\theta_1}(H \mid H_t, X_t)$$
$$X_{t+1} \sim P_{\theta_2}\ (X \mid H_{t+1})$$



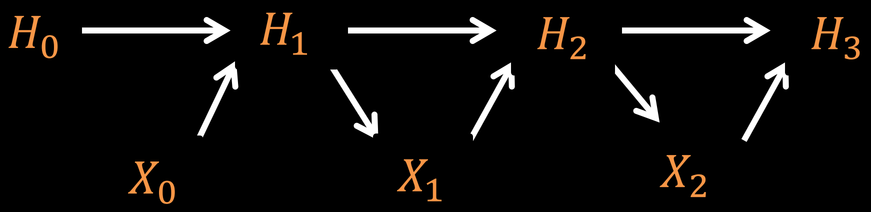
Learning $P\ (X \mid H)$ can be much easier than $P\ (X)$ by design.
Trainable using back-propagation

Model

GSN

P(X)

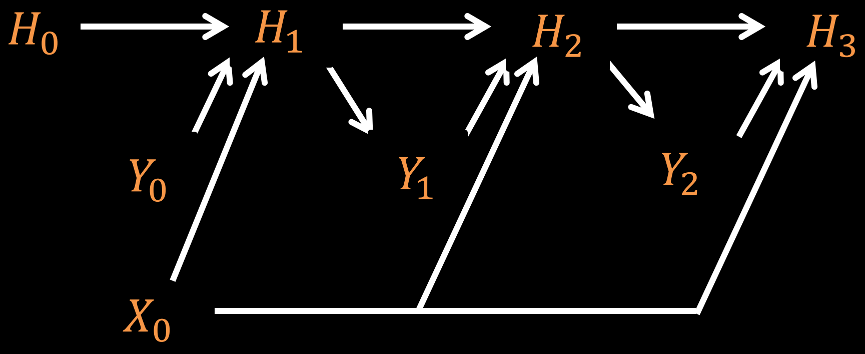
$$H_{t+1} \sim P_{\theta_1}(H \mid H_t, X_t)$$
$$X_{t+1} \sim P_{\theta_2}(X \mid H_{t+1})$$



Supervised GSN

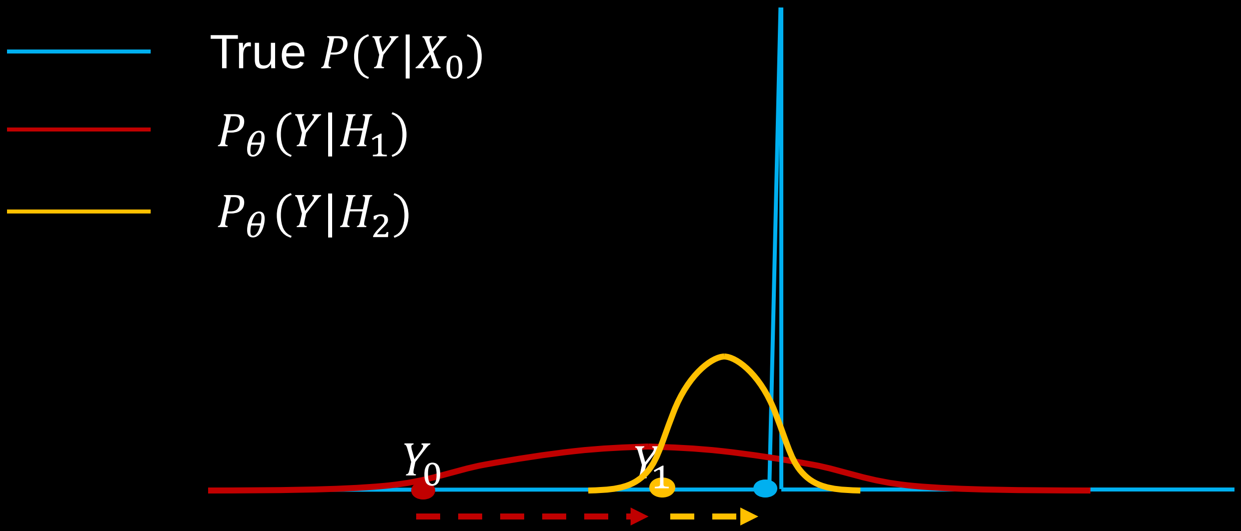
P(Y|X)

$$H_{t+1} \sim P_{\theta_1}(H \mid H_t, Y_t, X_0)$$
$$Y_{t+1} \sim P_{\theta_2}(Y \mid H_{t+1})$$



Learning $P\,(Y \mid H)$ can be much easier than $P\,(Y|X)$, utilizing previous state of the chain

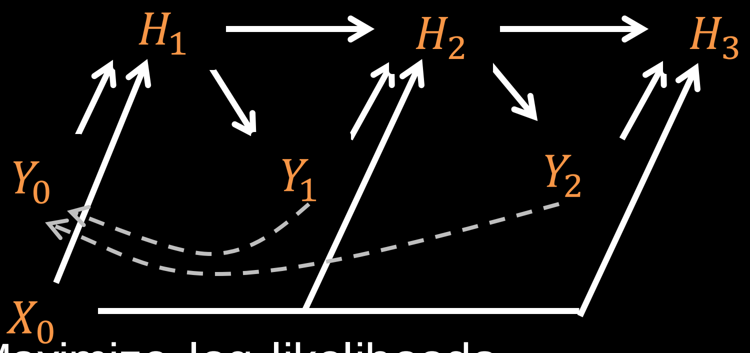
Model



Supervised GSN

P(Y|X)

$$H_{t+1} \sim P_{\theta_1}(H \mid H_t, Y_t, X_0)$$
$$Y_{t+1} \sim P_{\theta_2}(Y \mid H_{t+1})$$

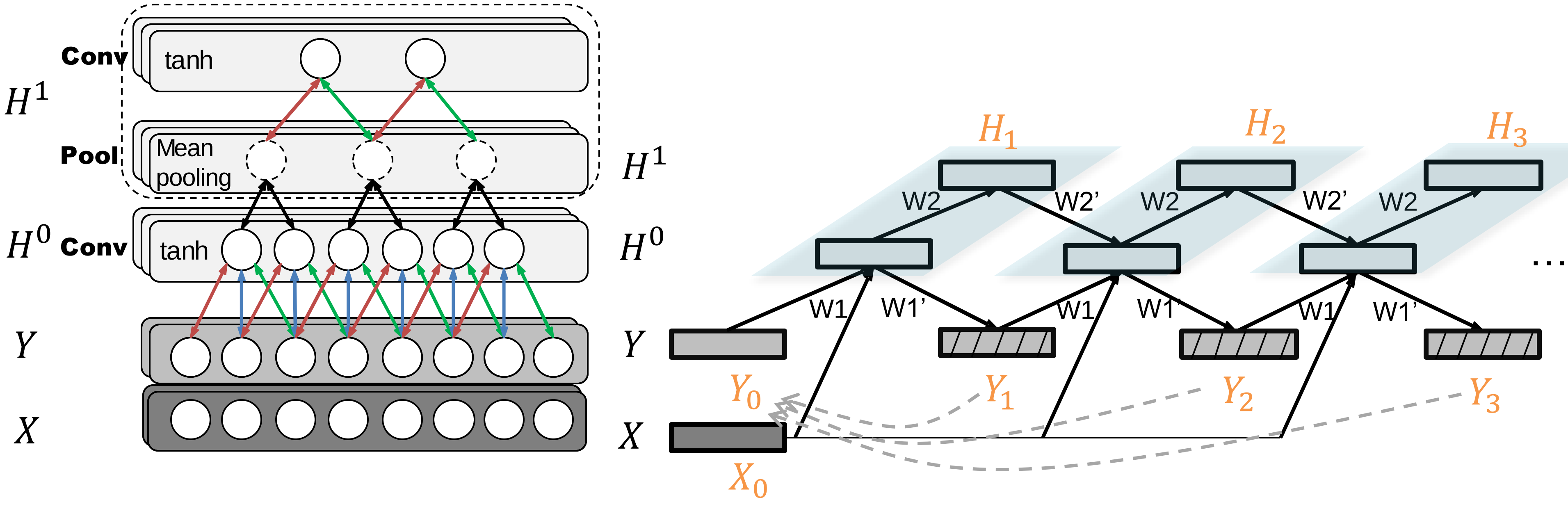


Maximize log-likelihoods

Model

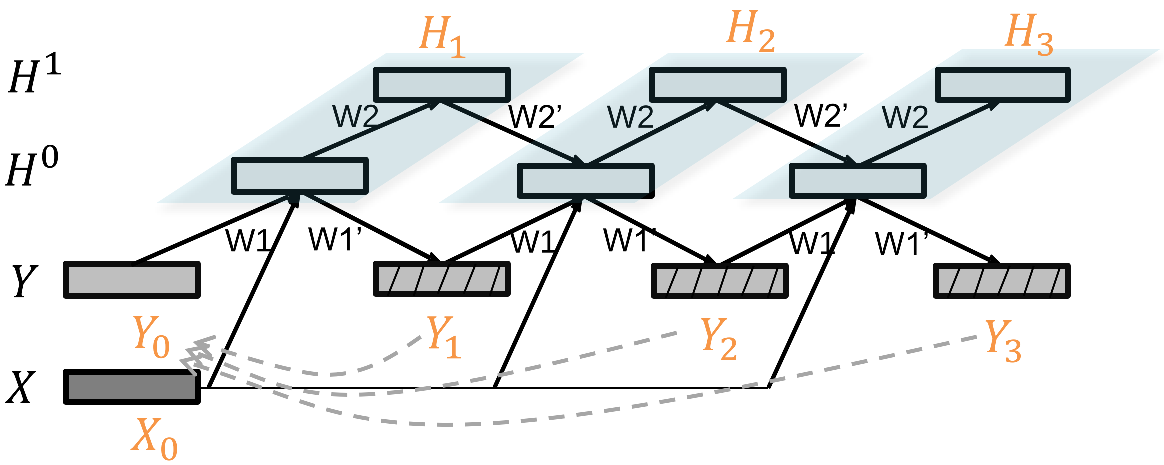
Architecture for protein secondary structure prediction

Multi-scale representation – multi-layer convolutional architecture
Local information sensitive – output unit at bottom layer



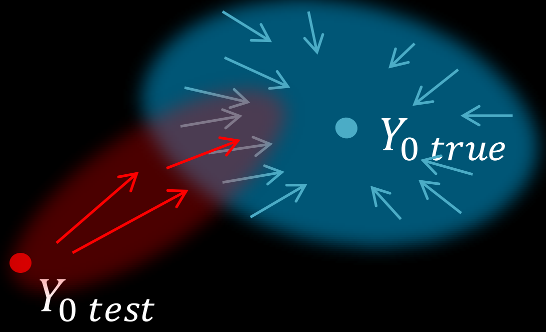
Training

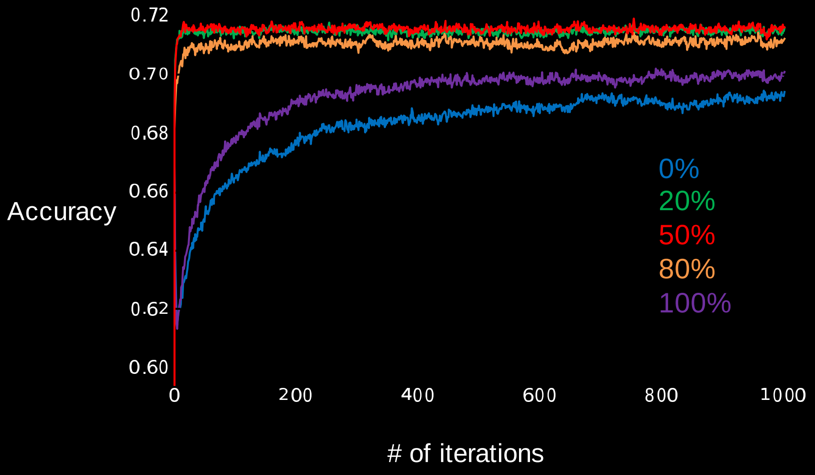
Experiments on initialization of chain during training



Initialize at a specified test initialization value for a subset of training batches:

- Optimal performance at 50% test initialization

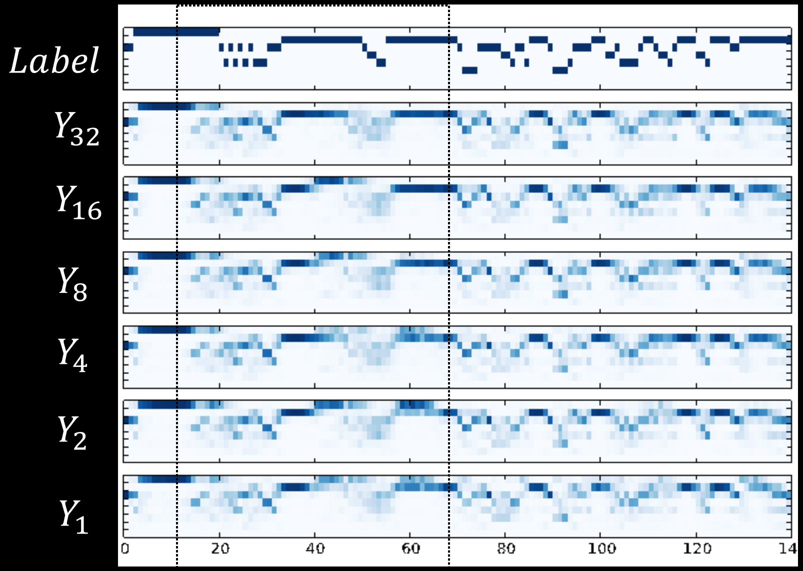




Performance

Cull PDB dataset (6133 proteins with <30% identity between any protein pairs);
available at www.princeton.edu/~jzthree/datasets

single protein prediction example



Performance through averaging iterative predictions:

CullPDB-30 test set	Overall Accuracy (8-class)
1 layer	0.714 ± 0.006
2 layers	0.720 ± 0.006
3 layers	0.721 ± 0.006

CB513 dataset	Overall Accuracy (8-class)
RaptorSS8/CNF	0.649 ± 0.003
Our method	0.664 ± 0.005

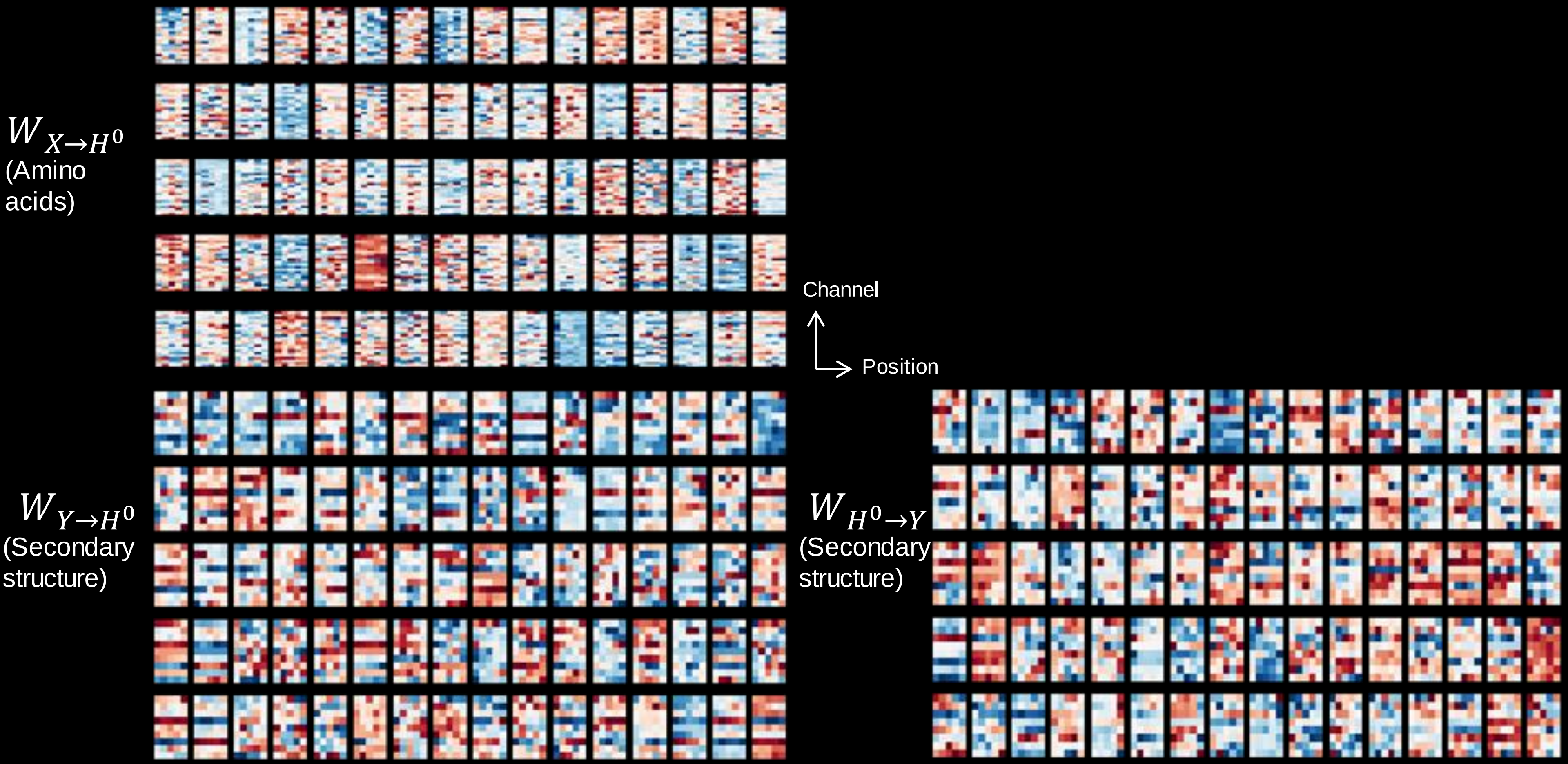
Summary

- We developed supervised convolutional GSN model for protein secondary structure prediction.

- Supervised GSN
  - Stochastic iterative prediction through Markov chain
  - Initialization trick improve both performance and convergence rate empirically

- Convolutional architecture for Supervised GSN
  - Combine high level representation and local prediction
  - Improved over previous best performance

- Filters: Layer1, $X, Y \leftrightarrow H^0$


$W_{X \rightarrow H^0}$
(Amino acids)
Channel
Position
$W_{Y \rightarrow H^0}$
(Secondary structure)
$W_{H^0 \rightarrow Y}$
(Secondary structure)