RNA secondary structure prediction with ***deep neural networks***

Advisor: Zhi J. Lu

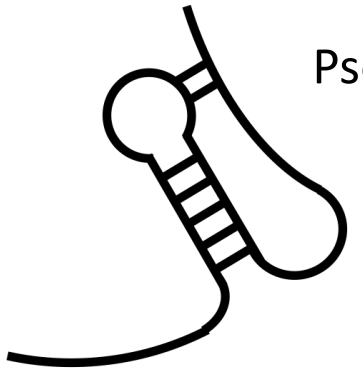
Presenter: **Zudi Lin**

2017 May 14

Background

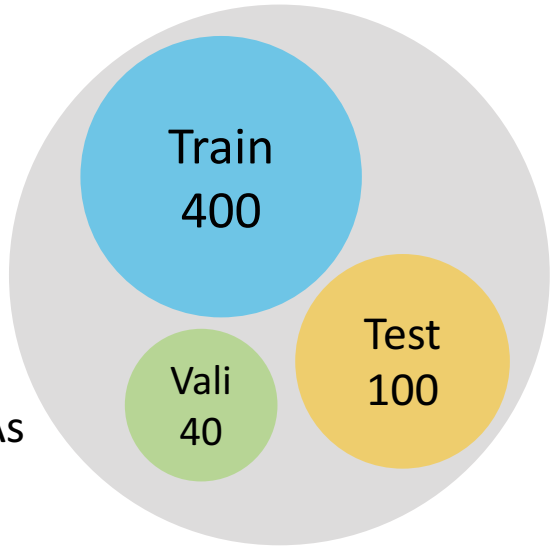
- **Dominant structure prediction methods**

1) Based on thermodynamic parameters
2) Nearest Neighbor Assumption
3) Dynamic Programming



Pseudo-knot

- **Dataset**

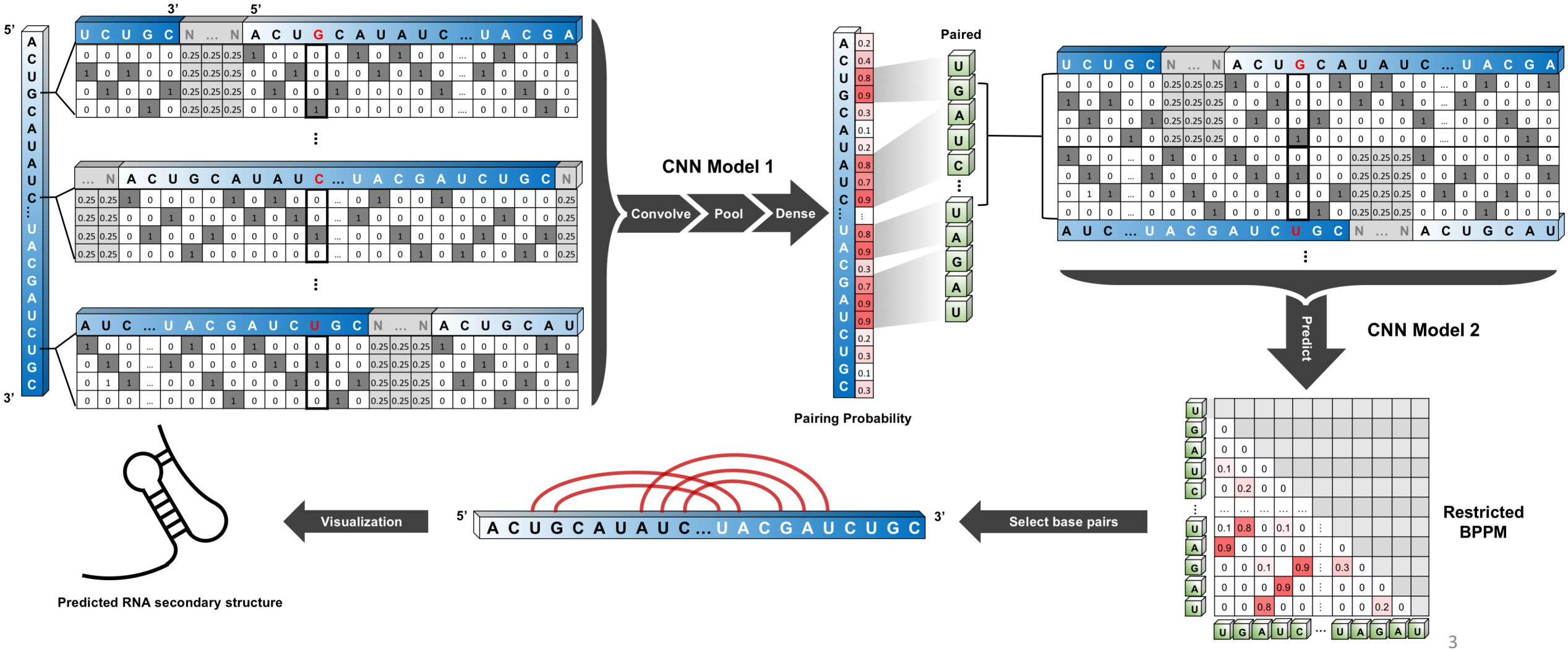


3324 RNAs

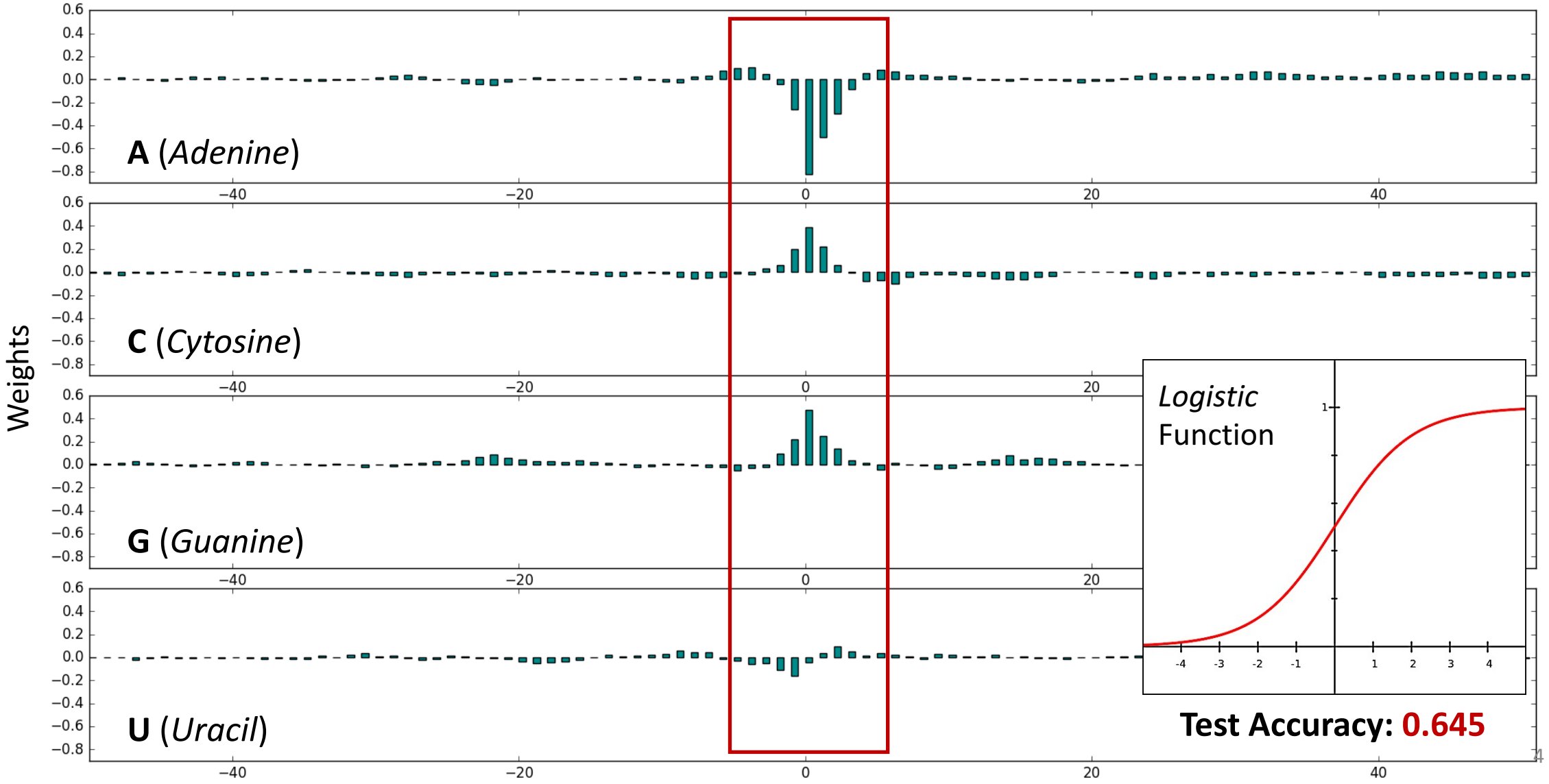
[1] Totally 3324 different RNA sequences were collected, including tRNA, rRNA, telomerase RNA, RNase *etc*..

[2] Filtered with *EMBOSS Needle* global alignment tool. Each pair of sequences have a similarity **< 60%.**

Prediction Pipeline

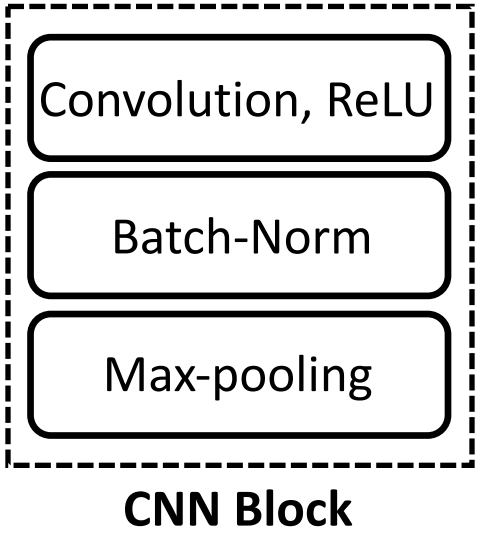


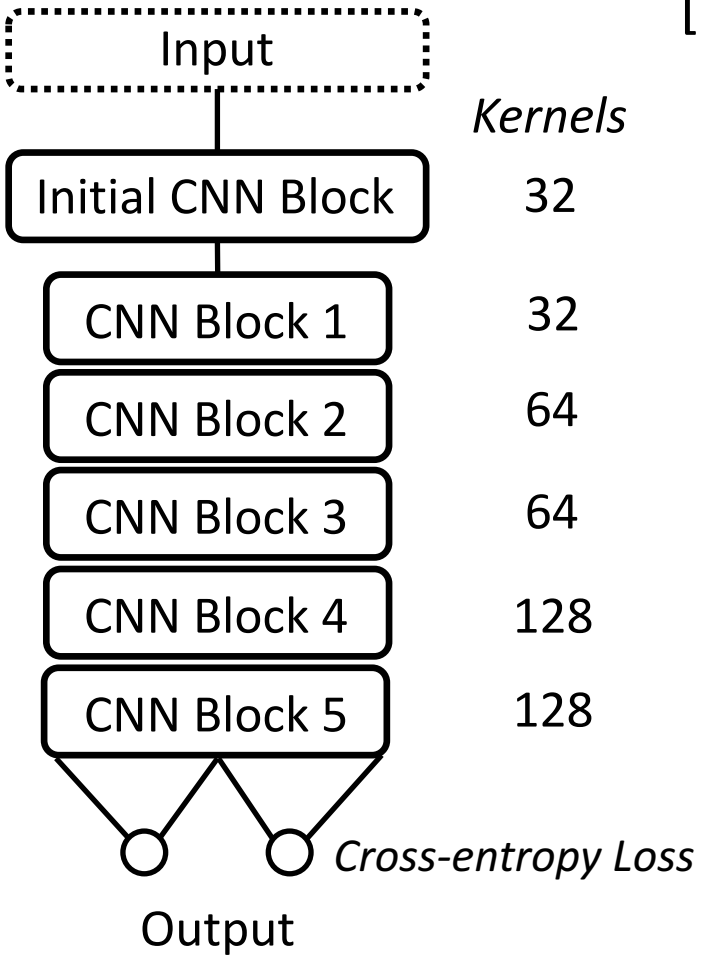
Logistic regression is ***not*** effective



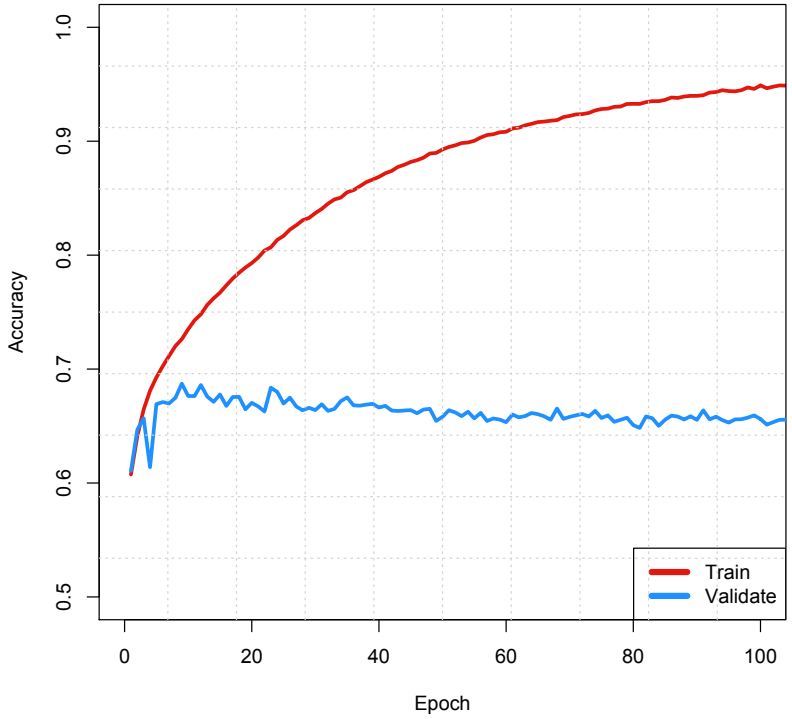
Network Architecture & Training Curve

[1]





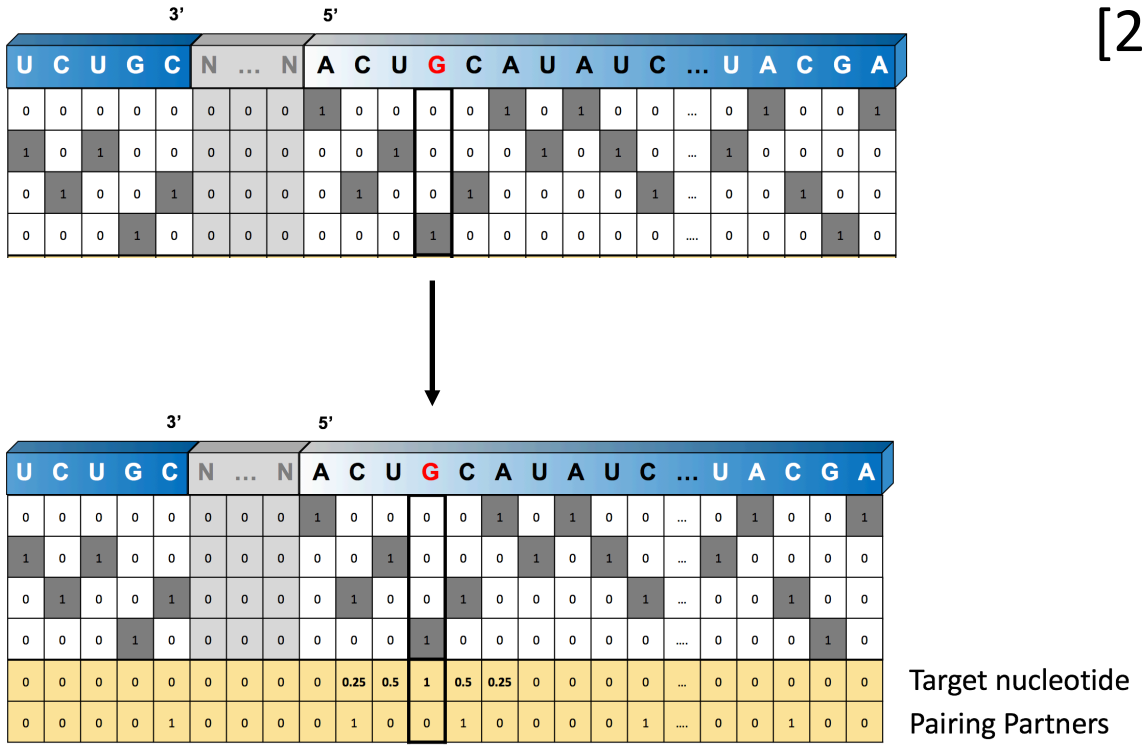
[2]



Training & Validation Accuracy

Solving the Over-fitting Problem

[1]



Circulative Representation

[2]

$$C = C_0 + \frac{\lambda}{2n} \sum_w w^2$$

L2 Regularization

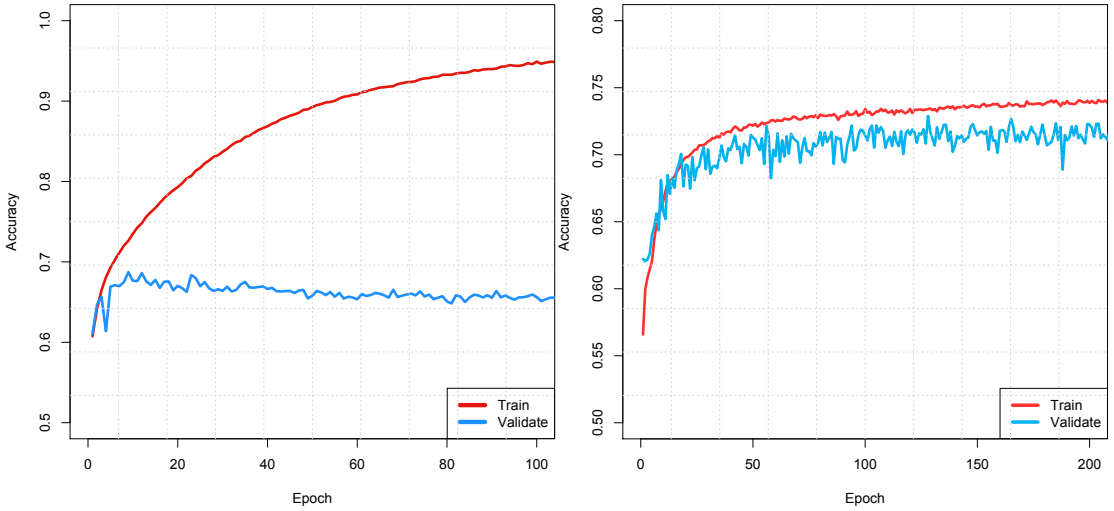
$$w \to \left(1 - \frac{\eta\lambda}{n}\right) w - \frac{\eta}{m} \sum_x \frac{\partial C_x}{\partial w}$$

$$b \to b - \frac{\eta}{m} \sum_x \frac{\partial C_x}{\partial b}$$

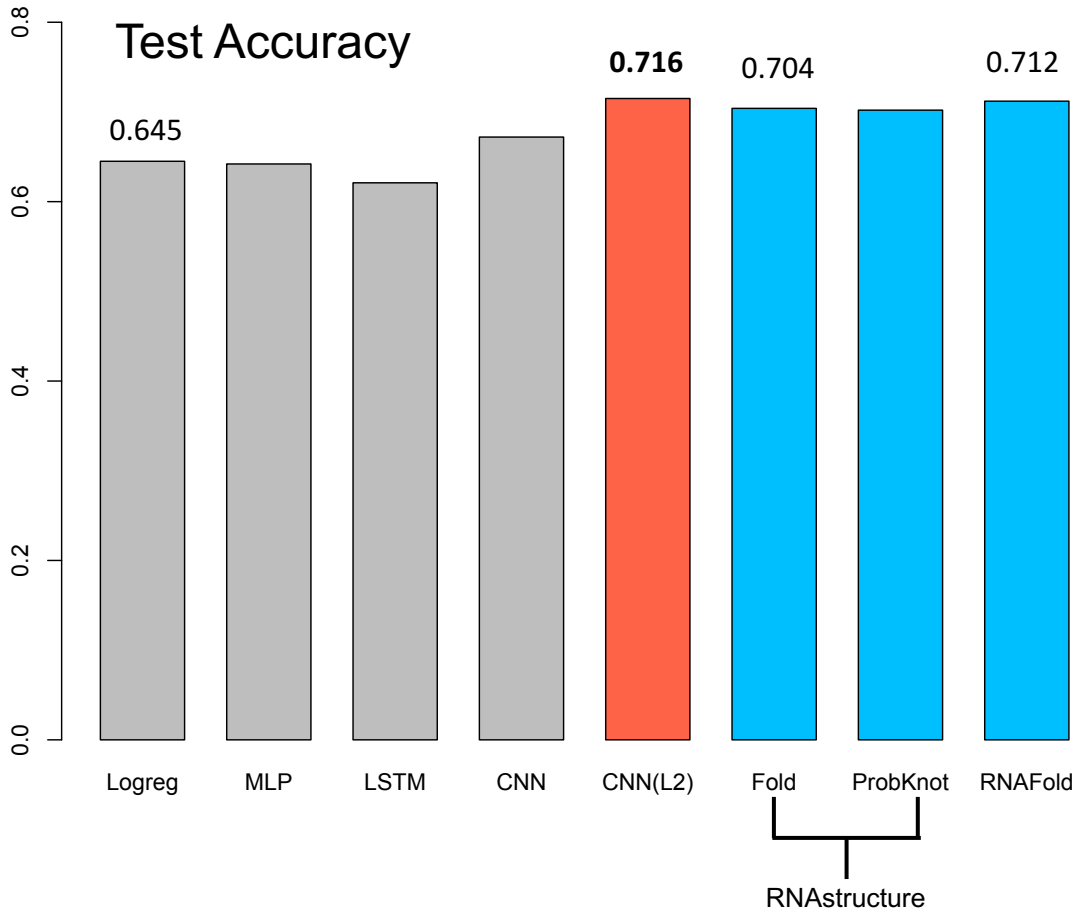
Stochastic Gradient Descent

One-dimensional prediction results

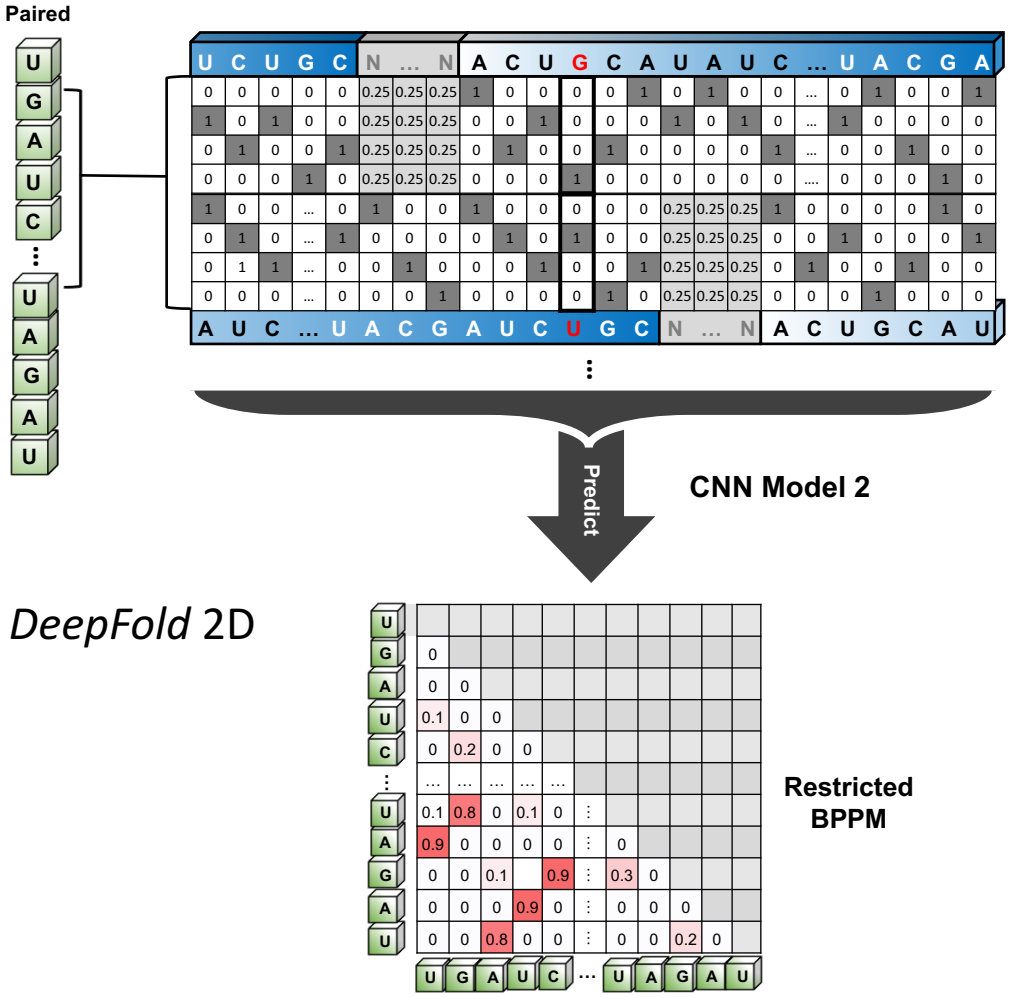
Training Curve



By utilizing L2 regularization and Gaussian noise, the validation accuracy is increased from **0.682** to **0.738**.



Two-dimensional structure prediction



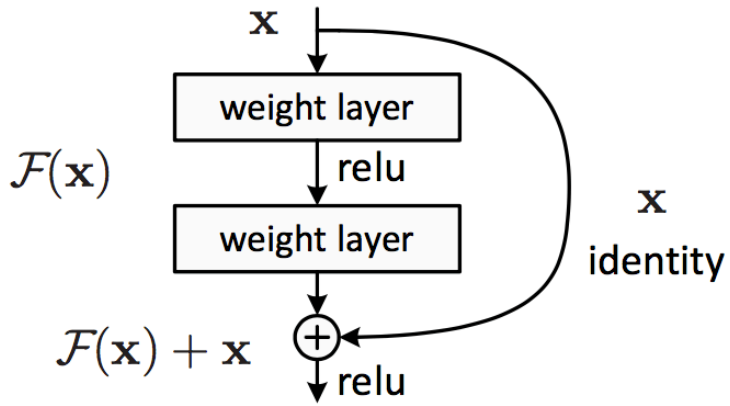
DeepFold 2D

[1] Severe Class Imbalance

a. Only consider canonical case pairing. Decrease neg-to-pos ratio from over 200 to 78.

b. Data-balanced training approach.

[2] Under-fitting problem



Residual learning framework

Two-dimensional structure prediction

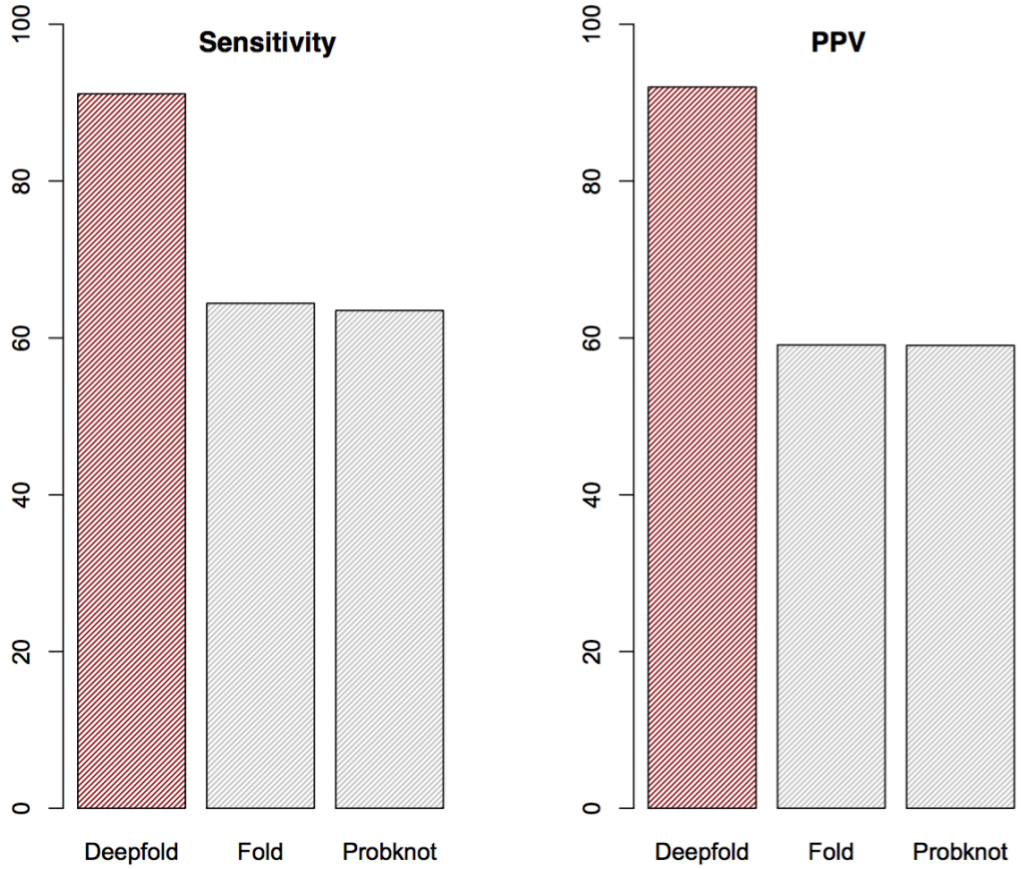
	Test Accuracy on positive samples	Test Accuracy on negative samples	$P(S^+\|T^+)$
MLP	0.75	0.81	0.06
CNN	**0.89**	0.70	0.04
Residual CNN	0.72	**0.98**	**0.31**

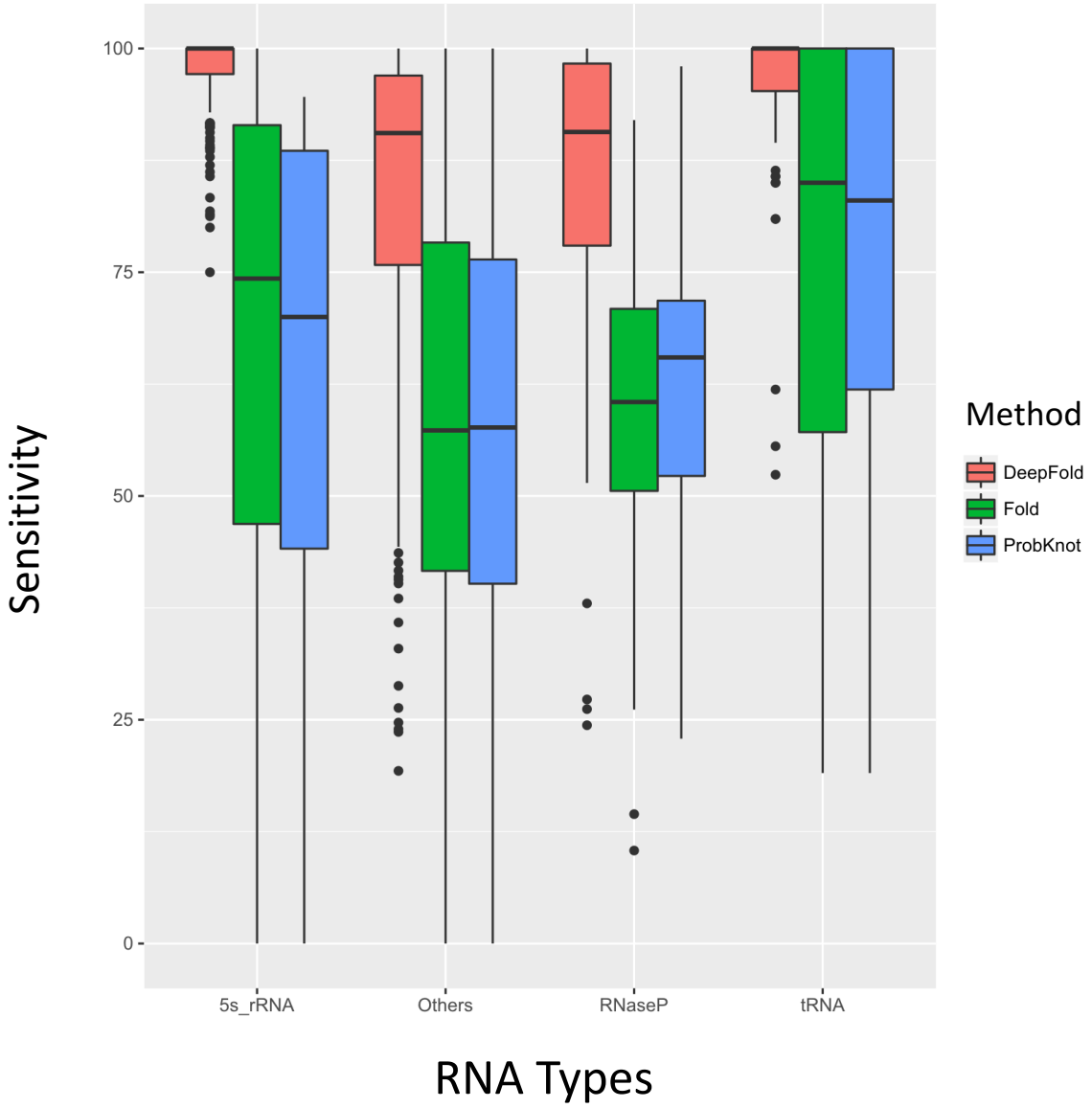
Stabilize the performance on negative samples and improve the performance on positive samples.

$$P(S^+|T^+) = \frac{P(T^+|S^+)P(S^+)}{P(T^+)}$$

$$P(S^+|T^+) = \frac{P(T^+|S^+)P(S^+)}{P(T^+|S^+)P(S^+)+P(T^+|S^-)P(S^-)} = \frac{P(T^+|S^+)\frac{1}{79}}{P(T^+|S^+)\frac{1}{79}+P(T^+|S^-)\frac{78}{79}}$$

Predictive Power





All RNA sequences, do not filter out sequences with high similarities.

Acknowledgement

Thanks to Prof. *Zhi Lu*, *Bin-Bin Shi*, *Boqin Hu*, *Yang Li* and other Lulab members for their support and guidance. Thanks to all the advisors in Tsinghua *Xuetang* Talent program for their mentoring.

Thank you all for listening!