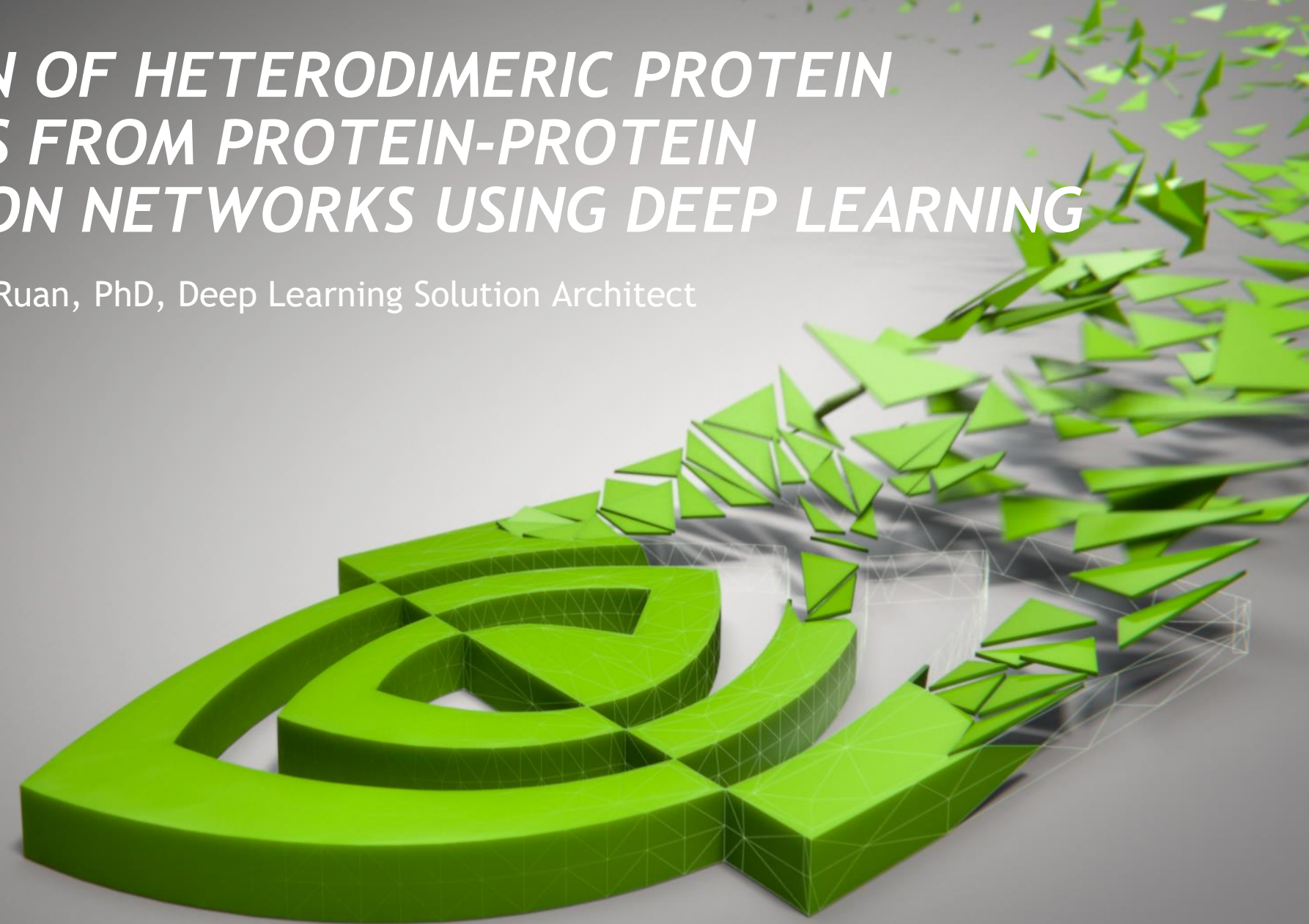
PREDICTION OF HETERODIMERIC PROTEIN COMPLEXES FROM PROTEIN-PROTEIN INTERACTION NETWORKS USING DEEP LEARNING

Peiying (Colleen) Ruan, PhD, Deep Learning Solution Architect

3/26/2018

NVIDIA



OUTLINE

Background

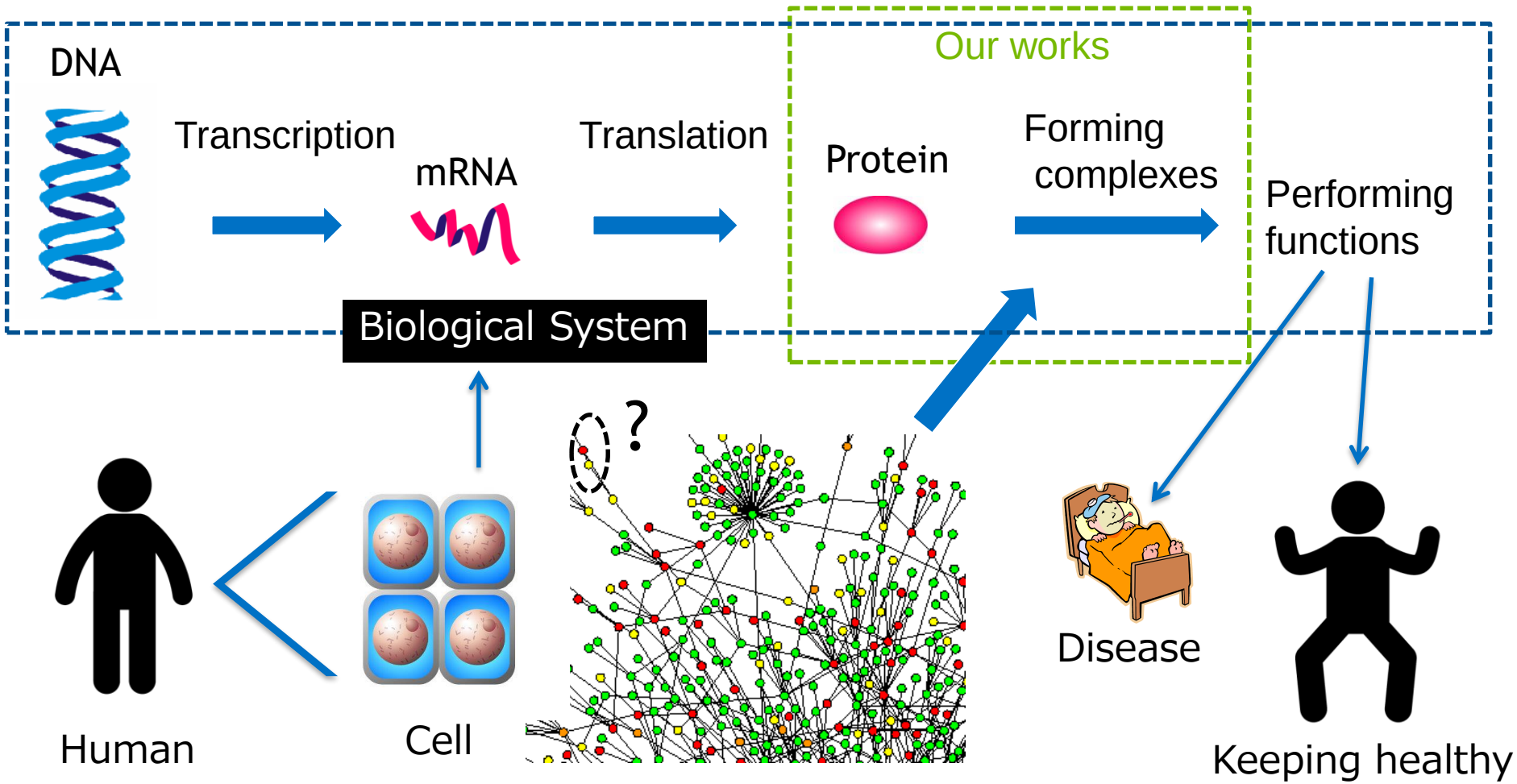
Method

Computational Experiments and Results

Conclusions

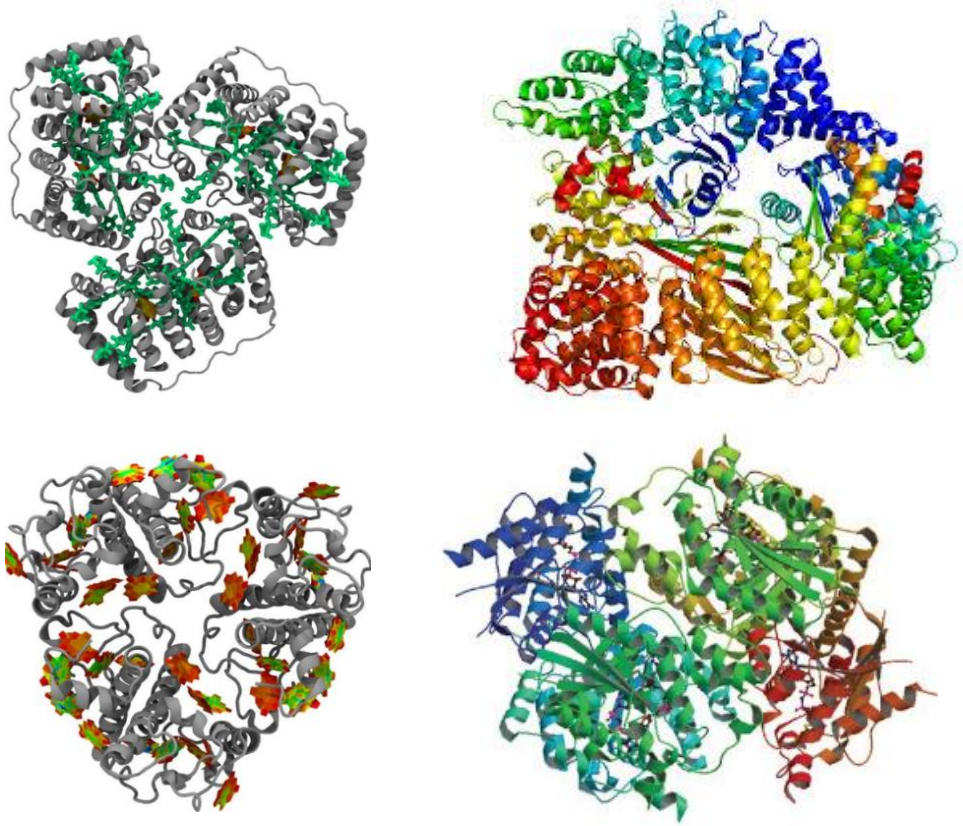
BACKGROUND

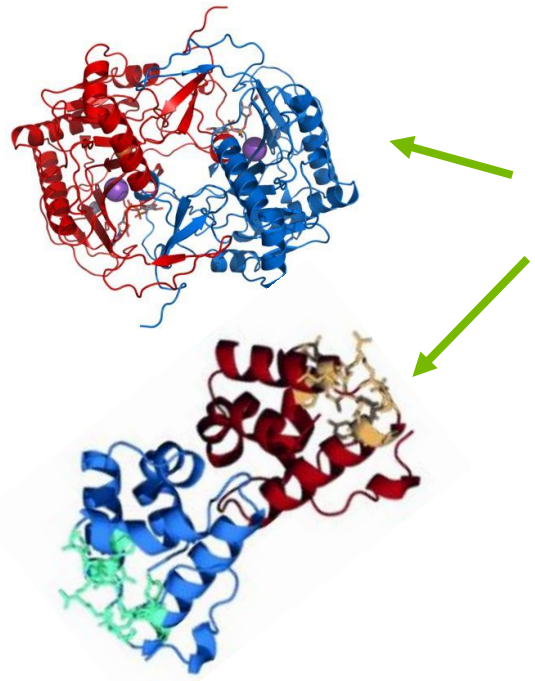
BACKGROUND


DNA
Transcription
mRNA
Translation
Protein
Forming complexes
Performing functions
Our works
Biological System
?
Human
Cell
Disease
Keeping healthy

BACKGROUND

What is heterodimer and why predict it?

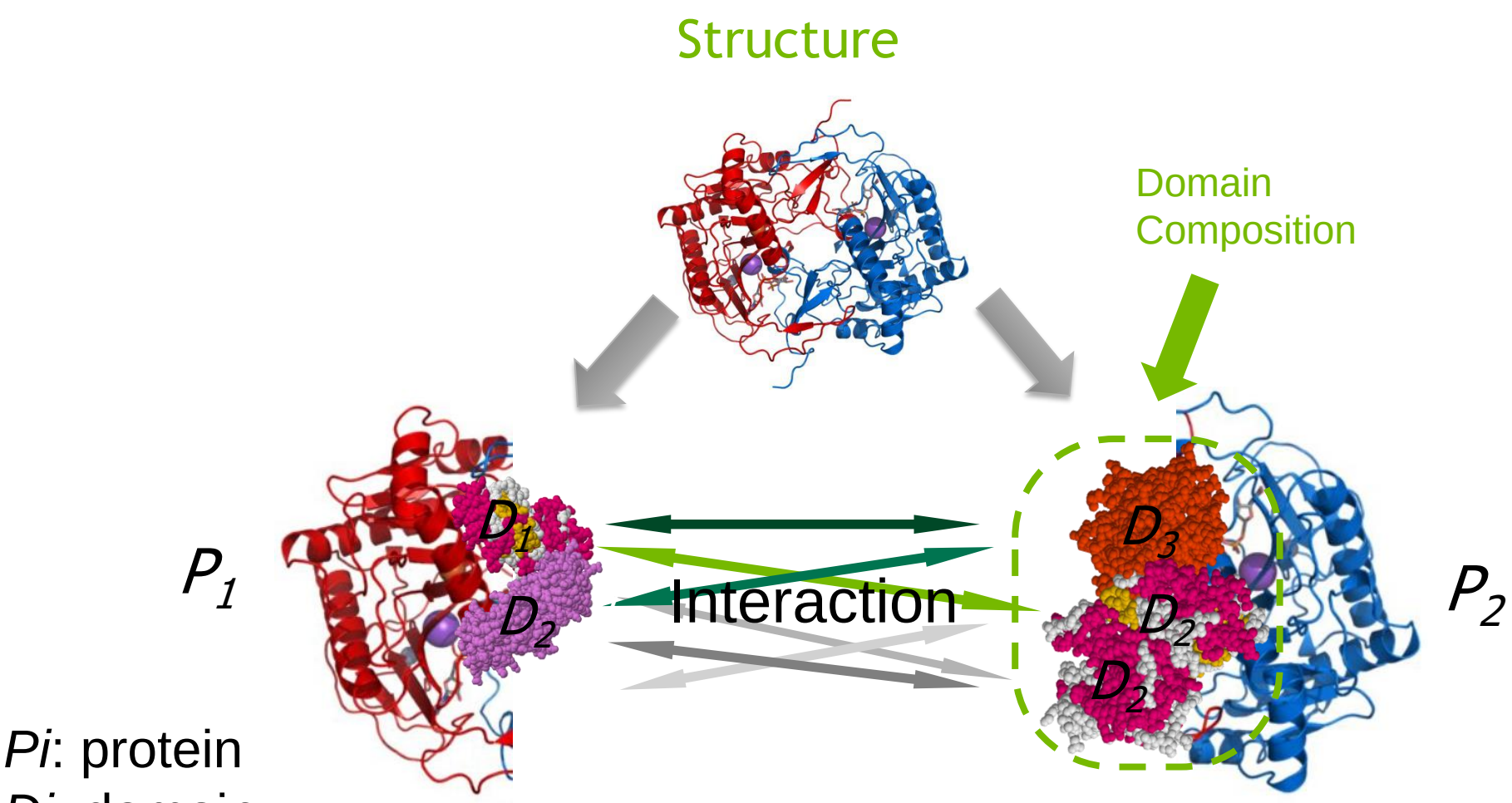




Heterodimers

Occupy 40% !!!

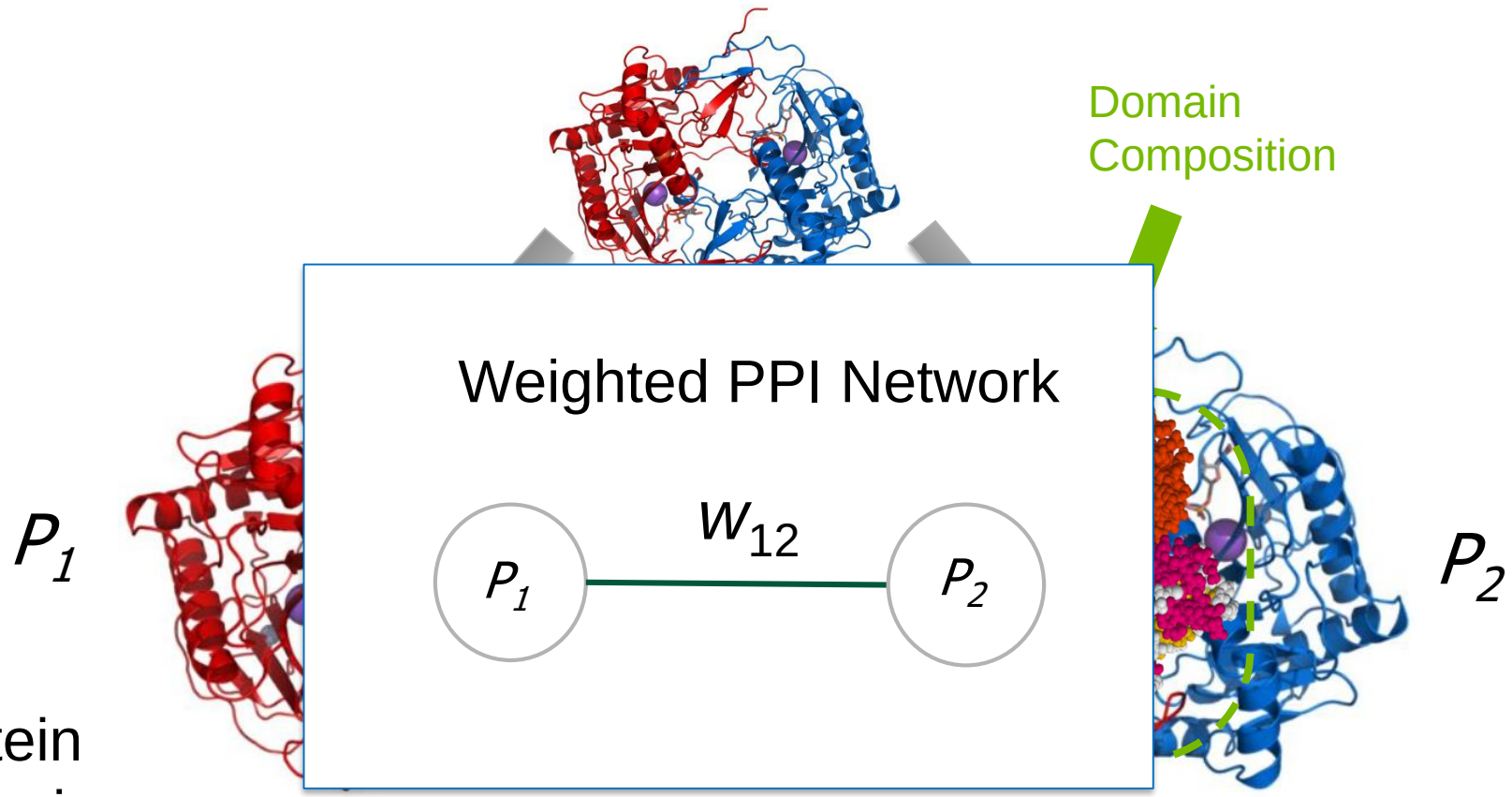
BACKGROUND



Pi: protein
Di: domain

BACKGROUND

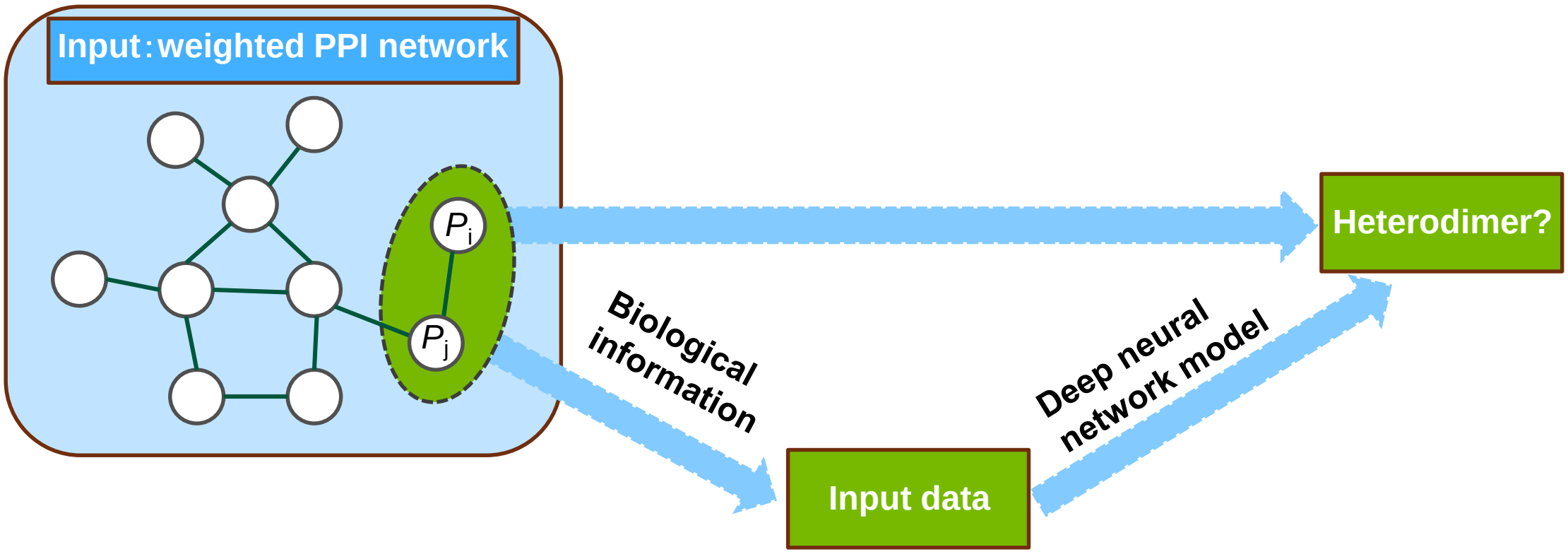
Structure



Pi: protein
Di: domain

METHOD

OVERVIEW OF THE PROBLEM



MULTIPLE INFORMATION + MULTIPLE DL MODELS

- Input data involving biological information
  - Protein-protein interaction (PPI)
  - Domain
  - Phylogenetic profile
- Deep neural network models including
  - Convolutional neural network (CNN)
  - Recurrent neural network (RNN)
  - CNN + RNN

PROTEIN-PROTEIN INTERACTION (PPI)

Table 1. Feature space mapping from two interacting proteins P_i, P_j and neighbors.

(F1)	w_{ij}
	The weights of interactions between the focused proteins.
	The maximum weights of interactions between either of focused proteins and a
	The minimum weights of interactions
	The maximum smaller weights of interactions
	The maximum differences of weights amon the neighboring weights.

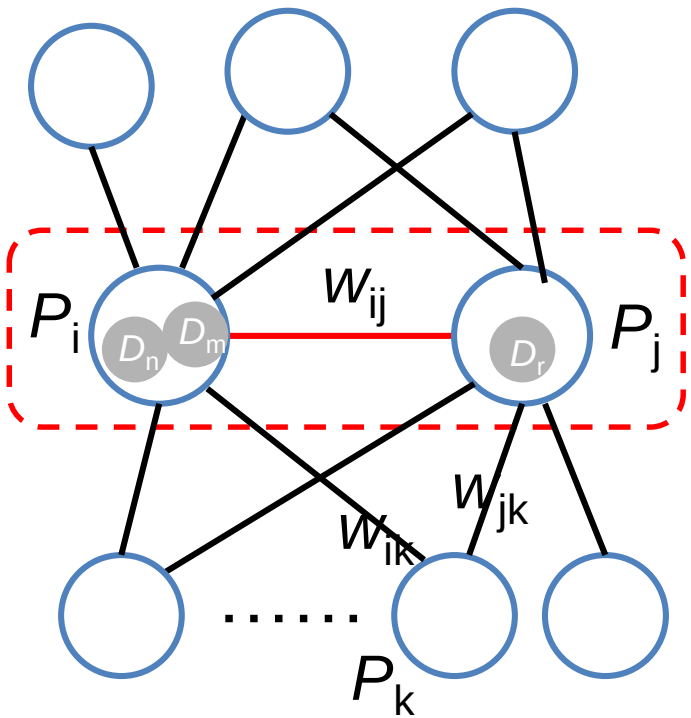
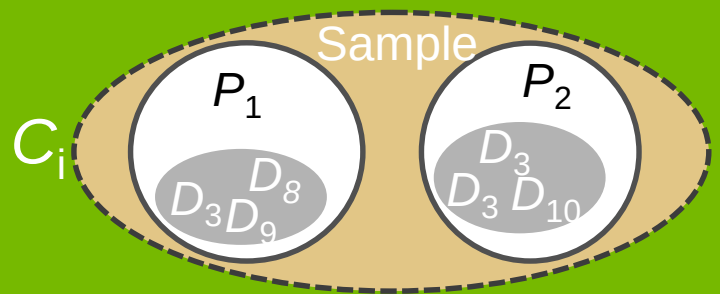


Figure 1. Example of a subgraph with an interacting protein pair and their neighboring proteins.

DOMAIN



Domain pair of protein complex C_j:
(D_3, D_3), (D_3, D_3), (D_3, D_{10}), (D_8, D_3) , (D_8, D_3) ,
(D_8, D_{10}) , (D_9, D_3) , (D_9, D_3) , (D_9, D_{10})

The whole domain pair sets for all complexes in the dataset

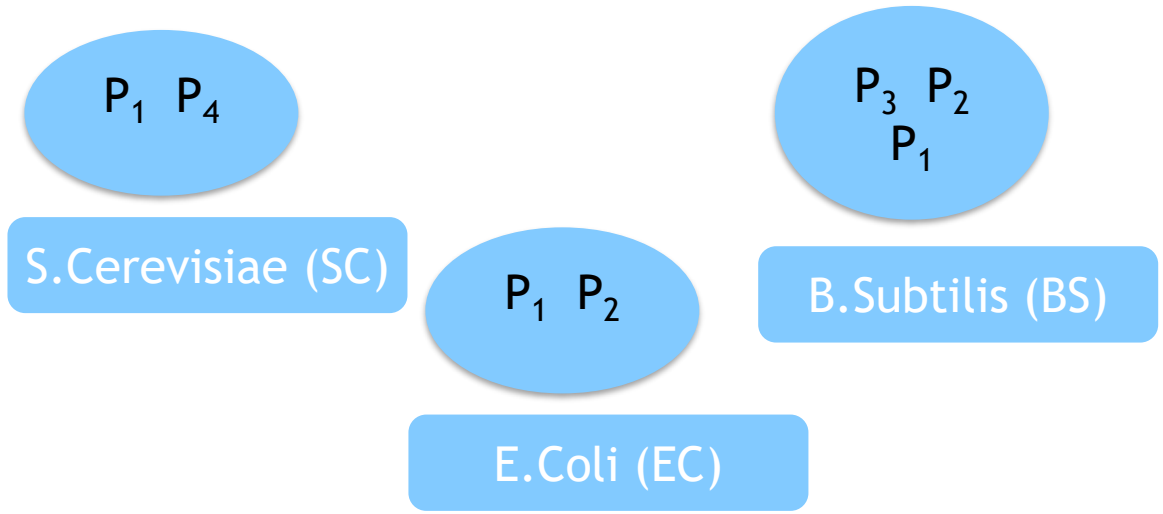
$$\{(D_1, D_1), (D_1, D_2), \ldots, (D_3, D_3), \ldots, (D_9, D_{10}), \ldots, (D_n, D_n)\}^{5295}$$

#domain pair is 5295

$$[C_j]=[\ 0 \quad 0 \quad \ldots, \quad 2 \quad \ldots, \quad 1 \quad \ldots, \quad 0\ ]$$

PHYLOGENETIC PROFILE

	SC	BS	EC
P_1	1	0	1
P_2	0	1	1
P_3	0	1	0
P_4	1	1	0



The whole organism for all complexes in the dataset

$$\{ SC,\ BS,\ EC,\ \ldots \}^{2717}$$

$$[C_j = Q(P_1, P_2)] = [\ 0\quad 0\quad 1\ ,\ \ldots\ ]$$

$$Q(a, b) = \min(a, b)$$

#organism is 2717

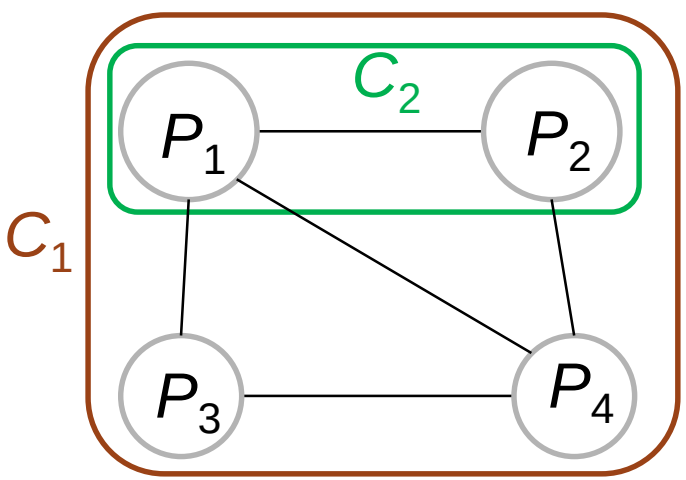
COMPUTATIONAL EXPERIMENTS

Databases

CYC2008: A manually curated comprehensive catalogue of yeast protein complexes, including **172(42%)** heterodimers.

WI-PHI: A PPI database with weights containing **49607** interacting protein pairs except self-interactions.

Positives and Negatives



- Positives: (P_1,P_2)
- Negatives: (P_1,P_3), (P_2,P_4), (P_3,P_4) and (P_1,P_4)

- #Sample: 5497

INPUT DATA

e.x.Domain property

The whole domain pair set for all complexes in the dataset

$$\{(D_{1,}\ D_1),\ (D_{1,}\ D_2),\ldots,\ (D_{3,}\ D_3),\ldots,\ (D_{9,}\ D_{10}),\ldots,\ (D_{n,}\ D_n)\}$$

Input data:

Label:

$[C_1]$=[0 0 ,…, 2 ,…, 1 ,…, 0] *0*

$[C_2]$=[0 1 ,…, 0 ,…, 0 ,…, 1] *1*

… …

$[C_{5497}]$=[0 0 ,…, 2 ,…, 1 ,…, 0] *0*

INPUT DATA

e.x. Domain + Phylogenetic profile

The whole (domain pair set + organism) for all complexes in the dataset

$$\{(D_1, D_1), (D_1, D_2), \ldots, (D_n, D_n), SC, BS, EC, \ldots\}^{5295+2717}$$

Input data:

Label:

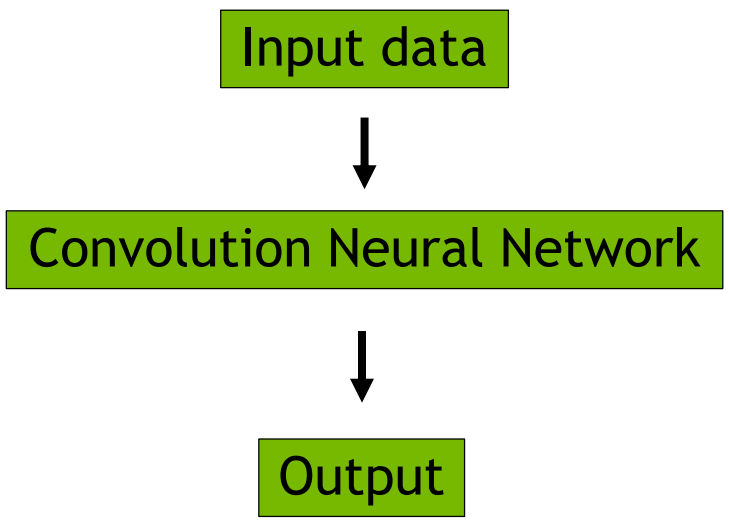
$[C_1]$=[0 0 ,…, 0 , 0, 0, 1, …] 0

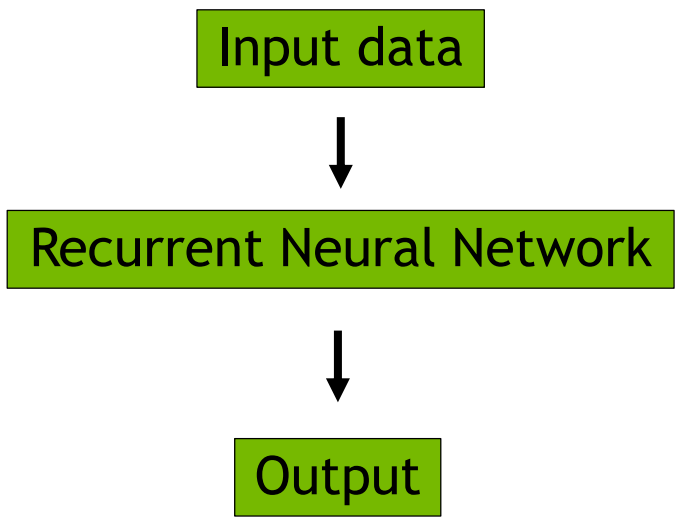
$[C_2]$=[0 1 ,…, 1 , 1, 0, 0, …] 1

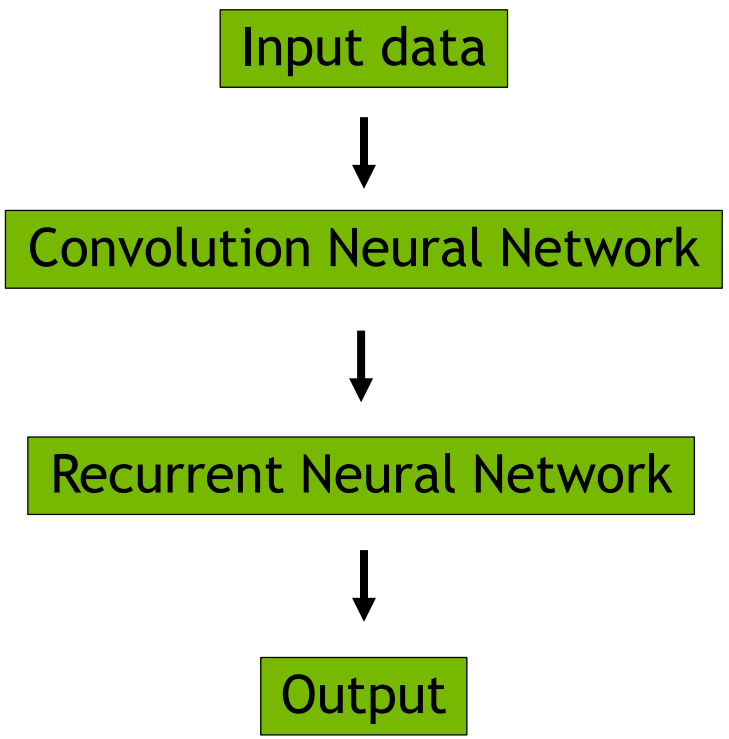
… …

$[C_{5497}]$=[0 0 ,…, 0 , 0, 1, 1, …] 0

MODELS







D. Quang et al., DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences, *Nucleic Acids Research,* 2016

RESULTS

PERFORMANCE MEASURES

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

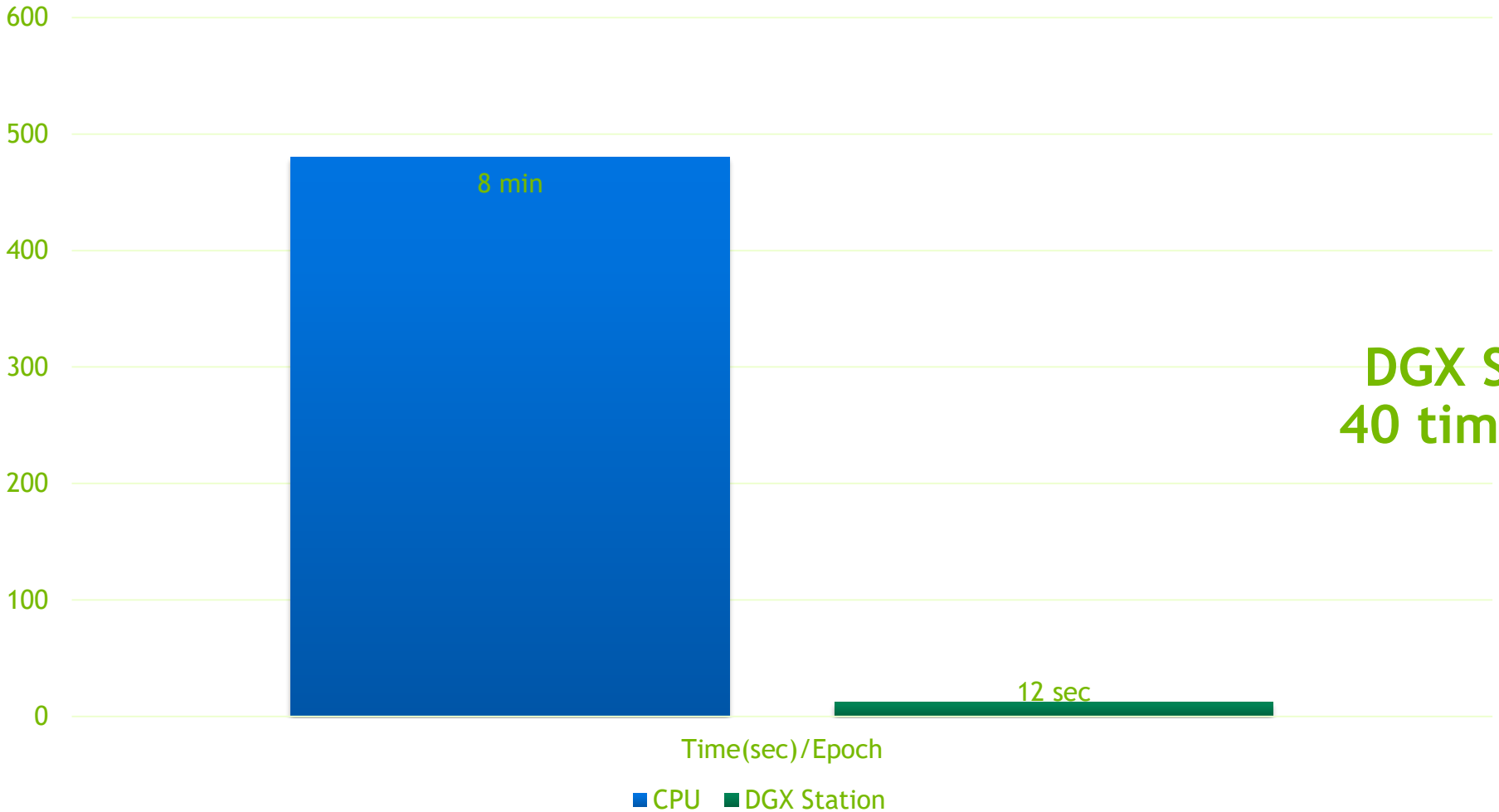
tp: true positive, tn: true negative, fp: false positive, fn: false negative

COMPARISON OF MODEL + INFORMATION

Models	Training accuracy	Training loss	Test accuracy	Evaluation score (F1)
CNN (domain)	0.80	1.311	0.79	0.68
CNN (domain+PPI)	0.84	1.124	0.81	0.69
CNN (domain+PPI+Phylogenetic profile)	0.83	0.912	0.81	0.69
RNN (domain+PPI+Phylogenetic profile)	0.71	2.334	0.72	0.66
CNN+RNN (domain+PPI+Phylogenetic profile)	0.86	0.865	**0.85**	**0.72**
Baseline method* SVM(PPI+domain)	0.65	-	**0.73**	**0.63**

*P. Ruan et al. Prediction of Heterodimeric Protein Complexes from Weighted Protein-Protein Interaction Networks Using Novel Features and Kernel Functions, *PLoS One*, 2013

CPU VS GPU



DGX Station is 40 times faster!!

CONCLUSIONS

- Applied deep learning to predicting heterodimeric protein complexes with multiple biological information
- The performance of hybrid model with multiple information is better than single model
- The speed of DGX station is 40 times faster than CPU

Thank you for your kind attention!

Email: cruan@nvidia.com