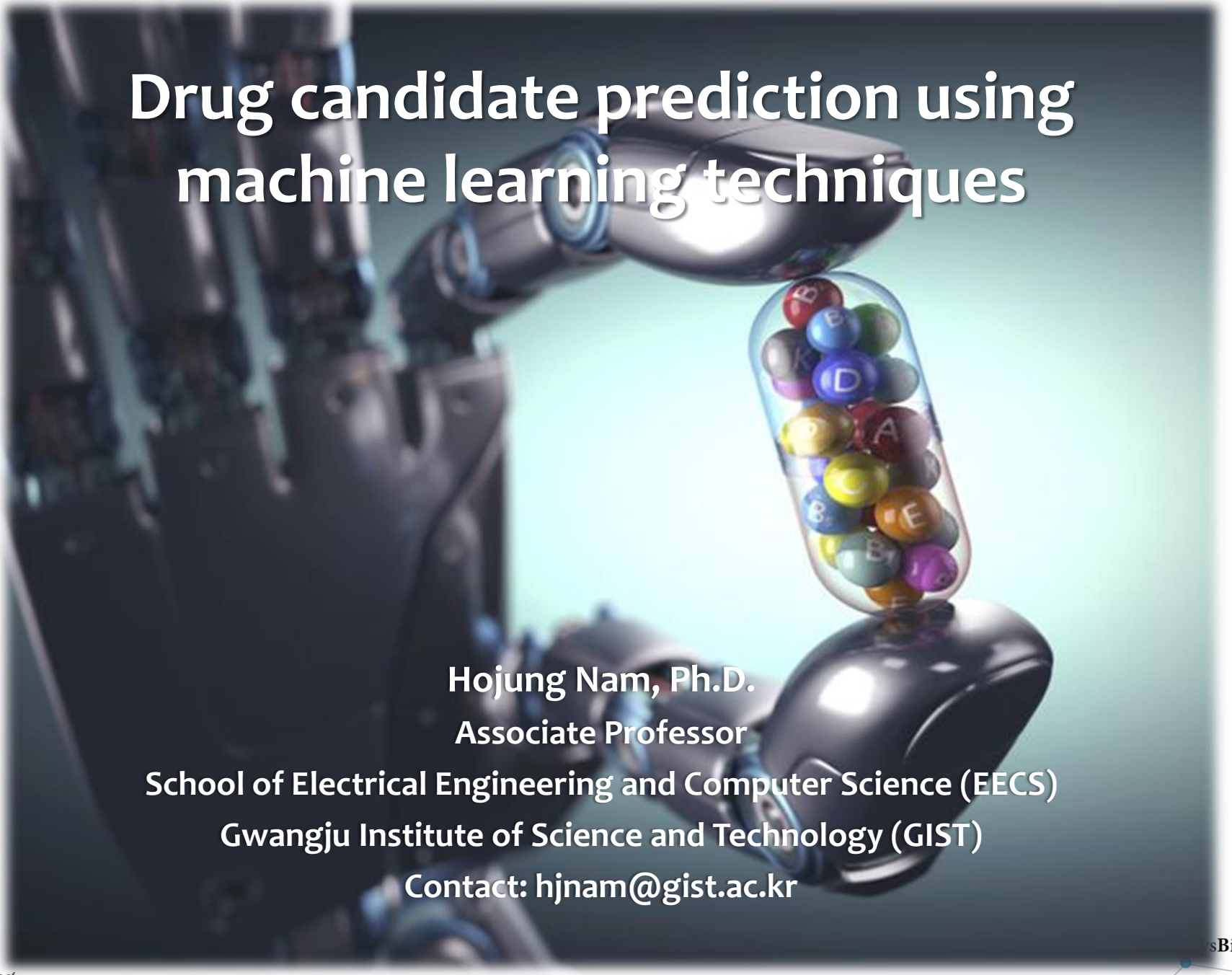
Drug candidate prediction using machine learning techniques



Hojung Nam, Ph.D.

Associate Professor

School of Electrical Engineering and Computer Science (EECS)

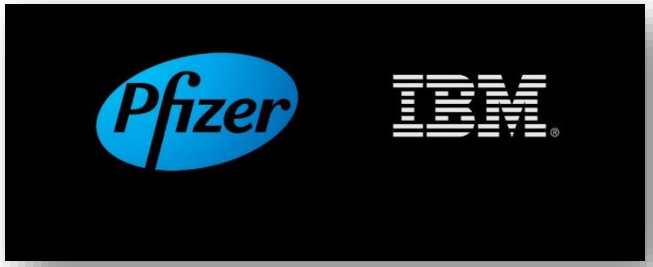
Gwangju Institute of Science and Technology (GIST)

Contact: hjnam@gist.ac.kr

Agenda

- Backgrounds - Drug development process
- Prediction of drug-target interactions using deep neural networks model

Pharmaceutical company



Pfizer & IBM Watson
: IBM and Pfizer to Accelerate Immuno-oncology Research with Watson for Drug Discovery

Janssen & BenevolnetAI (2016)
: BenevolentAI signs exclusive license agreement with Janssen for clinical-stage drugs

GSK & Exscientia (2017)
: GSK Launches Up-to-$43M AI-Focused Collaboration with Exscientia

Why NOW? Drug discovery?

Why is Deep Learning Hot Now?

Big Data Availability

facebook 350 millions images uploaded per day

Walmart 2.5 Petabytes of customer data hourly

YouTube 300 hours of video uploaded every minute

New ML Techniques

GPU Acceleration

CPU

- Few complex cores
- Specialized in Serial processing

GPU

- Many simple cores
- Built for Parallel processing (ex. Image)

CPU	Internal Structure	GPU
4~8	Cores	3000+
3~4 GHz	Clock Speed	~1.5 GHz
~1000 GFLOPS	Throughput	~10000 GFLOPS*

Gwangju Institute of Science and Technology

* FLOPS : Floating-Point Operations Per Second

ComSysBioLAB

TOP 3 HIGH-END NVIDIA GPUs

Name	GeForce GTX 1080 Ti	NVIDIA TITAN Xp	NVIDIA TITAN V
Architecture	Pascal	Pascal	Volta (Brand New)
CUDA cores	3584	3840	5120
Clock Speed	1582 MHz	1582 MHz	1455 MHz
VRAM Size	11GB GDDR5X (bus interface : 352 bit)	12GB GDDR5X (bus interface : 384 bit)	12GB HBM2[1] (bus interface : 3072 bit)
Memory Bandwidth	484.4 GB/s	547.6 GB/s	652.8 GB/s
Price	$699 USD	$1,199 USD	$ 2,999 USD
Tensor Cores[2]	n/a	n/a	640

[1] HBM2 memory is composed of higher bus interface than GDDR5X.
[2] Exclusive optimized cores for Deep learning operations.

Gwangju Institute of Science and Technology

ComSysBioLAB

Drug discovery?

Why is Deep Learning Hot Now?

Big Data Availability

New ML Techniques

GPU Acceleration

DATA DRIVEN DRUGS

Big Data ! (chemical compound, target)

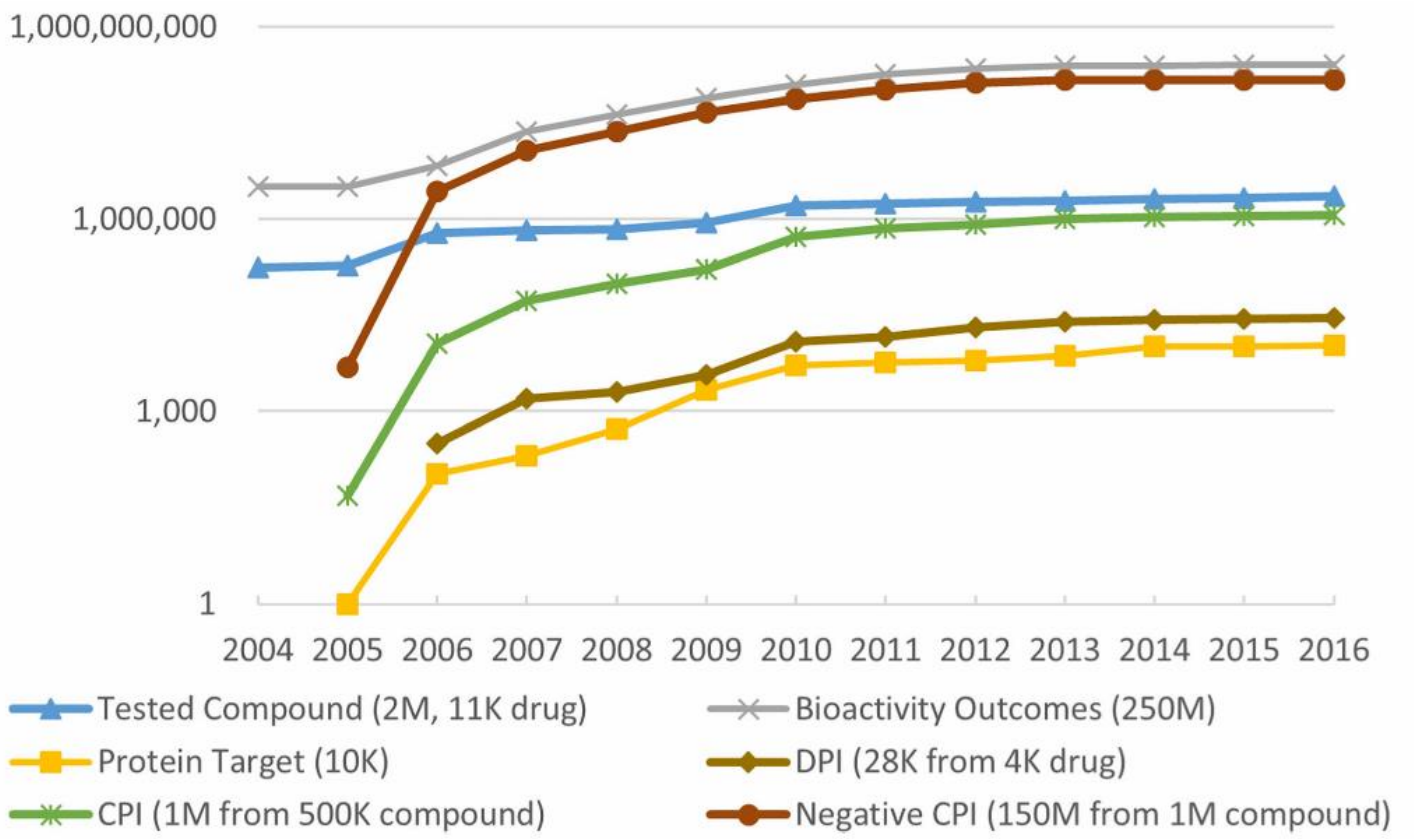


Fig. 2. The growth of biological data in PubChem BioAssay including biologically tested compounds, bioactivity outcomes, protein targets, drug-protein interactions (DPIs), compound-protein interactions (CPIs), negative compound-protein interactions (CPIs). The number in parenthesis is the total count of each data category. DPI and CPI are counted based on the confirmatory and literature-based assays

Cheng et al., The AAPS journal 2017

Big Data ! (beyond the compounds)

Opinion

Trends in Pharmacological Sciences September 2014, Vol. 35, No. 9

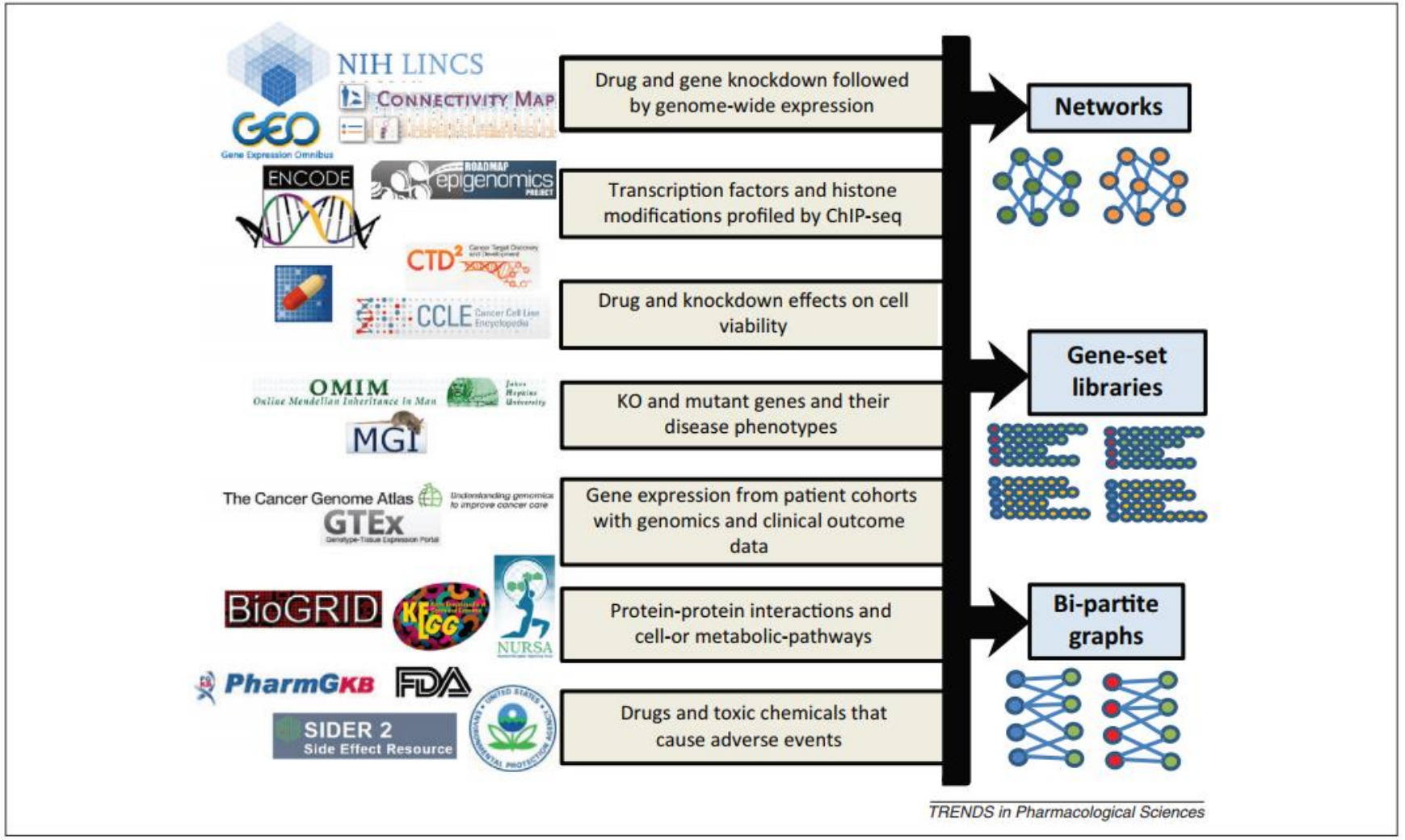
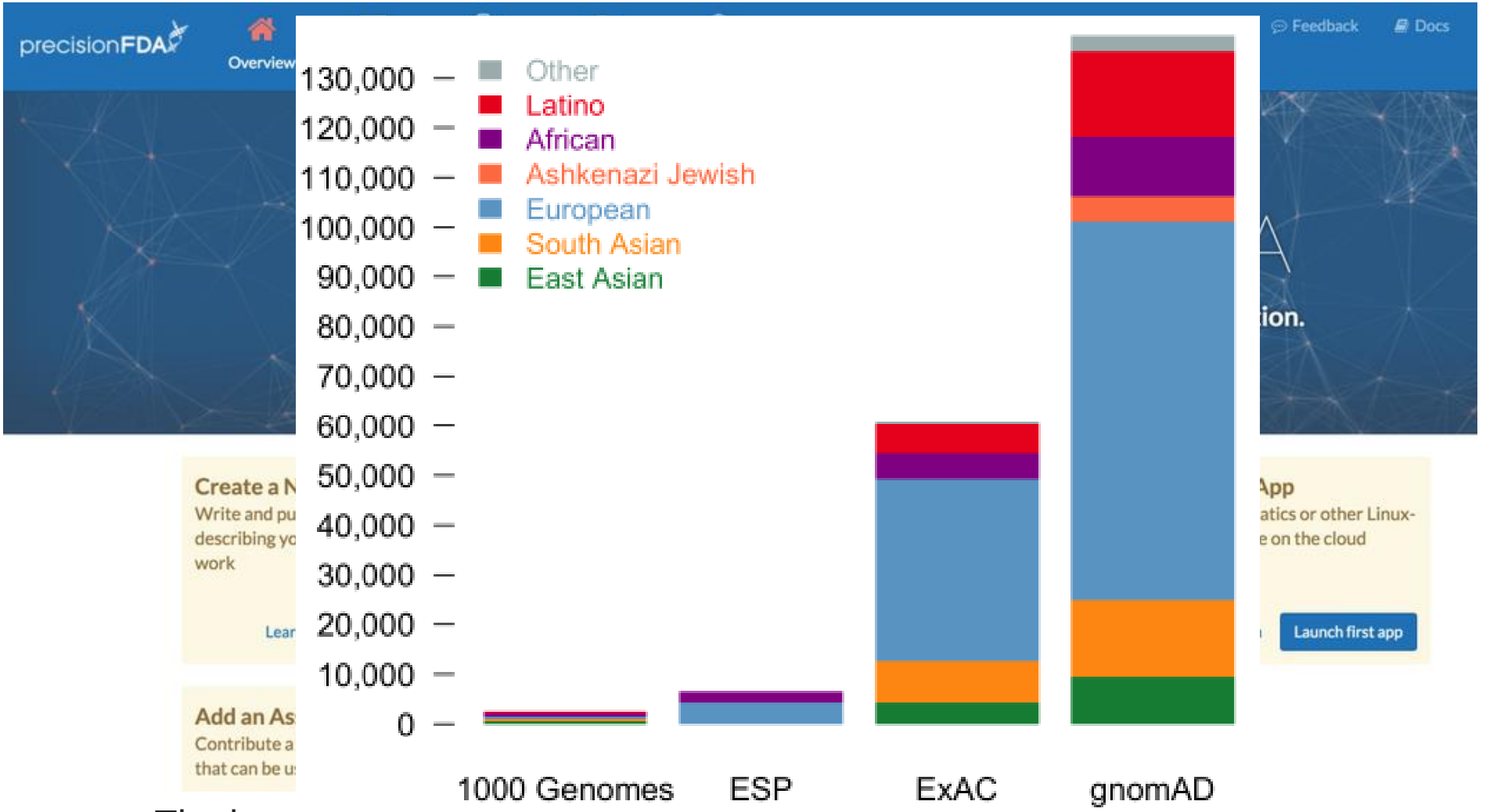


Figure 3. Resources from systems biology and systems pharmacology can be integrated by first identifying the various objects, their relations, and their data types and then converting the data into single-entity weighted networks, fuzzy-set libraries, or weighted multipartite graphs.

Considering individual genomes



The long-term goal of the platform is to streamline the process of evaluating tests leading to medical patients being able to get precise care based on their own individual genomic data

Gwangju Institute of Science and Technology

ComSysBioLAB

Drug discovery?

Why is Deep Learning Hot Now?

Big Data Availability

facebook 350 millions images uploaded per day

Walmart 2.5 Petabytes of customer data

YouTube hours of video uploaded every minute

New ML Techniques

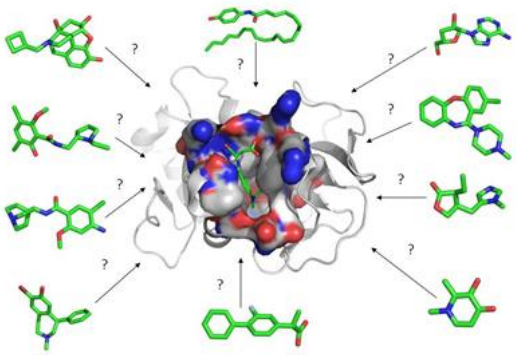
GPU Acceleration

Ingoo Lee, Jongsoo Keum, Hojung Nam*, "DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences", *Bioinformatics*, Under review.

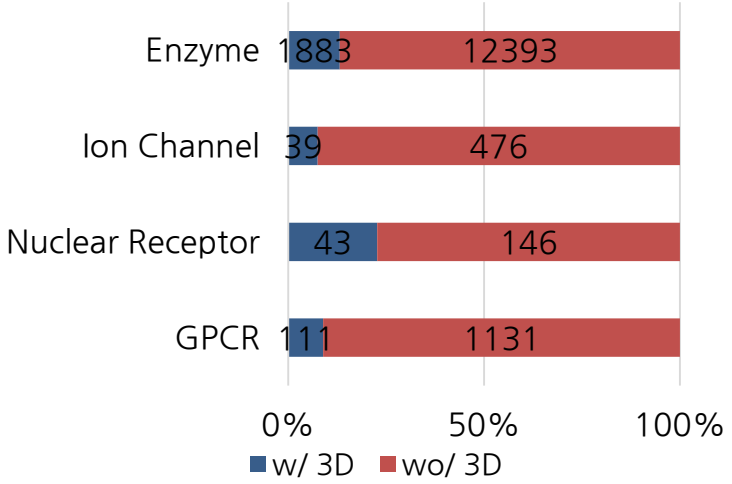
DRUG-TARGET INTERACTION

in-silico based DTI approaches

docking-based



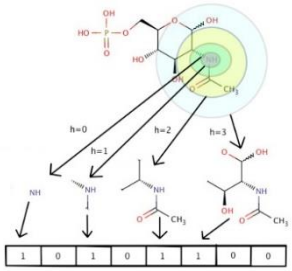
Need 3D structure



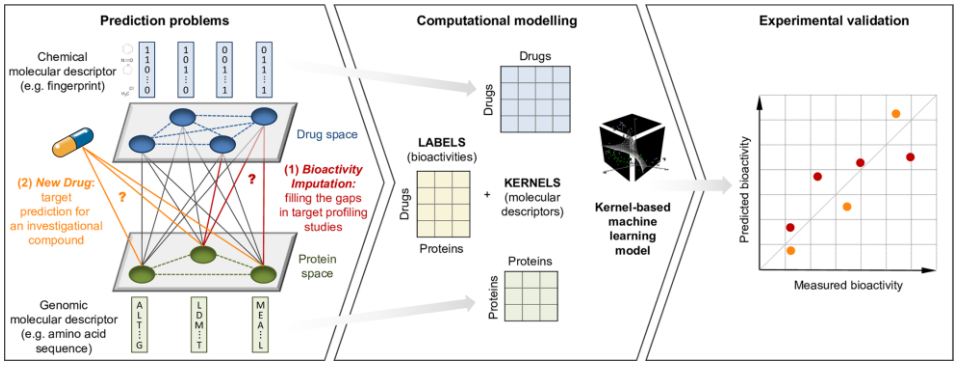
Can be applied to limited proteins
High Run time complexity

Machine learning based

Drug descriptor



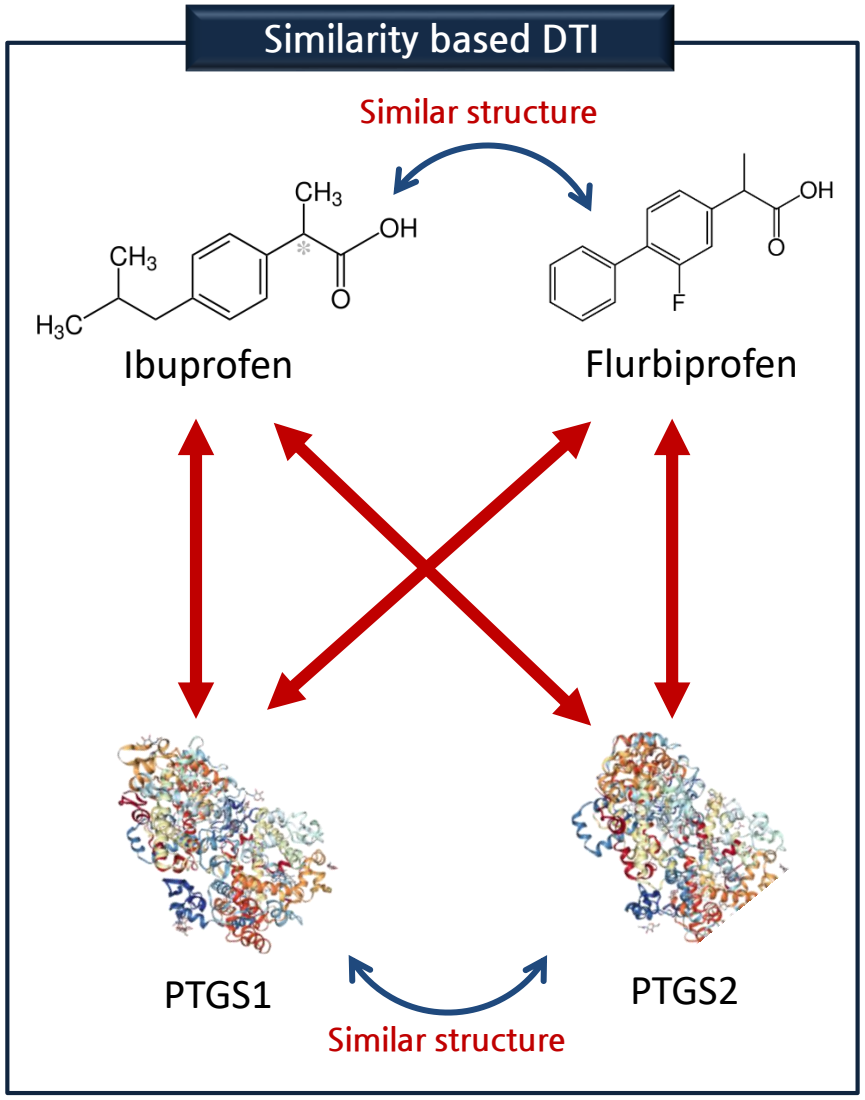
Protein descriptor

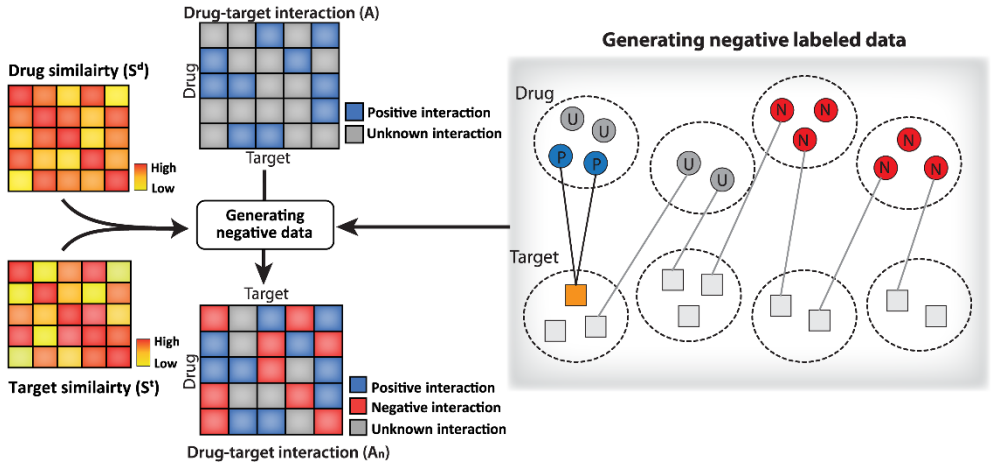


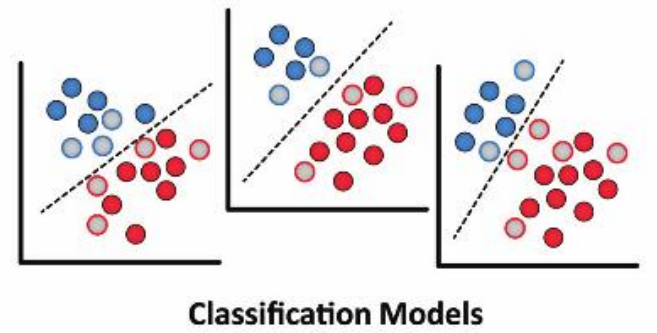
PLoS computational biology 13.8 (2017): e1005678

Can be applied to all general proteins
Low run time complexity

Similarity-based method



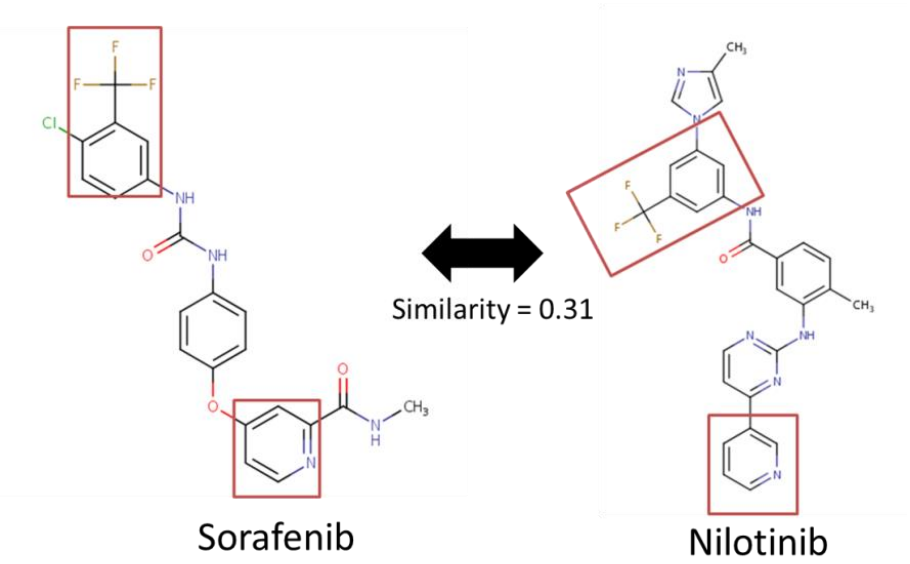




Jongsoo Keum, Hojung Nam*, "SELF-BLM: Prediction of drug-target interactions via self-training SVM", PLoS One, 2017

Problems of Similarity-based method

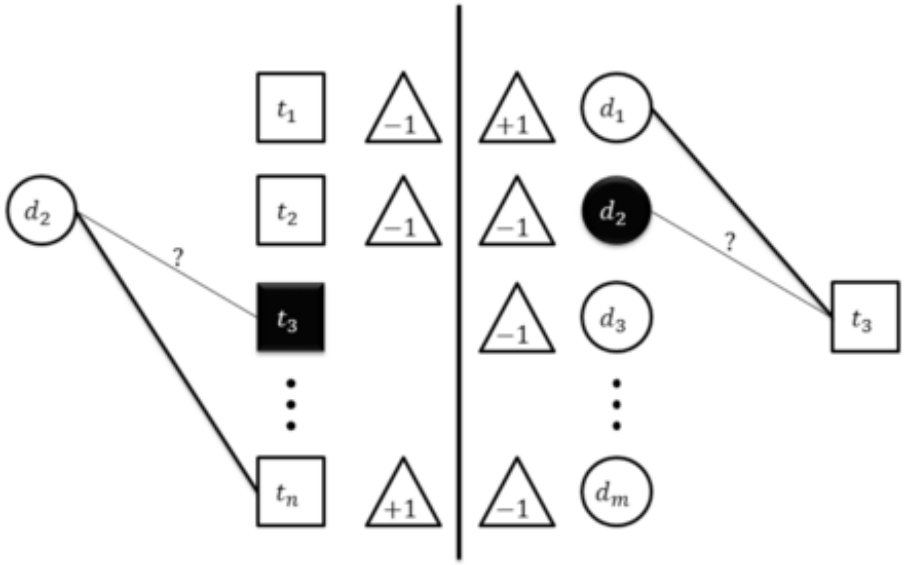
- Miss prediction



- Common important substructures, *but low similarity because of their proportion*

- Does not work well (**Low performance**)

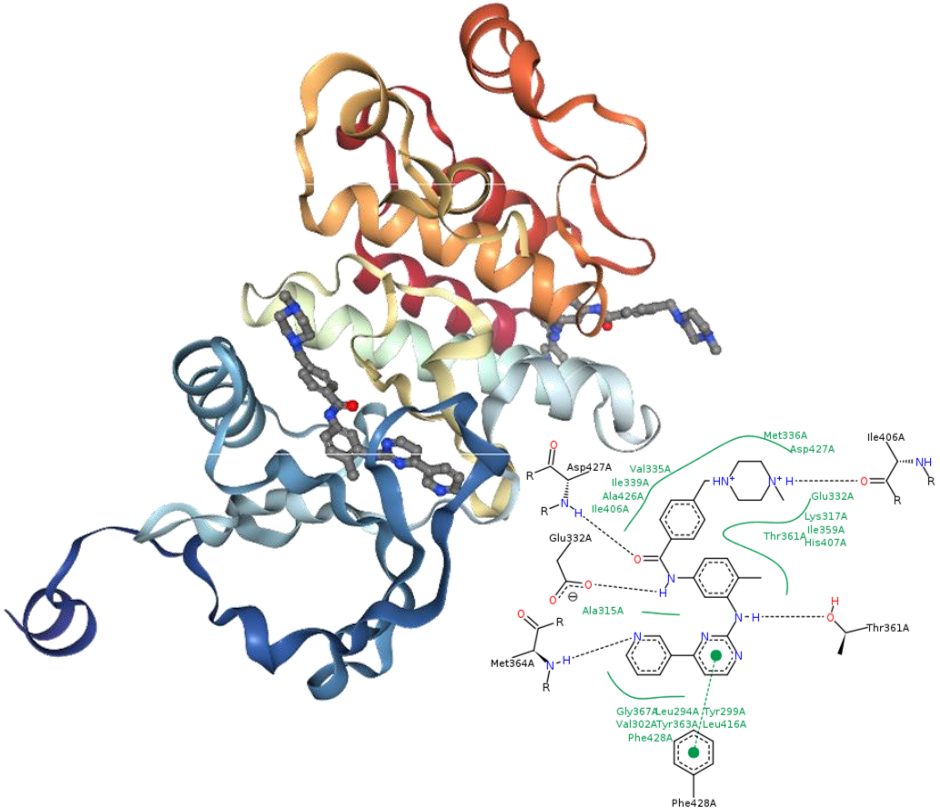
- High time complexity



- Large scale data for performance
- Time complexity: $O(n_d n_t + n_t n_d) \approx O(n^2)$

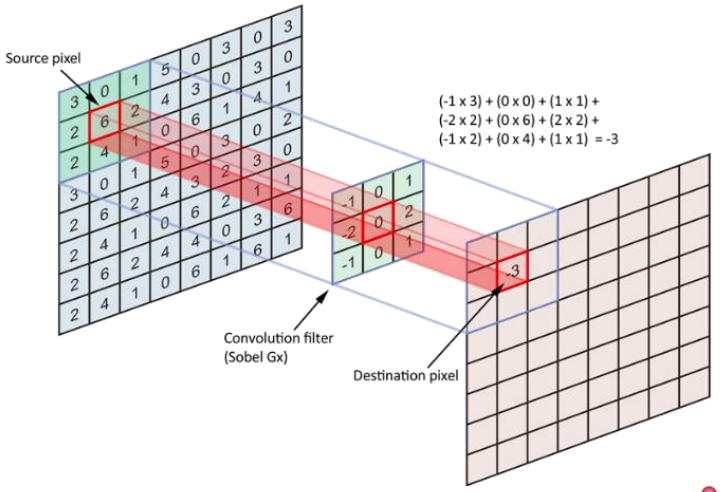
- Hard to train (**High time cost**)

DeepConv-DTI : CNN based model to detect binding regions



- 3D co-crystal structure of gleevec and its interaction with target protein

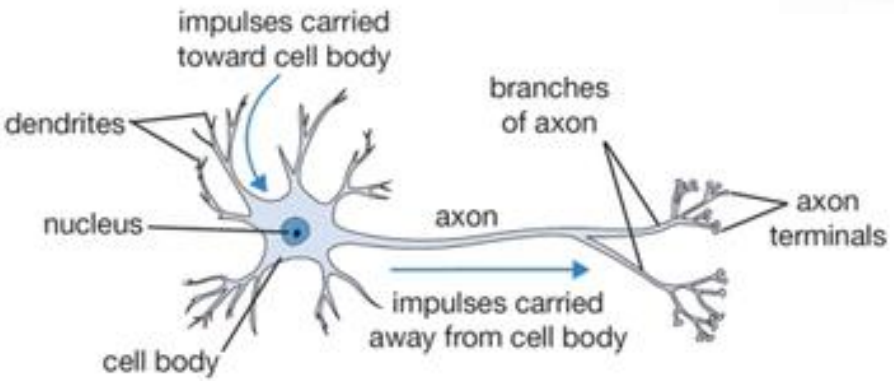
- **Binding region of target protein have a pattern to interact with drug**
- We can use patterns of binding regions to predict DTI for machine learning model
- Recently, **convolutional neural network** have received attraction to **extract local patterns**

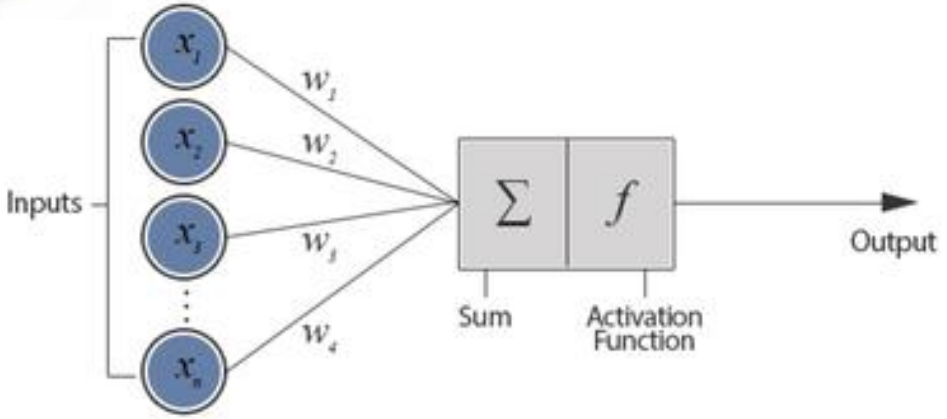


Gwangju Institute of Science and Technology

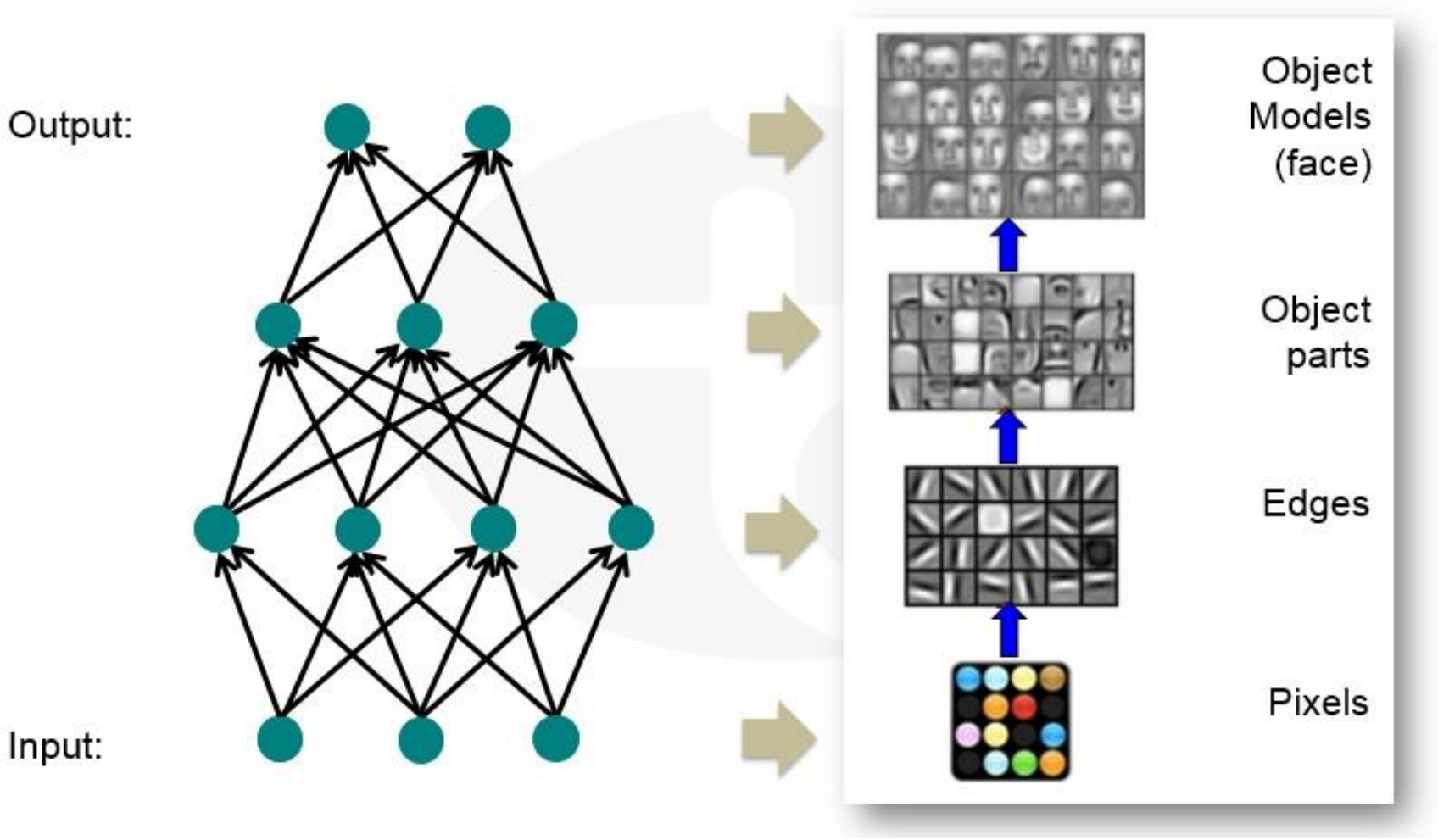
ComSysBioLAB

Neuron vs. Perceptron

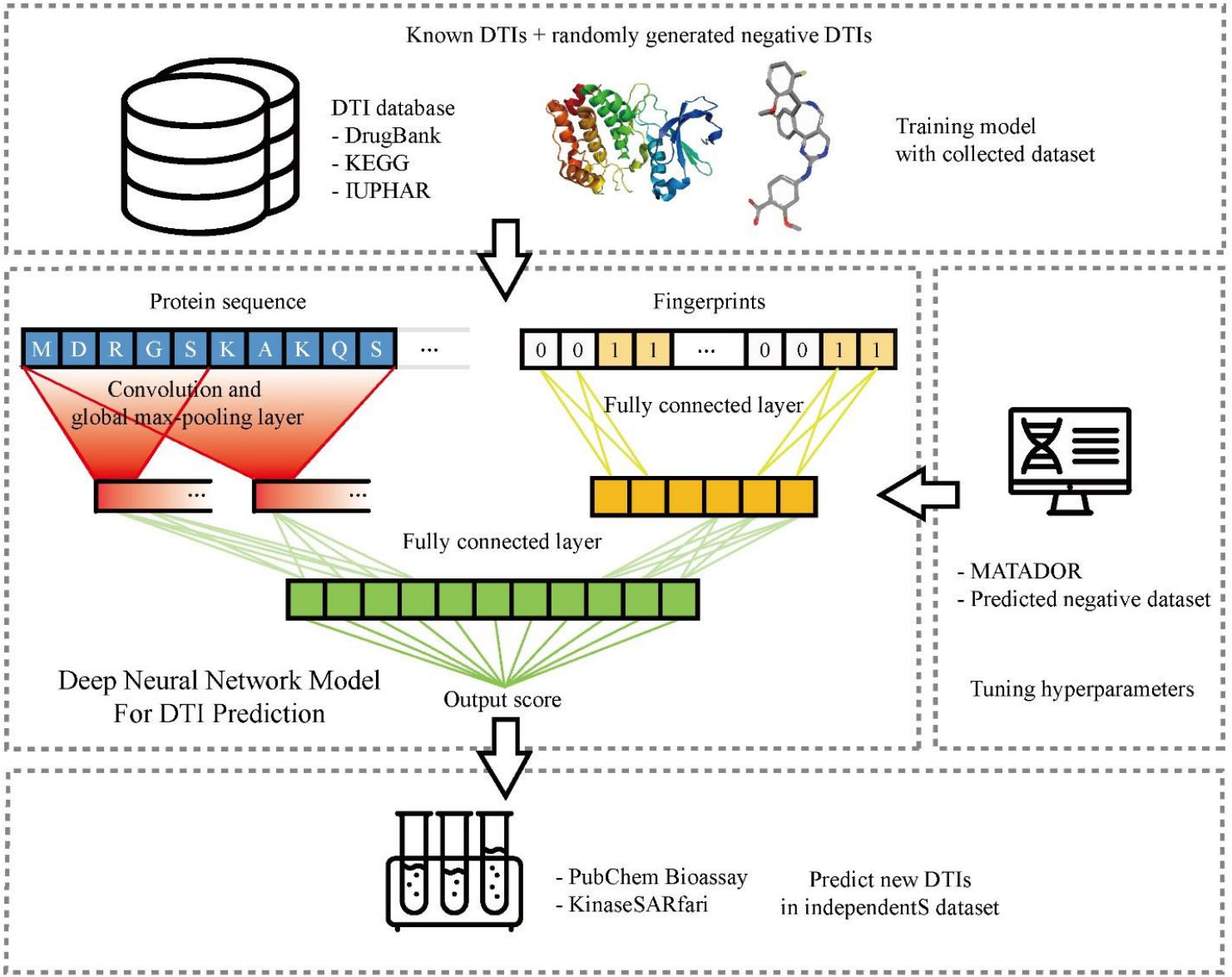




Deep Neural Network

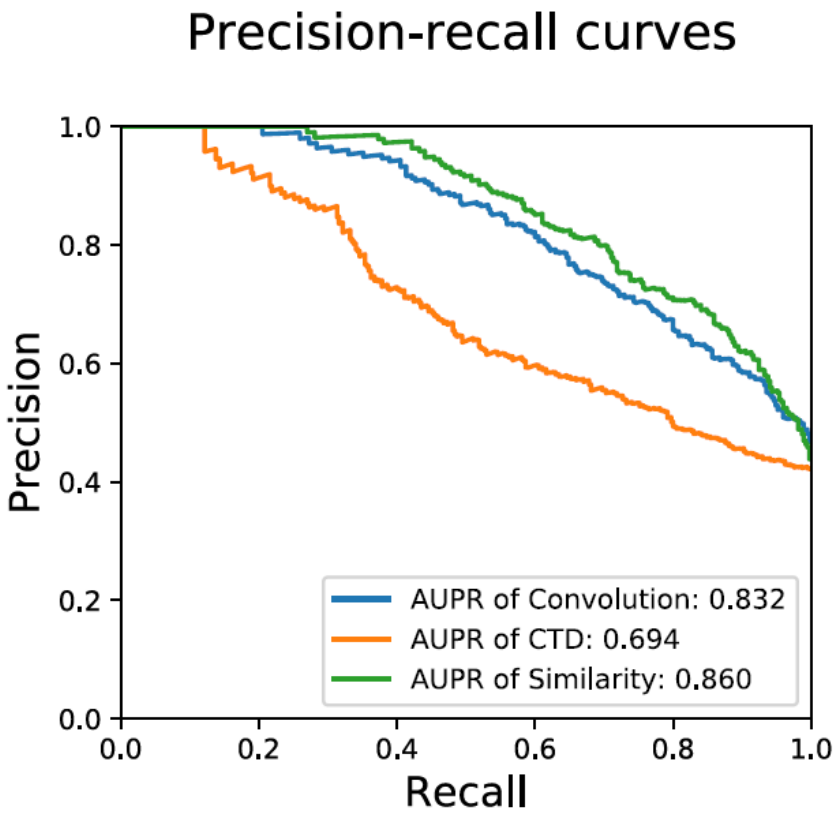


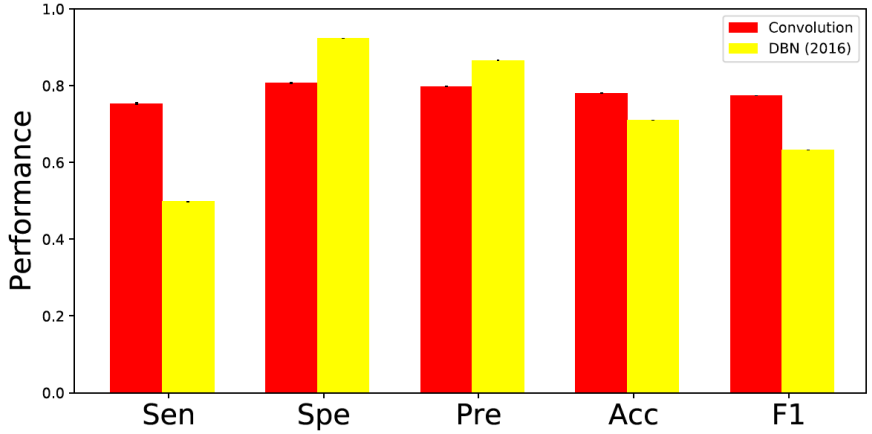
Overview


Known DTIs + randomly generated negative DTIs
DTI database
- DrugBank
- KEGG
- IUPHAR
Training model
with collected dataset
Protein sequence
M D R G S K A K Q S ...
Convolution and
global max-pooling layer
Fingerprints
0 0 1 1 ... 0 0 1 1
Fully connected layer
Fully connected layer
Deep Neural Network Model
For DTI Prediction
Output score
- MATADOR
- Predicted negative dataset
Tuning hyperparameters
- PubChem Bioassay
- KinaseSARfari
Predict new DTIs
in independentS dataset

Performance evaluation

- Comparison with other protein descriptor (vs CTD, similarity-based)



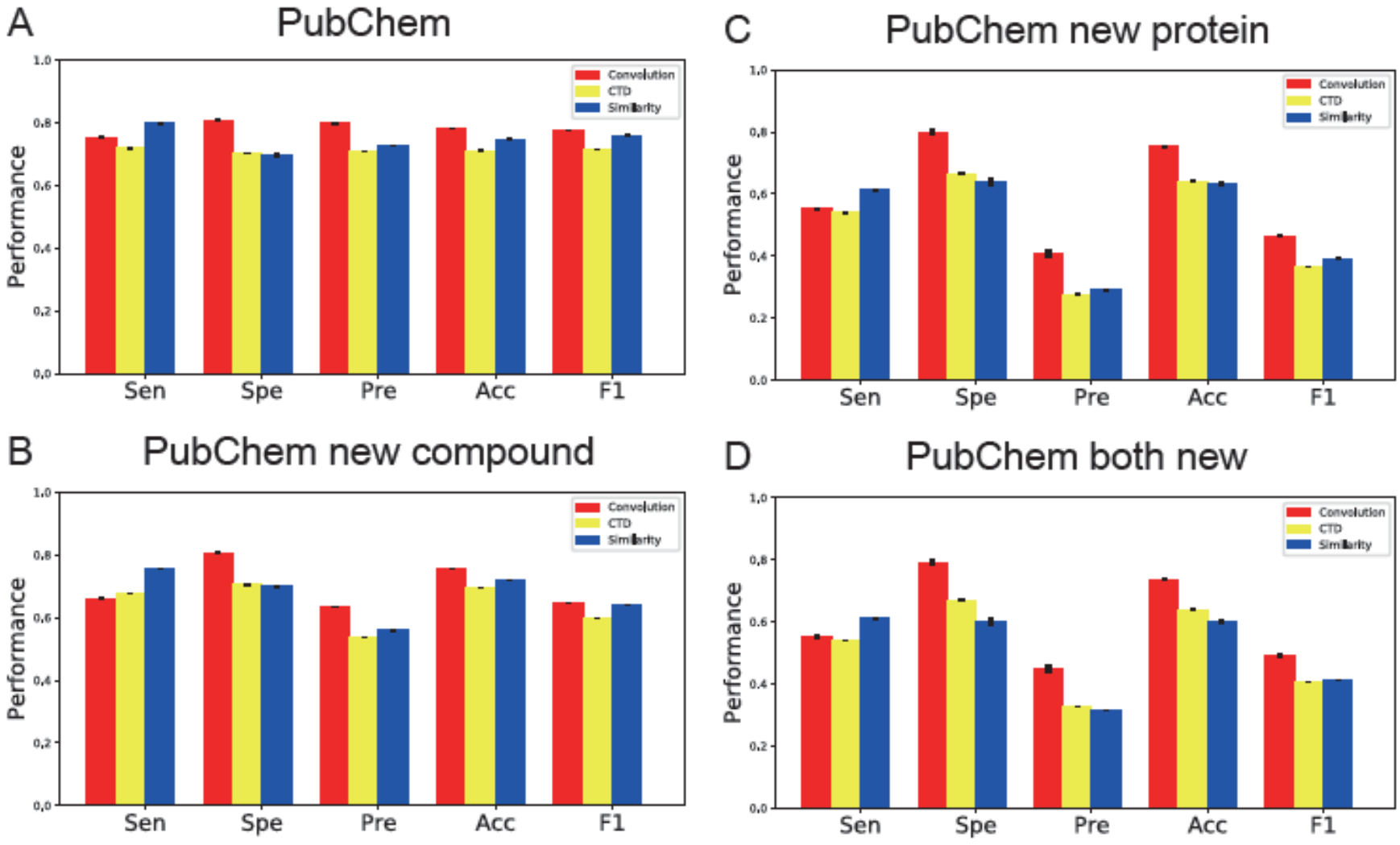


- Comparison with other DNN model (vs DBN)

Gwangju Institute of Science and Technology

ComSysBioLAB

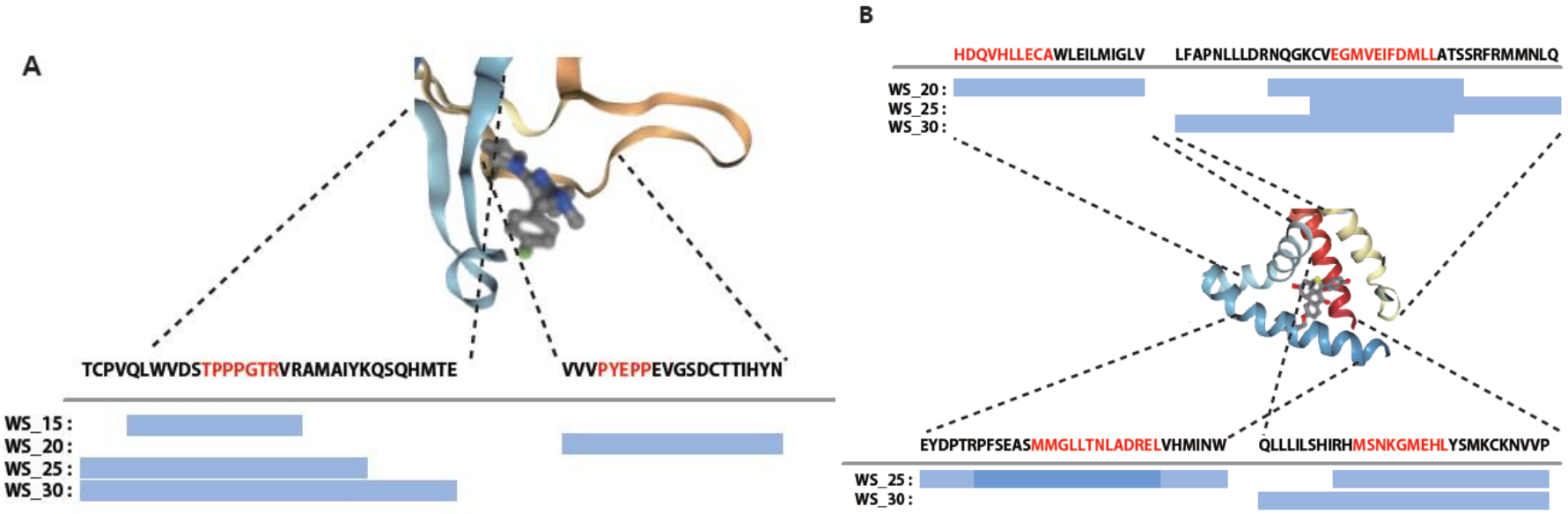
Performance evaluation



- Comparison with other protein descriptor (vs CTD, similarity-based)

Validation of extracted patterns

- Compare pooled convolution result with binding sites from sc-PDB



- cellular tumor antigen protein (P04637, P53_HUMAN)
- Estrogen receptor protein (P03372, ESR1_HUMAN)

- Pooled convolution results covers actual binding site

Summary (DTI)

- Propose a CNN based DTI prediction model
- The model show high prediction power in independent data sets
- The proposed model captures informative binding sites that contribute DTIs

Thank you
-Q&A-