KSBi-BIML 2026

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Recommendation Systems in Bioinformatics

송길태 _ 부산대학교

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 목적으로
제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우
발생하는 **모든 법적 책임은 행위자 본인에게 있음**을 알립니다.

KSBi-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics &Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의가 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

강의개요

Recommendation Systems in Bioinformatics

Netflix, YouTube 등에서 사용자의 흥미와 관심사에 대한 맞춤형 콘텐츠를 추천하는 인공지능 기법이 바로 추천 시스템이다. 본 강의에서는 이러한 추천 시스템에 대한 기본 개념을 이해하고 추천 시스템 기법이 Bioinformatics 분야에서 활용되는 사례를 배우는 것을 목표로 한다. 이를 통해 신약 개발 후보 물질 탐색이나 바이오 마커 발굴 등의 문제에 추천 시스템 기법을 적용할 있는 기초 역량을 갖추는 것을 목표로 한다.

강의는 다음의 내용을 포함한다.

- Recommendation systems 기본 개념 이해
  - Matrix factorization
  - Graph representation learning 기반 추천 시스템
- Recommendation systems적용을 통한 주요 bioinformatics 문제 해결
  - 표적 단백질 결합 후보 물질 탐색을 위한 recommendation systems
  - 바이오 마커 발굴을 위한 recommendation systems

* 강의 난이도: 중급

* 강의: 송길태 교수 (부산대학교 정보컴퓨터공학부)

Curriculum Vitae

Speaker Name: Giltae Song, Ph.D.

▶ **Personal Info**

Name	Giltae Song
Title	Associate Professor
Affiliation	Pusan National University

▶ **Contact Information**

Address	2 Busandaehak-ro 63 beon-gil, Geumjeong-gu
Email	gsong@pusan.ac.kr

Research Interest

Machine Learning, Computational Genomics, AI in Drug Discovery and Precision Medicine

Educational Experience

1999	B.S. in Computer Science, Seoul National University, South Korea
2001	M.S. in Computer Science and Engineering, Seoul National University, South Korea
2011	Ph.D. in Computer Science and Engineering, Pennsylvania State University, USA

Professional Experience

2001-2004	Instructor in Computer Science, Korea Naval Academy, South Korea
2012-2016	Post-doctoral scholar in Genetics, Stanford University, USA
2016-2020	Assistant Professor in Computer Science and Engineering, Pusan National University, South Korea
2020-	Associate Professor in Computer Science and Engineering, Pusan National University, South Korea
2020-	Director, the Center for Artificial Intelligence Research, Pusan National University, South Korea

Selected Publications (3 maximum)

1. Kibeom Kim, Juseong Kim, Minwook Kim, Hyewon Lee, Giltae Song*. Therapeutic Gene Target Prediction Using Novel Deep Hypergraph Representation Learning, Briefings in Bioinformatics, 2025.
2. Minwook Kim, Donggil Kang, Min Sun Kim, Jeong Cheon Choe, Sun-Hack Lee, Jin Hee Ahn, Jun-Hyok Oh, Jung Hyun Choi, Han Cheol Lee, Kwang Soo Cha, Kyungtae Jang, WooR I Bong, Giltae Song*, Hyewon Lee*, Journal of the American Medical Informatics Association, 2024.
3. Dohyeon Lee, Giltae Song*. FastqCLS: a FASTQ compressor for long-read sequencing via read reordering using a novel scoring model, Bioinformatics, 2022.

KSBI | 한국생명정보학회

Recommendation Systems in Bioinformatics

Giltae Song
Pusan National University

Recommendation Systems

❖ **Designed to recommend items** to the user **based on many different factors**

NETFLIX

Google News

YouTube

amazon

Pinterest

2

What Long-Tail tells?

- A physical world will display only the items that are **most popular**
- A on-line world is possible to **tailor the store to each individual customer**

Formal Model

- ***X*** = set of **Users**
- ***S*** = set of **Items**

- **Utility function** **u**: $X \times S \rightarrow R$
  - ***R*** = set of ratings
  - ***R*** is a totally ordered set
  - e.g., **0-5** stars, real number in **[0, 1]**

Utility Matrix

- The utility matrix represent a value that represents what is known about **the degree of preference of that user for that item**
- In most case, the utility matrix is **sparse**, meaning that **most entries are "unknown"**

	Iron Man	Shang-Chi	Minions	Toy Story
Jisoo	5		4	4
Jennie		1		
Rosé	2		4	
Lisa		3		5

5

Key Problems

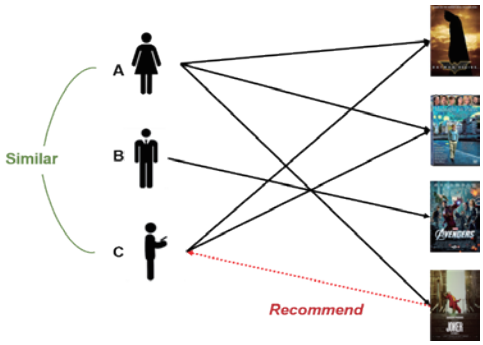
- **Gathering "known" ratings for matrix**
  → Explicit or Implicit
- **Extrapolate unknown ratings from the known ones**
  → Content-based, Collaborative, Latent factor based
- **Evaluating extrapolation methods**
  → How to measure success/performance of recommendation methods

6

Collaborative Filtering

- Focus on the similarity of the user rating for two items
- Identify similar users and recommend what similar users like is

- **Let s_{ij} similarity** of items ***i*** and ***j***, and r_{xj} rating of user ***x*** on item ***j***
- Select ***k***-nearest neighbors, ***N(i; x):*** items most similar to ***i*** that were rated by ***x***



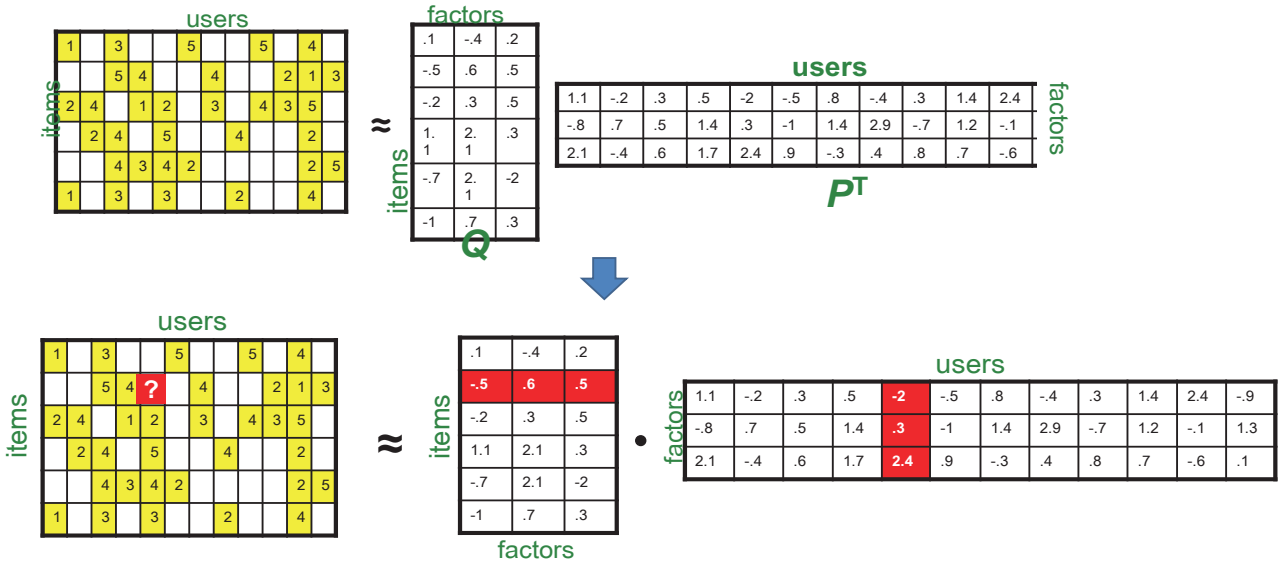
$$\hat{r}_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

Local & global effects

$$r_{xi} = b_{xi} + \frac{\sum_{j \in N(i;x)} s_{ij} \cdot (r_{xj} - b_{xj})}{\sum_{j \in N(i;x)} s_{ij}}$$

baseline $b_{xi} = \mu + b_x + b_i$

Latent Factor Models



Matrix factorization

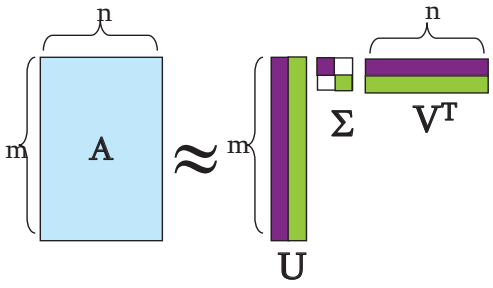
- Singular Value Decomposition (SVD)

  $A = R,\ Q = U,\ P^T = \Sigma V^T$

- SVD is not defined when entries are missing
- Use specialized methods to find P, Q

  $\min_{P,Q} \sum_{(i,x)\in R} (r_{xi} - q_i \cdot p_x)^2$

  (e.g. using stochastic gradient descent)



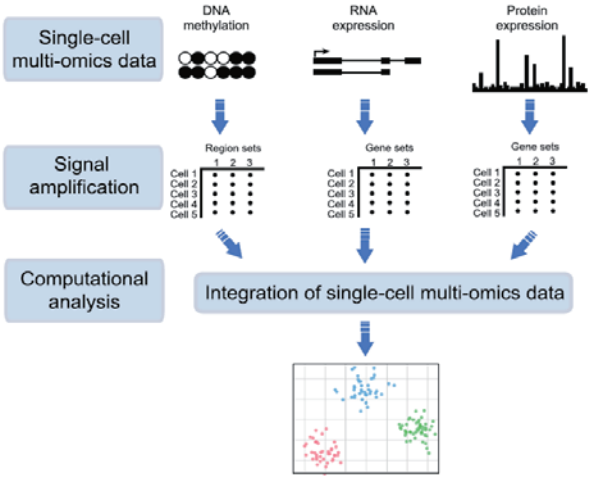
9

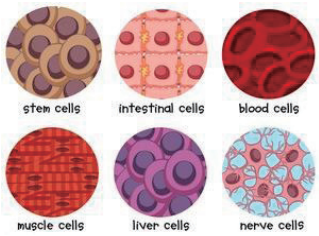
Recommendation systems for biomarker discovery

10

Integrative single cell analysis

- Identify and characterize cell types and their organization in space and over time



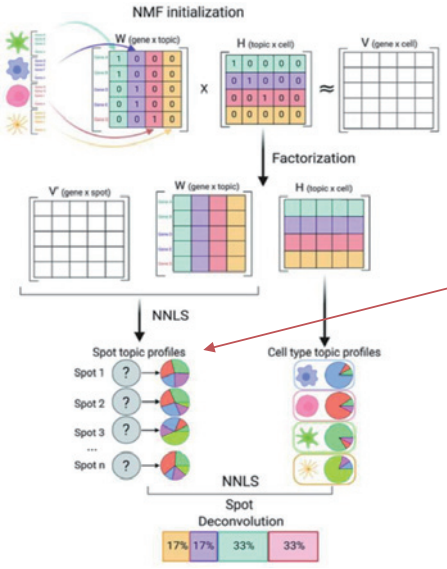


Cell type examples

Identification and characterization of cells using non-negative matrix factorization (NMF)

Single cell RNA-seq

Spatial transcriptomics

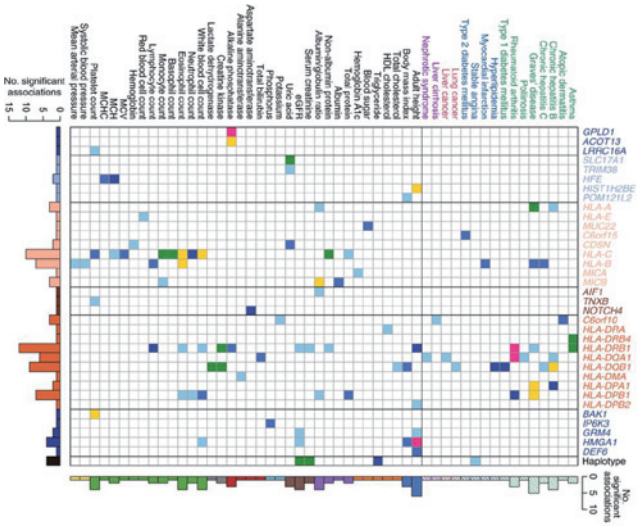


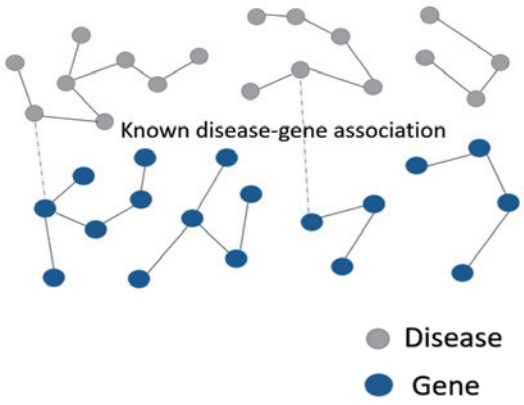
$V \sim W * H$ Using NMF

$V' \sim W * H'$ Using NN Least Squares

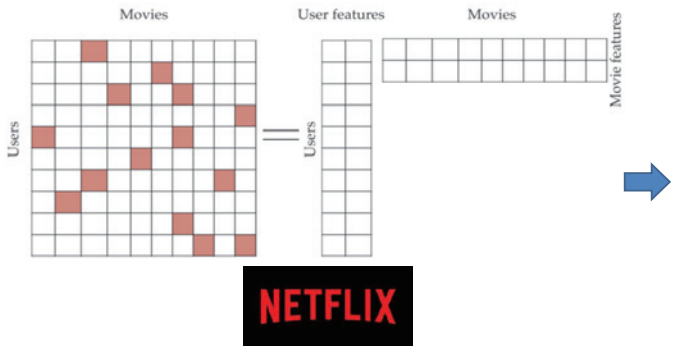
Elosua-Bayes et al., Nucleic Acids Res., 2021

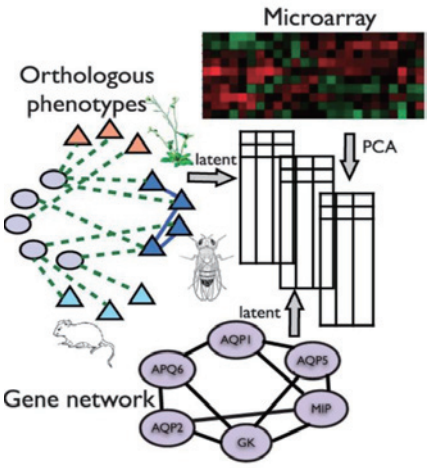
Disease-gene association





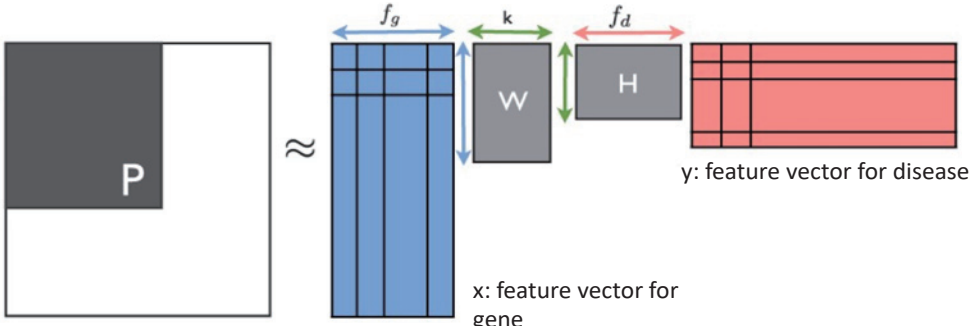
Different from Netflix problem





Gene-disease association prediction using matrix factorization

$$\min_{W \in \mathbb{R}^{f_g \times k}, H \in \mathbb{R}^{f_d \times k}} \sum_{(i,j) \in \Omega} \ell\left(P_{ij}, \mathbf{x}_i^T W H^T \mathbf{y}_j\right) + \frac{\lambda}{2}\left(\|W\|_F^2 + \|H\|_F^2\right)$$



gene−disease associations matrix $P \in \mathbb{R}^{N_g \times N_d}$

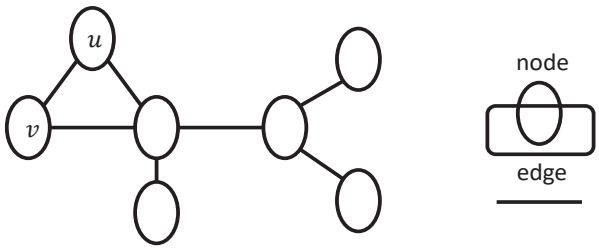
Natarajan *et al.*, Bioinformatics (2014)

Graph representation learning

16

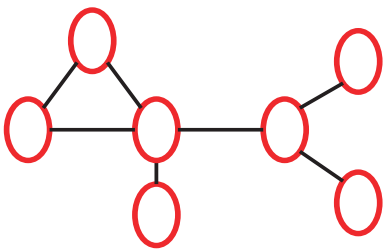
What is Graph data

- **Graph** $(\boldsymbol{G})$ data represents entities as **nodes** $(\boldsymbol{V})$ and the relationships between them as **edges** $(\boldsymbol{E})$
  - $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$,
  - $u \in V$, $v \in V$, and $(\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E}$
- Graphs are a general language for describing entities with relations/interactions



Nodes (V)

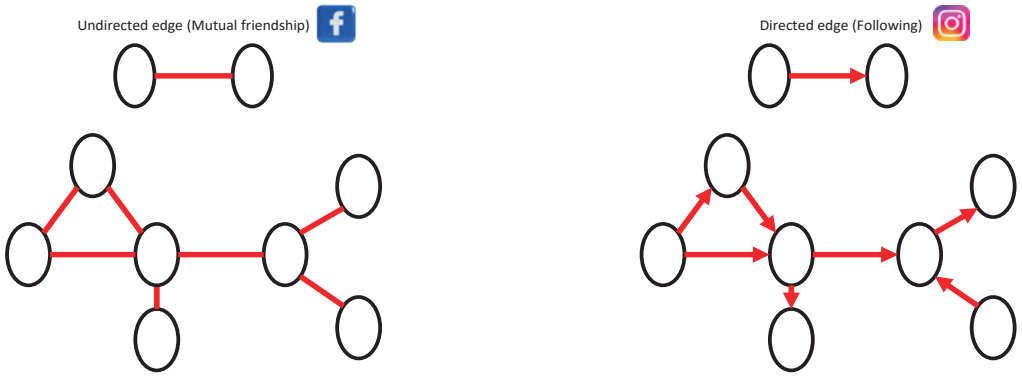
- Nodes (also called vertices) represent entities or objects
  - In a social network, each person would be a node
- Node can have attributes or properties
  - Node representing a person could have attributes like name, age, location, etc.



출처: https://medium.com/analytics-vidhya/social-network-analytics-f082f4e21b16

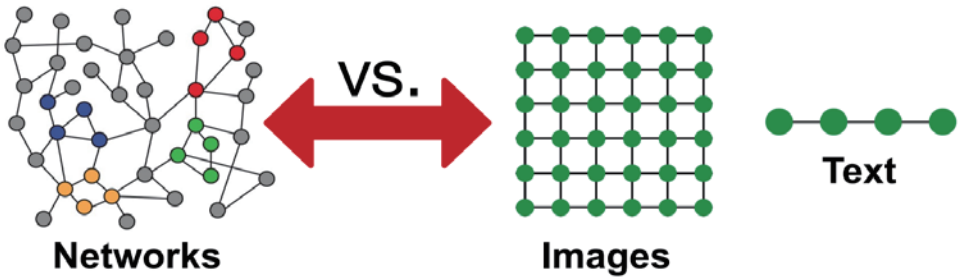
Edges (E)

- Edges (or links) are the connections between nodes.
  - They represent the relationships or interactions between these entities
- Edges can be undirected or directed
  - **Undirected edge** implies a two-way relationship (like a mutual friendship on Facebook)
  - **Directed edge** shows a one-way relationship (like a follower on Instagram)



Graph data vs. other data types

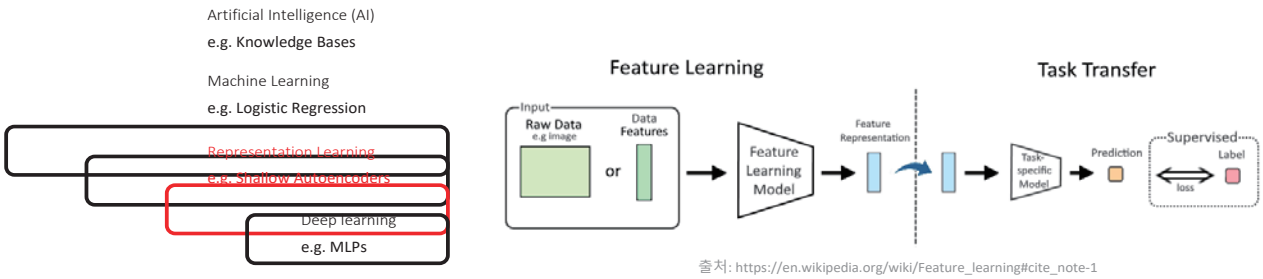
- Networks are complex
  - Graphs are irregular and intricately connected, lacking a uniform structure unlike the regularity of images or the sequence of text
- No fixed node ordering or reference point
  - Graphs do not have a natural order or fixed starting point for processing, as opposed to the structured layout of text and images



출처: https://web.stanford.edu/class/cs224w/slides/01-intro.pdf

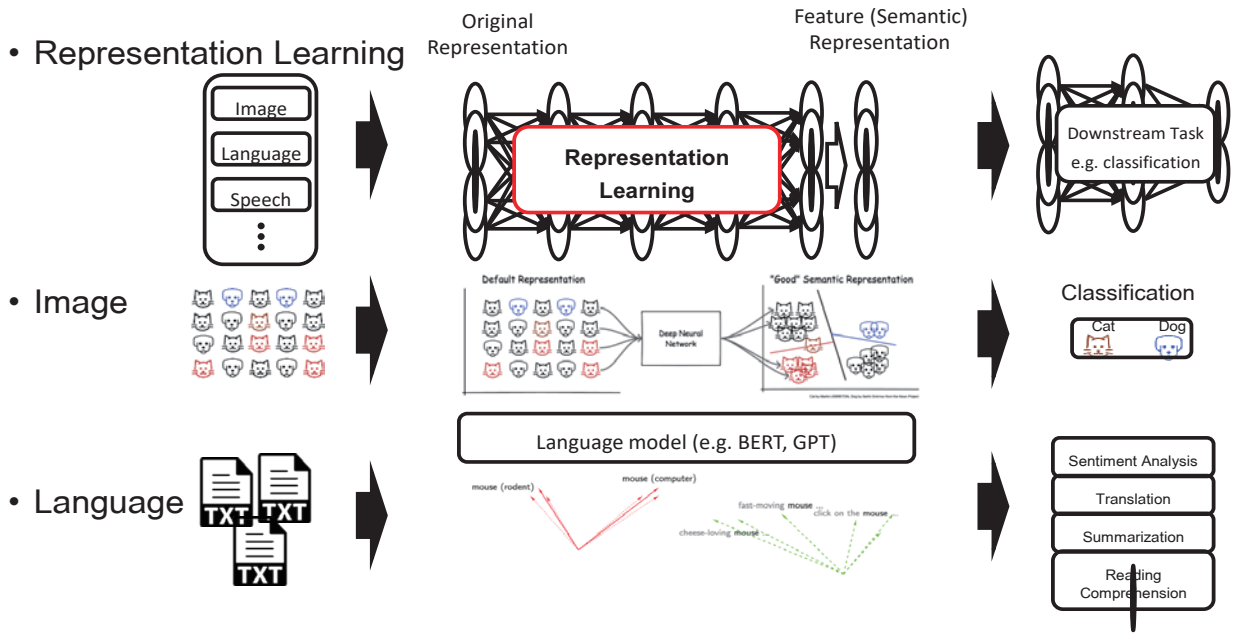
Representation Learning

- In machine learning, **feature learning** or **representation learning** is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data
  - Representation
    - It refers to the abstracted form of raw data, transformed into a format (like feature vectors or embeddings) that a deep learning model can efficiently process and learn from.



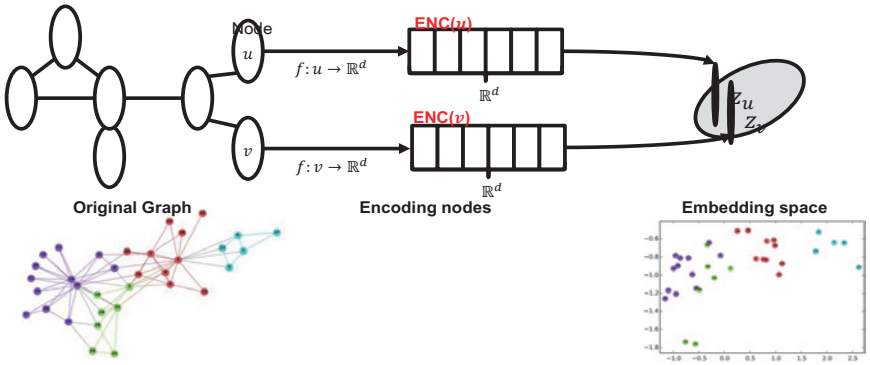
출처: https://en.wikipedia.org/wiki/Feature_learning#cite_note-1

Representation Learning



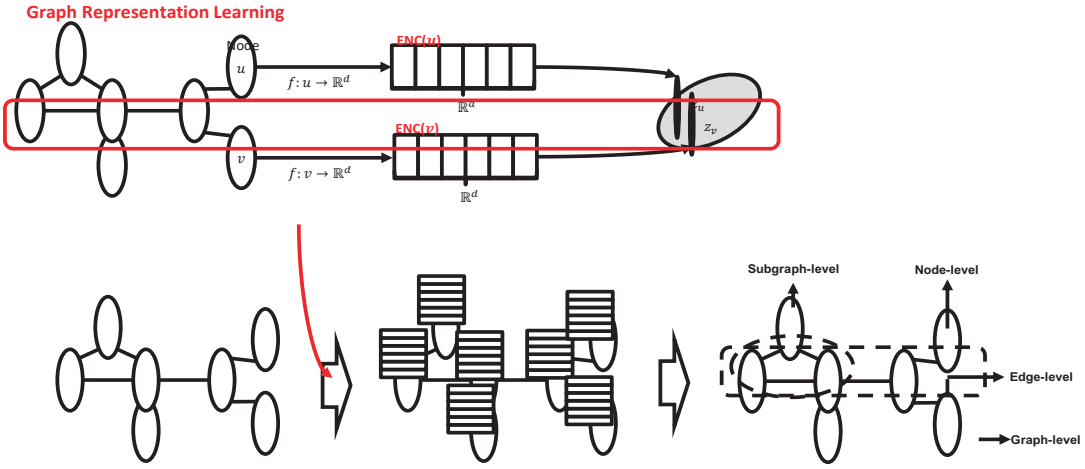
Graph Representation Learning (GRL)

- **Graph representation learning (GRL)** aims to effectively encode high-dimensional sparse graph-structured data into low-dimensional dense vectors (**embeddings**).
  - **Embeddings**: Map nodes into an embedding space
    - Similarity of embeddings between nodes indicates their similarity in the network.



Graph Representation Learning (GRL)

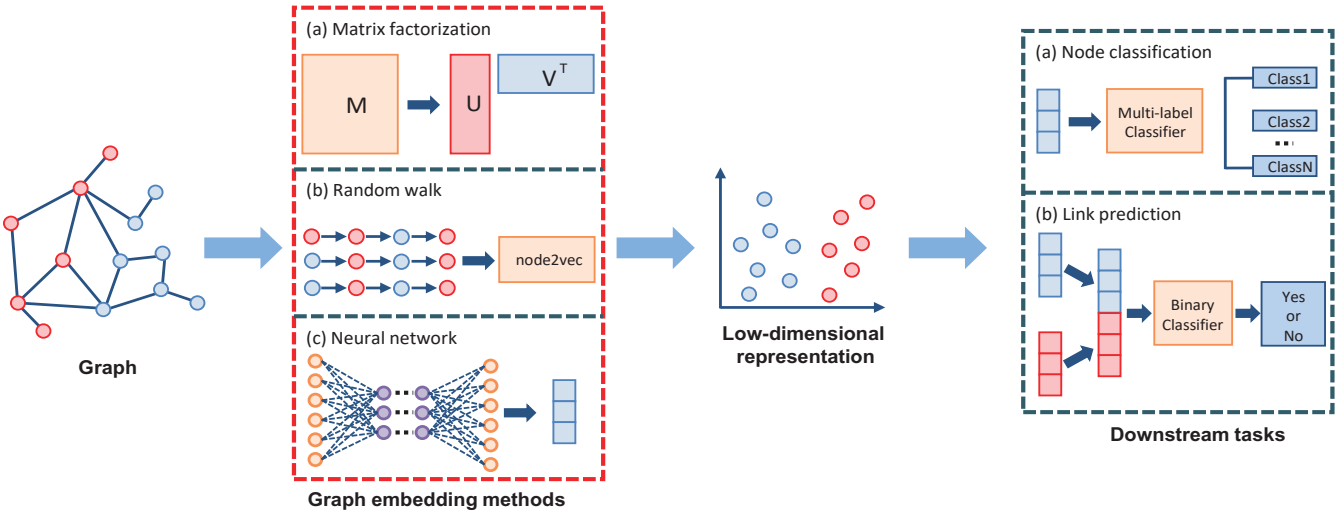
- **Graph representation learning (GRL)** aims to effectively encode high-dimensional sparse graph-structured data into low-dimensional dense vectors (**embeddings**).
  - **unsupervised GRL** : Matrix factorization, Random walk, Neural network
  - **self-supervised GRL** :GNN (GNN, Graph Convolutional Networks, Graph ATtention networks)



Unsupervised GRL

25

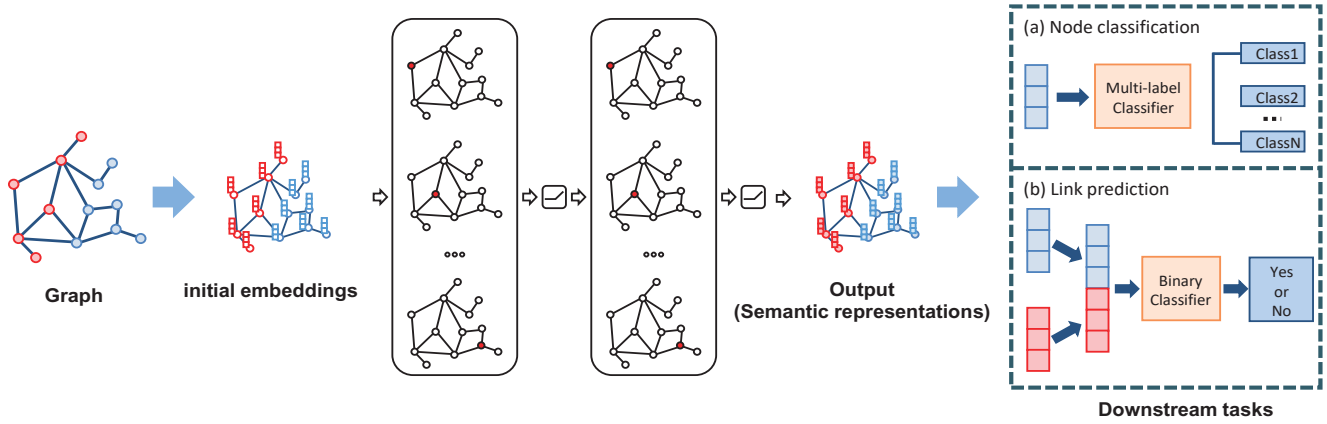
- Network Embedding
- Learning node embeddings without labels



Semi-Supervised GRL

26

- End-to-end learning (with Graph Neural Network)
- Learning node embeddings with partial labels



Graph Neural Network (GNN)

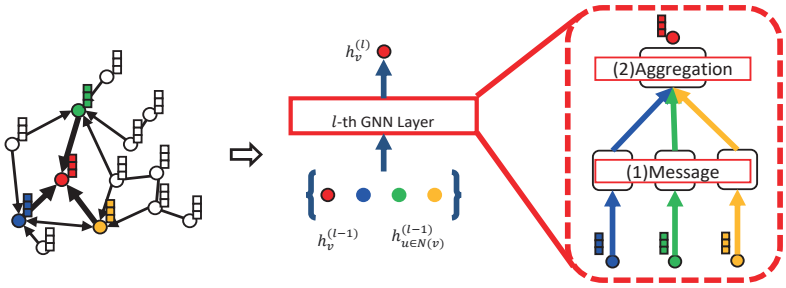
27

- Two-step process
  - **Message**: Each node will create a message, which will be sent to other nodes later

$$m_u^{(l)} = \mathrm{MSG}^{(l)}\left(h_u^{(l-1)}\right)$$

  - **Aggregation**: Node v will aggregate the messages from its neighbors u

$$h_v^{(l)} = \mathrm{AGG}^{(l)}\left(\left\{m_u^{(l)}, u \in N(v)\right\}\right)$$



GNN: Graph Convolutional Network (GCN)

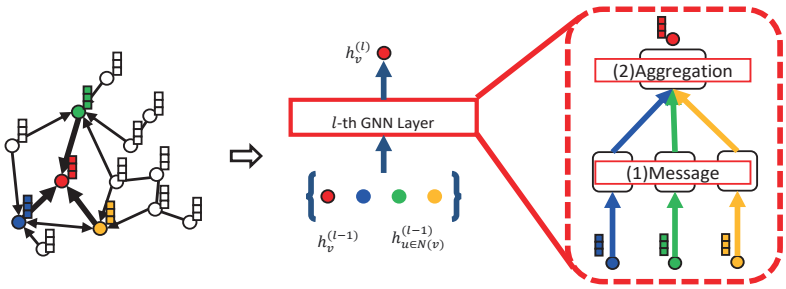
28

- Two-step process
  - **Message**: Node features are updated using a weight matrix $W^{(l)}$

$$m_u^{(l)} = \frac{W^{(l)} h_u^{(l-1)}}{\sqrt{|N(u)||N(v)}}$$

  - **Aggregation**: Messages from neighbors are summed and an activation function is applied.

$$h_v^{(l)} = \sigma\left(\sum_{u\in N(v)\cup v} m_u^{(l)}\right)$$



GNN: Graph Attention Network (GAT)

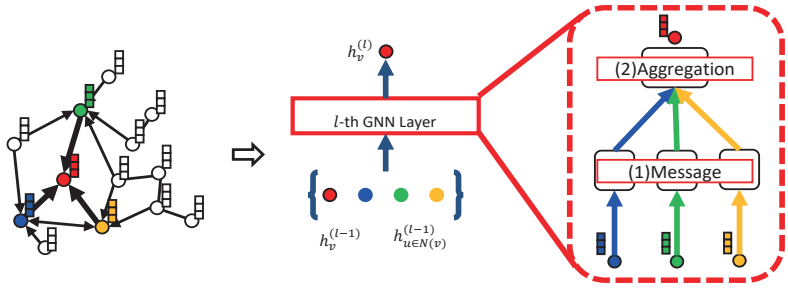
- Two-step process
  - **Message**: Each node's features are transformed and an attention score is computed between pairs of u and v

$$\alpha_{vu} = \text{LeakyReLU}(\boldsymbol{a}^T[Wh_v \| Wh_u])$$

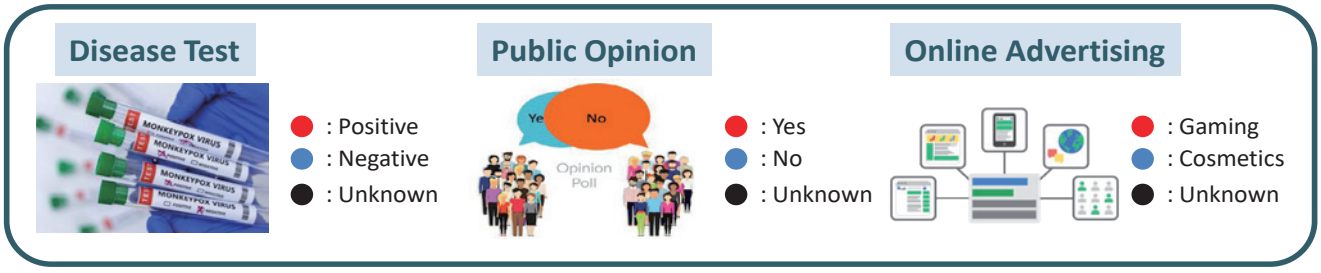
where || is concatenation and $\boldsymbol{a}$ is a weight vector

  - **Aggregation**: Normalize attention scores, then aggregate neighbors' features weighted by the attention scores.

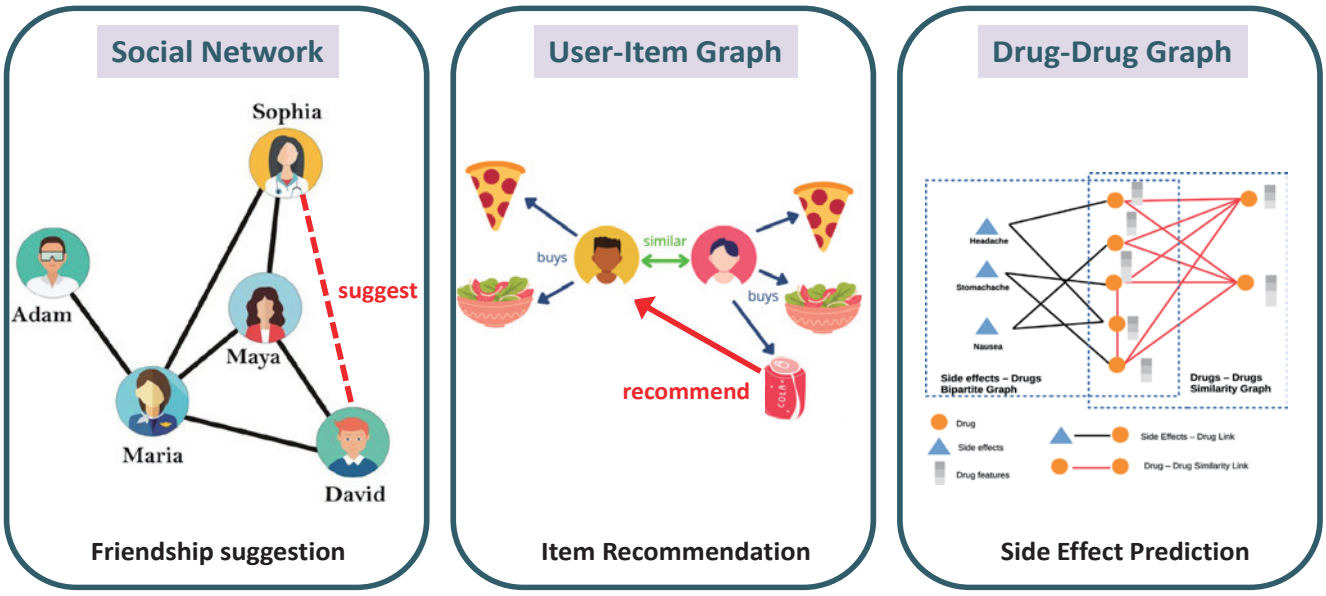
$$h_v^{(l)} = \sigma\left(\sum_{u \in N(v)} \text{softmax}_u(\alpha_{vu}) W^{(l)} h_u^{(l-1)}\right)$$



Node Classification



Link Prediction

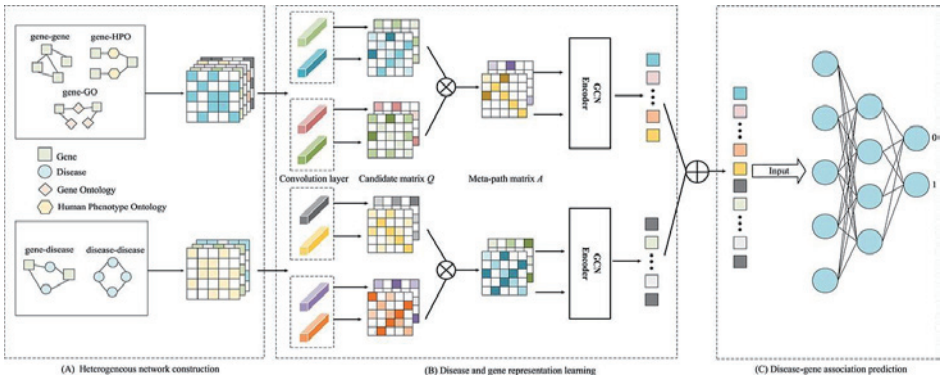


Heterogeneous network representation learning

Heterogeneous network representation learning

33

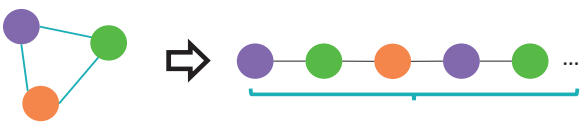
- Researchers suggested heterogeneous graph represetation learning methods to utilize abundant heterogeneous biological information when predicting whether a gene is asso-ciated with a disease.
- **Diease Gene association Prediction – DGP-PGTN (Yang Li et al, 2023)**



Hypergraph

34

- A network structure for capturing high-order relationships among multiple nodes (Yifan Feng *et al.*, 2019), which implies that the network can indeed replace metapaths (multi-hop heterogeneous node sequences).



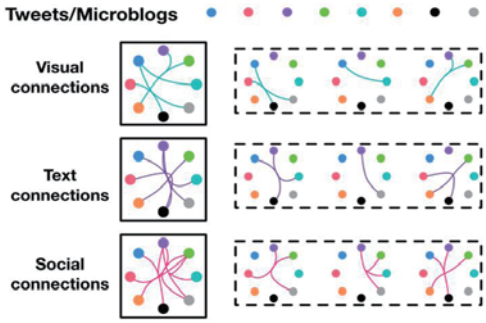
- The above hyperedge represents common relationship among three nodes.
- The connection implies a multi-hop node sequences.

- The hyperedge represents the relationship among above three nodes (endlessly overlapped 2-hop node sequences).

Hypergraph

35

- Suitable network structure for processing multi-modal information.



⇨ Overlap hypergraphs of three modalities

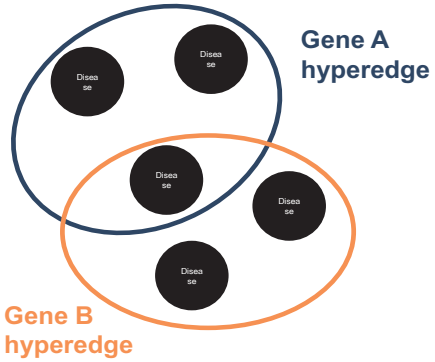
Above pictures are from the paper 'Hypergraph Neural Networks (Yifan Feng *et al*., 2019, AAAI)'

Hypergraph Interaction Transformer (HIT)

36

- "heterogeneous" methods are limited to only predicting for binary classification only
- To apply a multiclass classifications, HIT is proposed

⇨



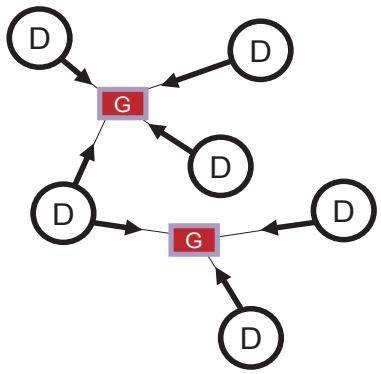
The association type between **a gene** and a **disease**

Kibeom Kim et al., *Briefings in Bioinformatics* 2025

How HIT operates?

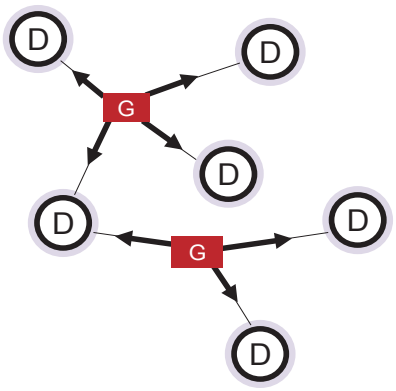
37

Intra **Mutual Attention**



- The Attention mechanism updates hyper-edge (common relationship)'s information using node information.

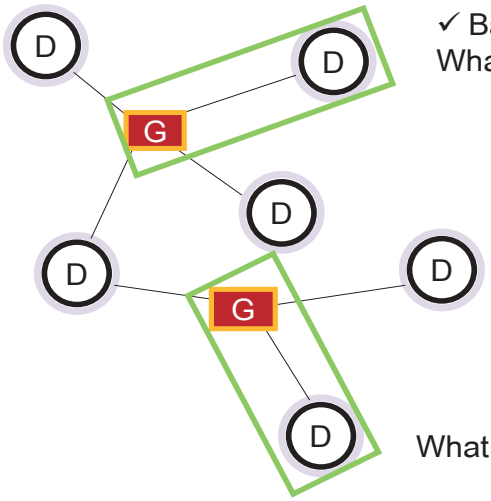
Inter **Mutual Attention**



- The Attention mechanism updates node's information using hyper-edge information.

How HIT operates?

38



✓ Based on the network structure,
What is the type of the association between **G** and **D**?

What is the type of the association between **G** and **D**?

HIT Performance

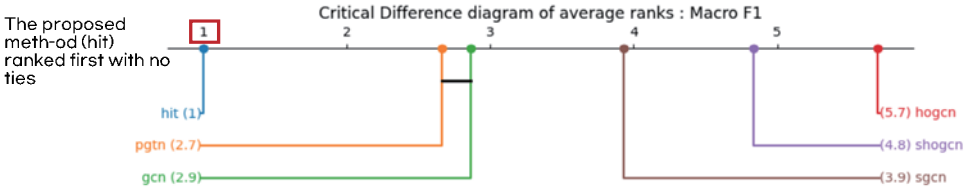
- HIT achieved the best performance against other baseline models

Table 5 The performance comparison between HIT and the baselines. The top results are highlighted in bold.

Method	Macro F1	Precision	Recall	F1(therapeutic)	F1(biomarker)	F1(NA)	Accuracy	Micro F1
GCN-S	$0.7470_{\pm 0.016}$	$0.7482_{\pm 0.018}$	$0.7482_{\pm 0.015}$	$0.6808_{\pm 0.018}$	$0.6915_{\pm 0.029}$	$0.8687_{\pm 0.006}$	$0.7982_{\pm 0.013}$	$0.7982_{\pm 0.013}$
HOGCN-S	$0.5497_{\pm 0.187}$	$0.6047_{\pm 0.168}$	$0.5569_{\pm 0.173}$	$0.4798_{\pm 0.224}$	$0.3768_{\pm 0.303}$	$0.7926_{\pm 0.066}$	$0.6873_{\pm 0.101}$	$0.6873_{\pm 0.101}$
GCN	$0.7613_{\pm 0.011}$	$0.7651_{\pm 0.012}$	$0.7589_{\pm 0.012}$	$0.6862_{\pm 0.012}$	$0.7194_{\pm 0.019}$	$0.8783_{\pm 0.005}$	$0.8125_{\pm 0.008}$	$0.8125_{\pm 0.008}$
HOGCN	$0.3377_{\pm 0.077}$	$0.4508_{\pm 0.179}$	$0.3846_{\pm 0.047}$	$0.2079_{\pm 0.158}$	$0.0487_{\pm 0.091}$	$0.7565_{\pm 0.007}$	$0.6151_{\pm 0.016}$	$0.6151_{\pm 0.016}$
DGP-PGTN	$0.772_{\pm 0.007}$	$0.7757_{\pm 0.011}$	$0.7706_{\pm 0.012}$	$0.7021_{\pm 0.024}$	$0.7384_{\pm 0.014}$	$0.8754_{\pm 0.008}$	$0.8165_{\pm 0.008}$	$0.8165_{\pm 0.008}$
HIT	$\mathbf{0.8431}_{\pm 0.01}$	$\mathbf{0.8373}_{\pm 0.014}$	$\mathbf{0.85}_{\pm 0.008}$	$\mathbf{0.7385}_{\pm 0.013}$	$\mathbf{0.8418}_{\pm 0.015}$	$\mathbf{0.9491}_{\pm 0.002}$	$\mathbf{0.8926}_{\pm 0.007}$	$\mathbf{0.8926}_{\pm 0.007}$

Proposed model

- The performance gap between HIT and other baselines is statistically meaningful

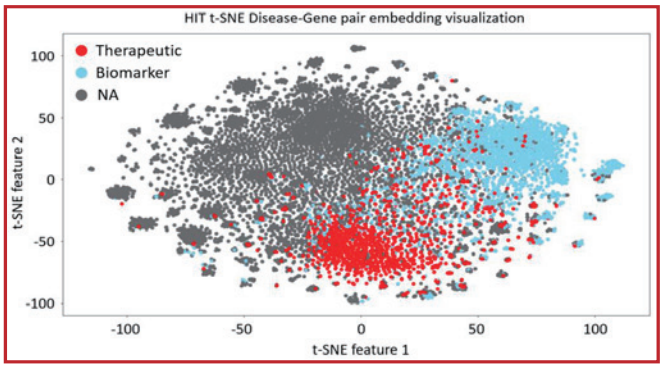


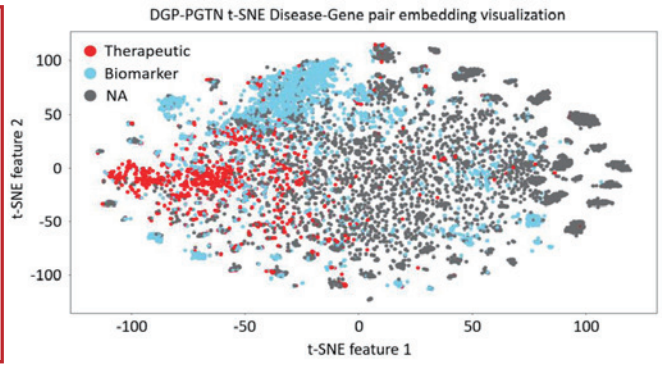
- Black-horizontal tie means no statistical difference between model performances

Kibeom Kim et al., *Briefings in Bioinformatics* 2025

T-SNE visualization

- How HIT distinguishes each data instances into three groups
- Gene-Disease pair (relationship) instances





- The suggested model (left) classifies data instances into clear three categories.

Kibeom Kim et al., *Briefings in Bioinformatics* 2025

Ablation study

- In ablation experiment, the proposed model always shows better performance than other baselines, even with less association information.

Table 7 Results of ablation experiments. Each value is the average Macro F1 score on five random seeds. Top results are highlighted in bold.

	HIT	DGP-PGTN	GCN
w/o GO	**0.8435**$_{\pm 0.01}$	0.7737$_{\pm 0.009}$	0.7525$_{\pm 0.012}$
w/o DO	**0.846**$_{\pm 0.013}$	0.7641$_{\pm 0.007}$	0.7515$_{\pm 0.017}$
w/o HPO	**0.8447**$_{\pm 0.012}$	0.7586$_{\pm 0.008}$	0.7384$_{\pm 0.012}$
w/o HPO & GO	**0.8482**$_{\pm 0.007}$	0.7576$_{\pm 0.01}$	0.7397$_{\pm 0.015}$
w/o GO & DO	**0.8436**$_{\pm 0.01}$	0.769$_{\pm 0.012}$	0.7558$_{\pm 0.01}$
w/o DO & HPO	**0.8497**$_{\pm 0.012}$	0.7626$_{\pm 0.007}$	0.7412$_{\pm 0.017}$
w/o GO & DO & HPO	**0.8477**$_{\pm 0.016}$	0.7616$_{\pm 0.013}$	0.7403$_{\pm 0.011}$
Original	**0.8431**$_{\pm 0.01}$	0.772$_{\pm 0.007}$	0.7613$_{\pm 0.011}$

- In all ablation settings, HIT always achieved the best performance against other baselines.

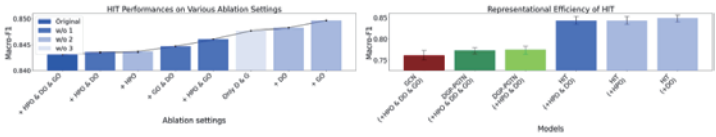


Fig. 13 HIT performance on different ablation settings (**Left**) and representational efficiency of HIT (**Right**).

- Above Fig. 13 derives from the Table 7.
- The left figure above describes the proposed model's performance change tendency with various dataset s-ettings.
- In the right figure, HIT always achieved the best performance against other baselines.

Kibeom Kim et al., *Briefings in Bioinformatics* 2025

Explainability of HIT

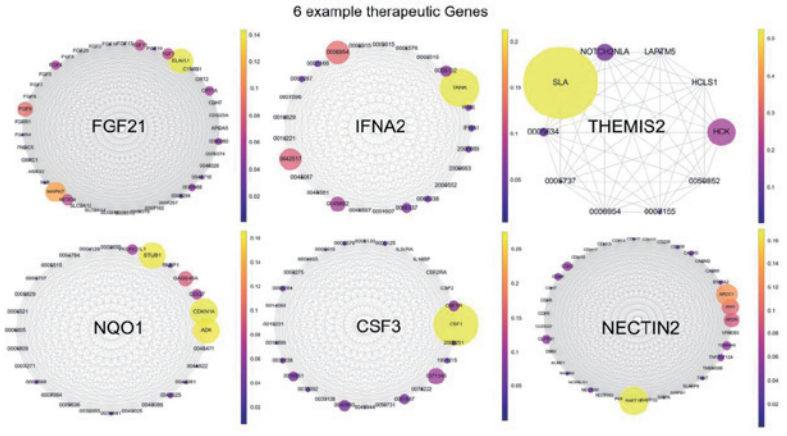


Fig. 14 Complete graph-based attention score visualization. The gene names and GO ids are used to annotate the corresponding biological entities. Referring to the DisGeNET dataset, *FGF21* is identified as a therapeutic gene target for *Steatohepatitis*, *IFNA2* for *Thyroid carcinoma*, *THEMIS2* for *Malignant neoplasm of the breast*, *NQO1* for *Diabetic Nephropathy*, *CSF3* for *Familial Non-Hodgkin Lymphoma*, and *NECTIN2* for *Ovarian neoplasm*.

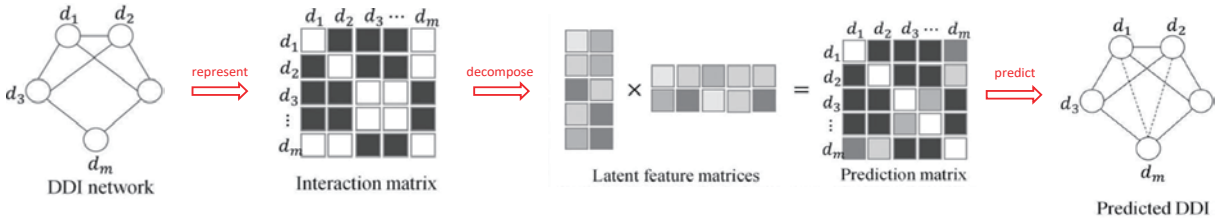
- HIT provides explanations on its decision making based on complete graph analogy of hypergraphs.
- Fig. 14 shows which information does HIT consider as important when classifying six genes (FGF21, IFNA2, THEMIS2, NQO1, CSF3, and NECTIN2) as therapeutic targets of diseases.

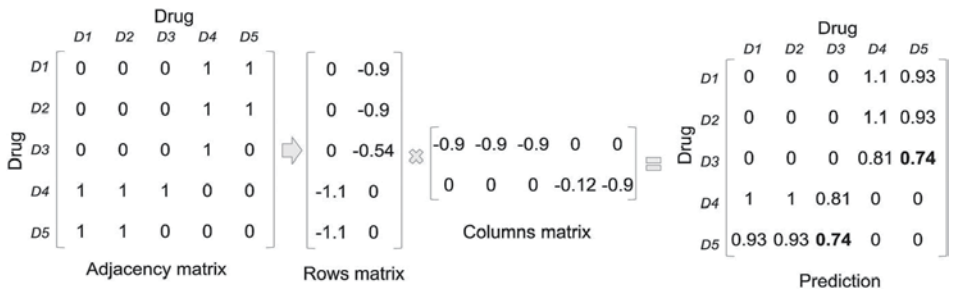
Kibeom Kim et al., *Briefings in Bioinformatics* 2025

Recommendation systems for drug candidate discovery

43

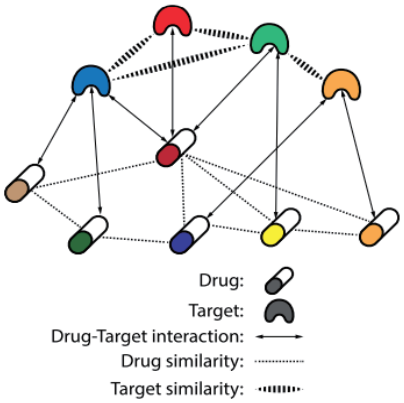
Matrix factorization for Drug-Drug Interaction(DDI)

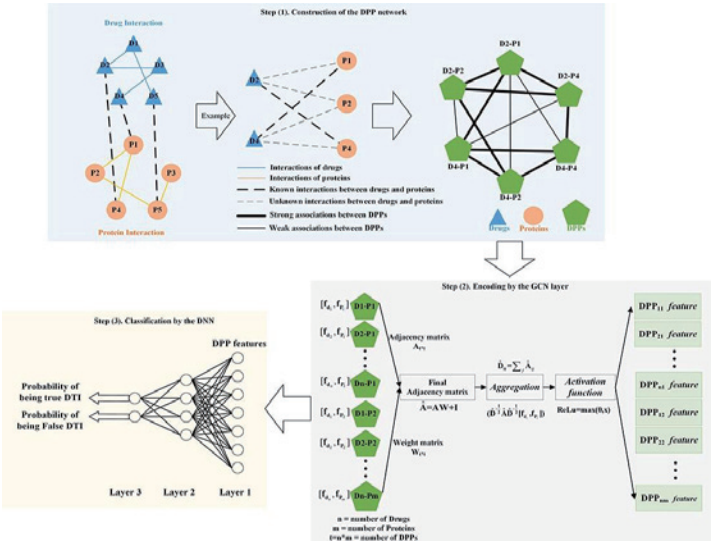




Zhang *et al.*, *J. Biomedical informatics* (2018)

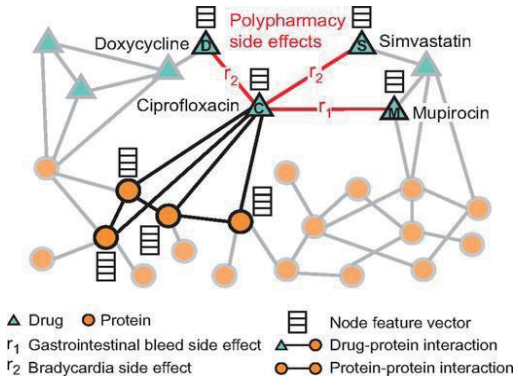
Drug-Target interaction using Graph Convolutional Network (GCN)

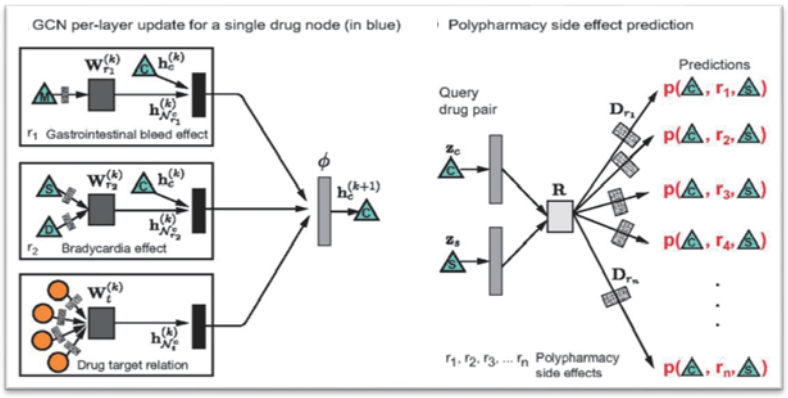




Zhao et al. Briefings in Bioinformatics (2020)

Modeling polypharmacy side effects using Graph Convolutional Network (GCN)





Zitnik et al. Bioinformatics (2018)

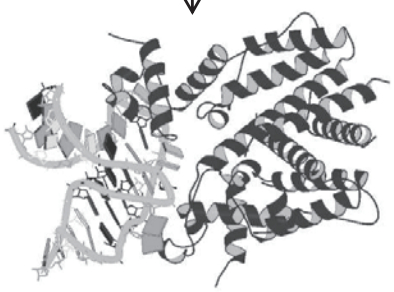
Drug-Target interaction

in silico selection of RNA aptamers about target protein

based on discriminative classifier and Monte-Carlo tree search (MCTS)

	RNA-aptamer	Target Protein
In Medicine,	신약 후보 물질	바이오마커
In Molecular Biology,	짧은 RNA 염기 접힘 구조	아미노산 접힘 구조
In Computer Science,	N A-C-C-G-G- … -G-U-U-A-C-G	M R-H-K-D-E- … -S-T-Q-C-M-A

Aptamer-Protein Complex
압타머-단백질 복합체



RNA aptamer

N

A-C-C-G-G- … -G-U-U-A-C-G

Target protein

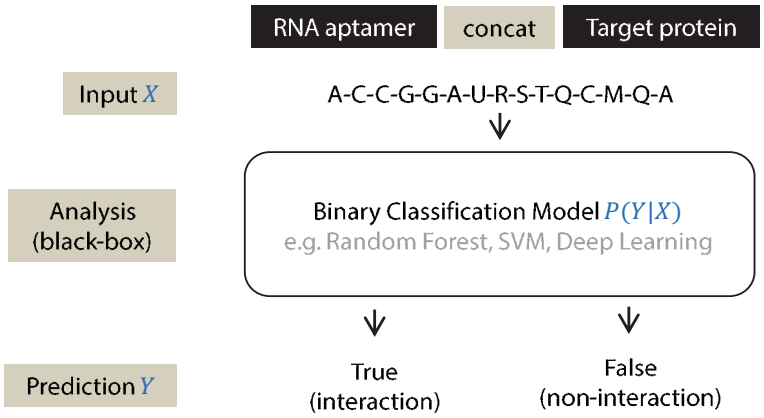
M

R-H-K-D-E- … -S-T-Q-C-M-A

① 결합하는 두 서열(sequence) 사이의 패턴은 무엇인가?
② 어떤 압타머 서열이 표적 단백질에 결합을 잘 하는가?

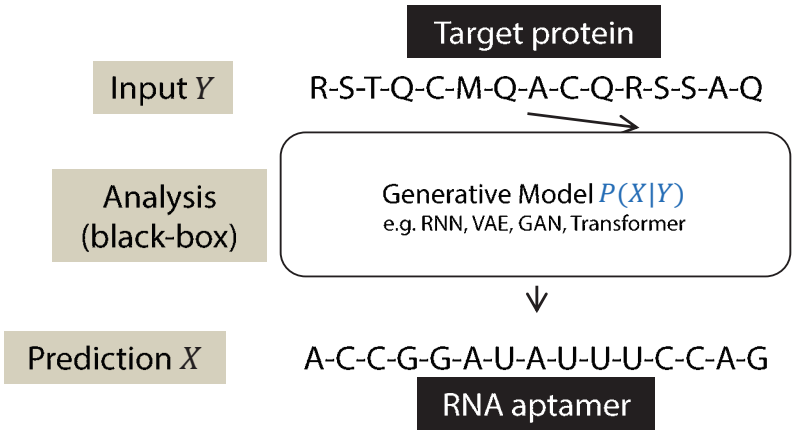
Discriminative Modeling

① 결합하는 두 서열(sequence) 사이의 패턴은 무엇인가?



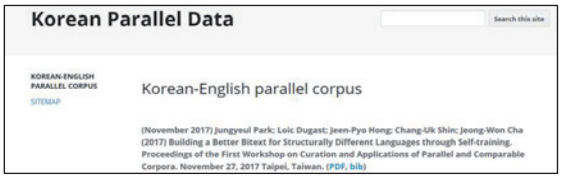
Generative Modeling

② 어떤 압타머 서열이 표적 단백질에 잘 결합하는가?



Limitation – Small Dataset

② 어떤 압타머 서열이 표적 단백질에 잘 결합하는가?
→ 생성 모델(Generative Model)



https://sites.google.com/site/koreanparalleldata/

번역 된 문장(sentence) 생성을 위한
데이터 ~ 약 97,000개

Table 1. The two benchmark API datasets that are used fo

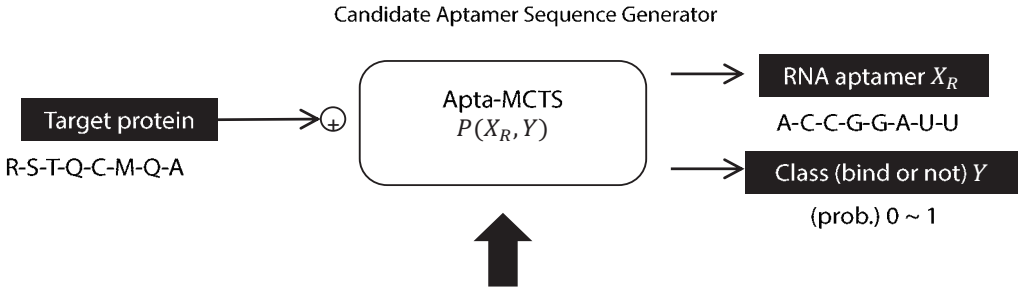
Source	Number of positive pairs	Number of negative pairs
[18]	580	1740
	145	435
[20]	157	493
	56	56

Note that we obtained two pre-trained API classifiers for the aptame

압타머(aptamer) 생성을 위한
데이터 ~ 약 1,000개

*Benchmark dataset for discriminative models

Apta-MCTS



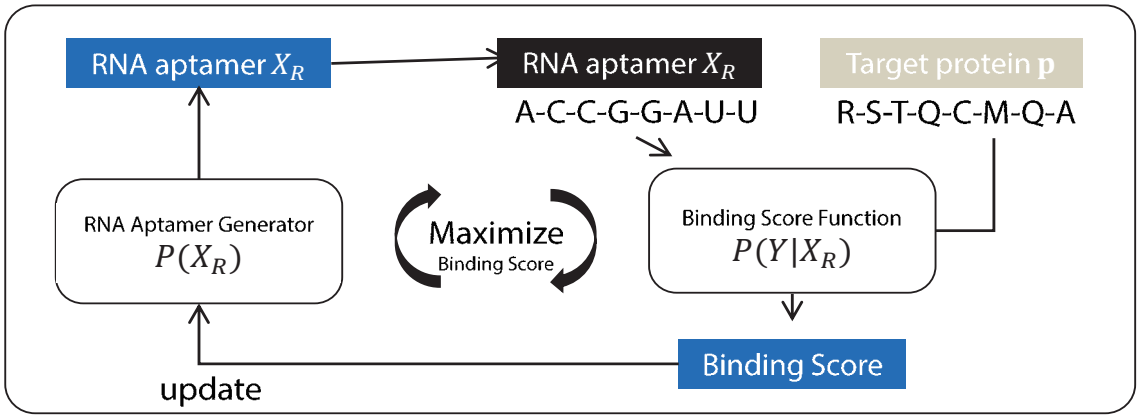
Benchmark Dataset for Discriminative Modeling

# of Positive Aptamer-Protein sequence pair	# of Negative Aptamer-Protein sequence pair	# of Proteins	# of Aptamers	Length of Protein Sequences (avg/std)	Length of Aptamer Sequences (avg/std)
약 1,000 쌍	약 2,800 쌍	166 개	709 개	393.2 / 325	51.2 / 25.5

Lee et al. PLOS ONE (2021)

Apta-MCTS

$$P(X_R, Y) = P(X_R)P(Y|X_R)$$



Lee et al. PLOS ONE (2021)

Binding (Interaction) Score Function

- RNA aptamer sequence

$X_R = (x_1, x_2, x_3, \dots, x_N)$

$x_i \in \{r_1, r_2, r_3, r_4\}$

Length of aptamer N

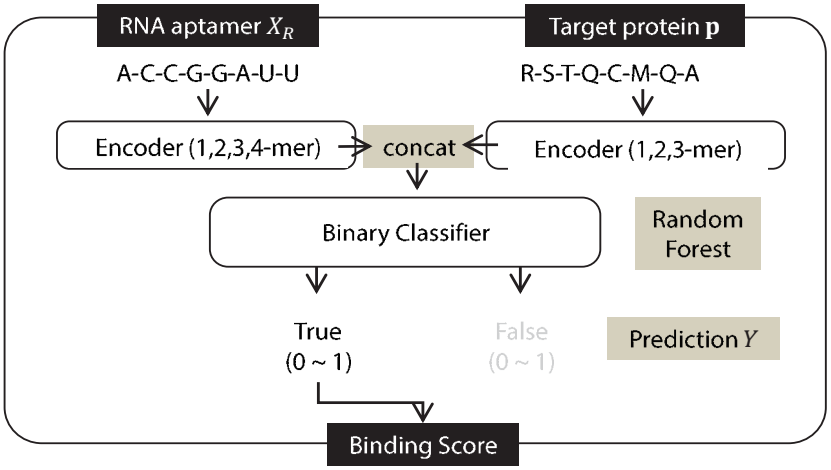
- Target protein sequence

$X_P = (x_1, x_2, x_3, \dots, x_M)$

$x_i \in \{p_1, p_2, p_3, \dots, p_{19}, p_{20}\}$

Length of protein M

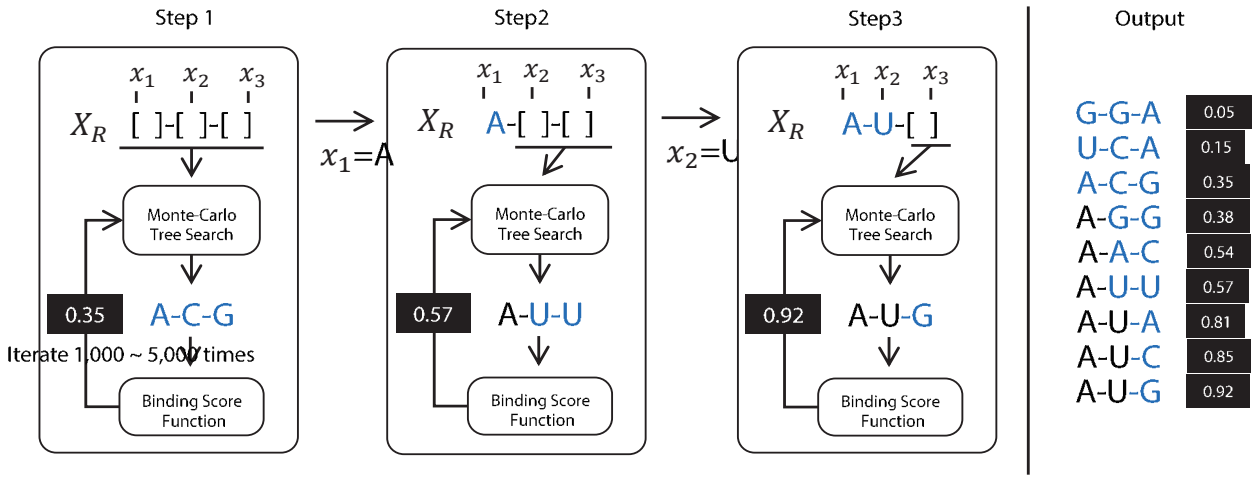
Binding Score Function $P(Y|X_R)$



Lee et al. PLOS ONE (2021)

RNA Sequence Generator – Iterative forward sequencing using MCTS

Ex. Length of aptamer $N = 3$ $\quad X_R = (x_1, x_2, x_3) \quad x_i \in \{r_1, r_2, r_3, r_4\}$

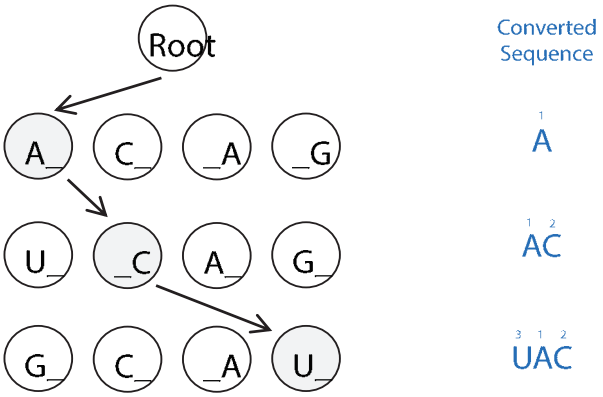


Lee et al. PLOS ONE (2021)

Monte-Carlo Tree Search with Bidirectional Nucleotides

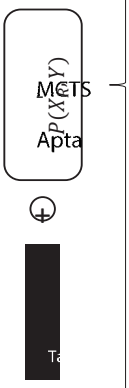
Ex. Length of aptamer $N = 3$ $\quad X_R = (x_1, x_2, x_3) \quad x_i \in \{r_1, r_2, r_3, r_4\}$

Bidirectional RNAs : (_A, _C, _G, _U, A_, C_, G_, U_)

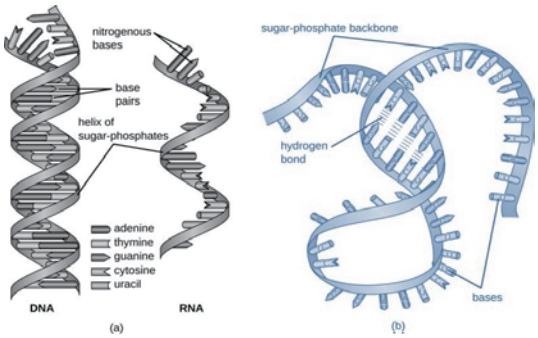


Lee et al. PLOS ONE (2021)

Post-processing

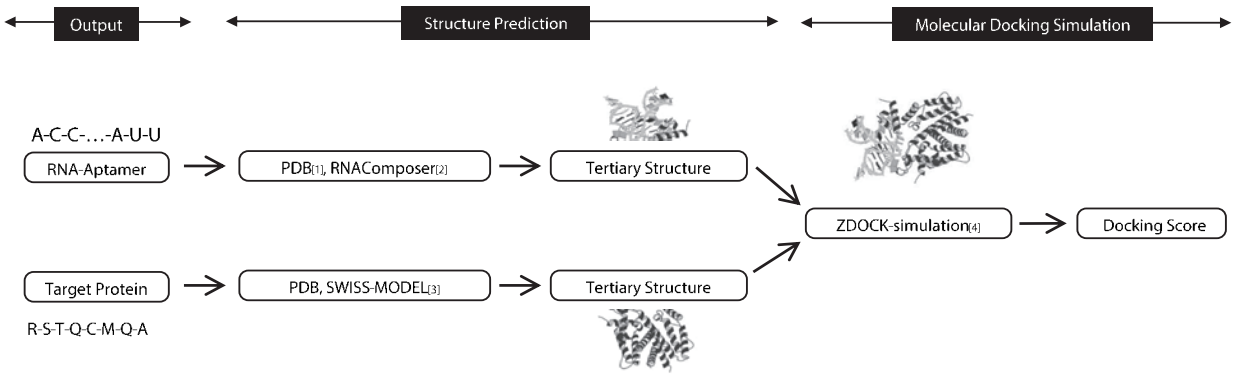


RNA aptamer X_R	Binding Score
A-C-C-…-A-U-U	0.94
A-C-C-…-A-G-U	0.93
A-G-C-…-A-U-C	0.93
C-C-U-…-A-C-U	0.91
G-C-C-…-A-U-U	0.90
G-C-C-…-A-U-U	0.89
A-G-C-…-A-G-U	0.88
A-C-G-…-A-U-U	0.88
G-C-C-…-A-U-U	0.84
A-C-C-…-G-U-U	0.84
A-C-C-…-G-U-U	0.83
U-C-C-…-A-U-U	0.81
A-U-C-…-A-U-U	0.80
A-C-C-…-U-U-U	0.74

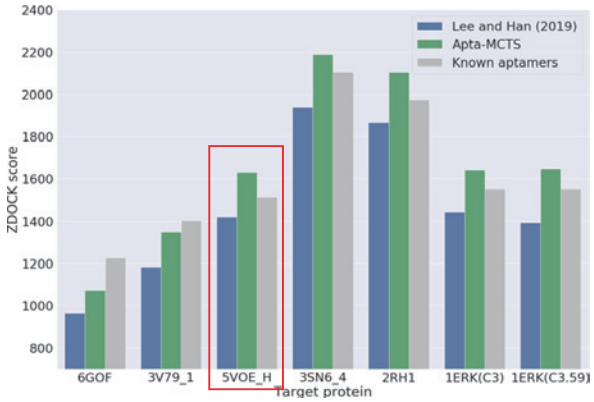


같은 2차 구조를 갖는 압타머 들 중
가장 높은 Score를 갖는 것들만 최종 선별

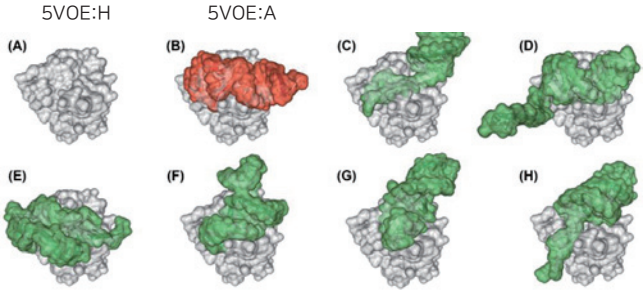
Validation Procedure



Validation – PDB structures



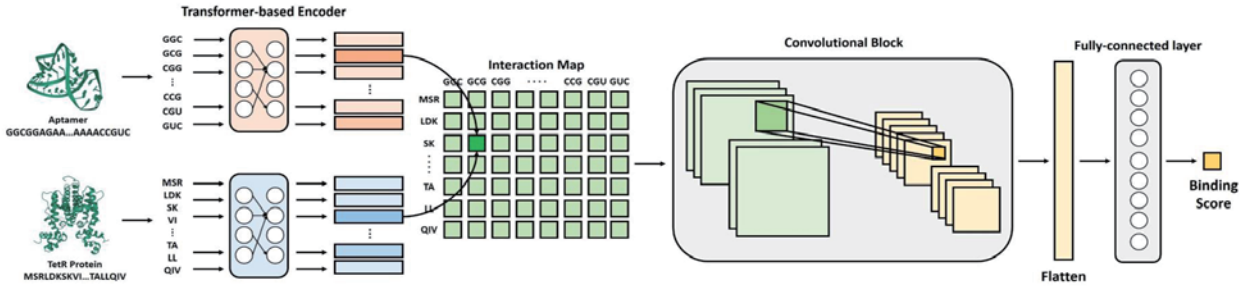
일부 주요 표적 단백질들에 대하여
후보 압타머 생성 및 도킹 시뮬레이션 수행



시뮬레이션 결과
구조적으로 유사한 위치에 결합하는 것을 확인 가능

Lee et al. PLOS ONE (2021)

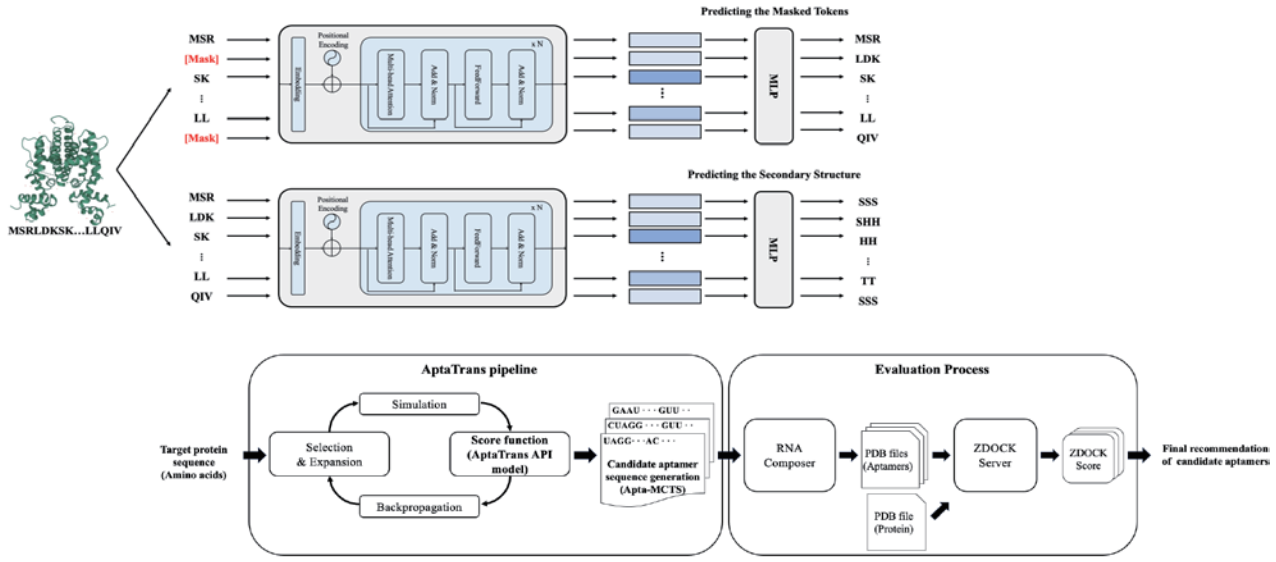
Apta-Trans : Aptamer-Protein interaction prediction using Transformer



- Transfer learning-based Aptamer-Protein Interaction prediction Model
  - A model that predicts the interaction between aptamers and proteins
  - Sequence embedding using a Transformer-based encoder, interaction map, and convolutional neural network (CNN)
  - Enhanced data representation through transfer learning with a Transformer-based encoder

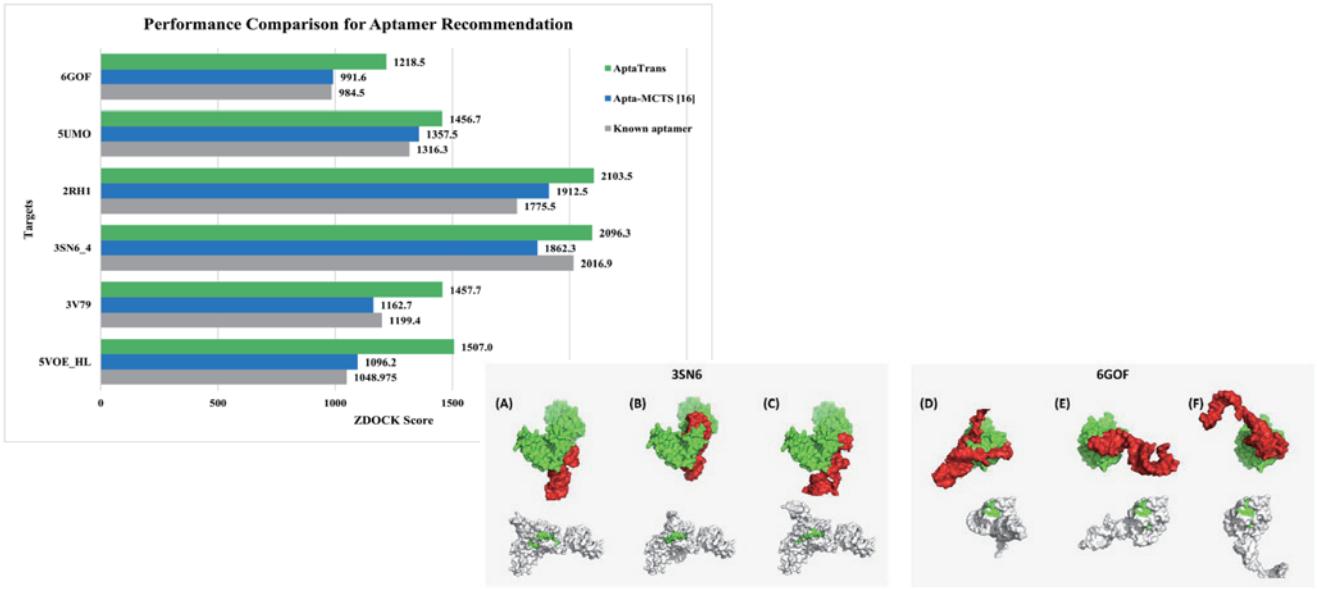
Shin *et al. BMC Bioinformatics (2023)*

Apta-Trans: Aptamer-Protein interaction prediction using Transformer



Shin *et al. BMC Bioinformatics (2023)*

Performance evaluation of Apta-Trans



Shin *et al. BMC Bioinformatics (2023)*

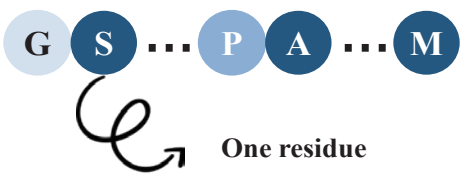
Recommendation systems for drug candidate discovery

63

Protein-ligand binding residue prediction

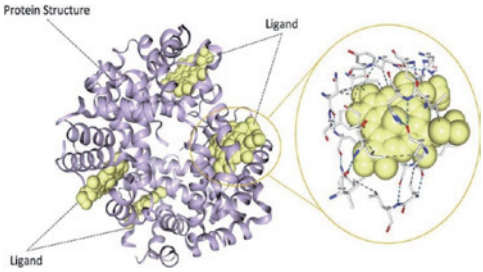
- Process of identifying specific amino acid residues
  within a protein that interact with a ligand molecule
  - ✓Residue : a single amino acid within a protein sequence
  - ✓Ligand : a ligand is a substance that forms a complex with a biomolecule to serve a biological purpose such as a durg

- An example of a residue



Importance of binding residue prediction

- Biological Significance
  - ✓Identifying binding residues determines how proteins function and how they interact with ligands
- Drug Discovery Relevance
  - ✓Key residues narrow down the search space for potential drug candidates during **virtual screening**
    - Virtual screening : a computational technique used to identify potential drug candidates by evaluating their interaction with target proteins



Zhao, Jingtian, Yang Cao, and Le Zhang. "Exploring the computational methods for protein-ligand binding site prediction." *Computational and structural biotechnology journal* 18 (2020): 417-426.

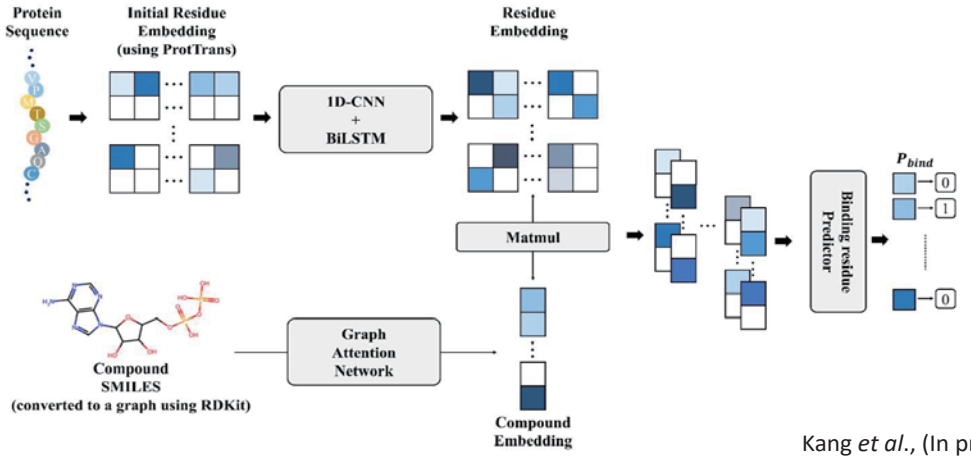
Protein-ligand binding residue dataset

- sc-PDB
  - ✓ An Annotated Database of Druggable Binding Sites from the Protein DataBank
- PDBbind
  - ✓ to provide a comprehensive collection of experimentally measured binding affinity data for all biomolecular complexes deposited in the Protein Data Bank (PDB)

PDB ID	Sequence	SMILES (Ligand Representation)	Binding index
5vsl	KSVVKS ... VQLLII	(=O)OP([O])([O])=O) ... (O	5, 7, 8, ... , 35, 36, 37, 38
1u9l	AHAIFT ... KYDED	CC(C)C(NC(=O) ... C1CCC	177, 179, ... , 212, 213, 214
3mg0	TTIVGVKF ... AASTY	CC(C) C=O) ... C1=C=B(O)O	52, 53, 54, ... , 74, 75, 89

Proposal Method

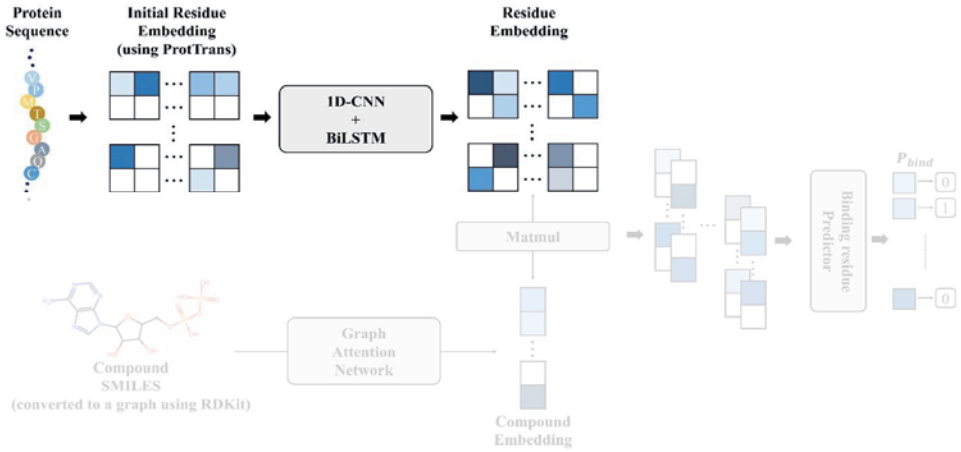
- Predicts the binding score between each residue in the protein and the ligand
- Consists of the following stages
  - ✓ the residue embedding stage, the ligand embedding stage, and the binding prediction stage



Kang *et al.*, (In preparation) 2025

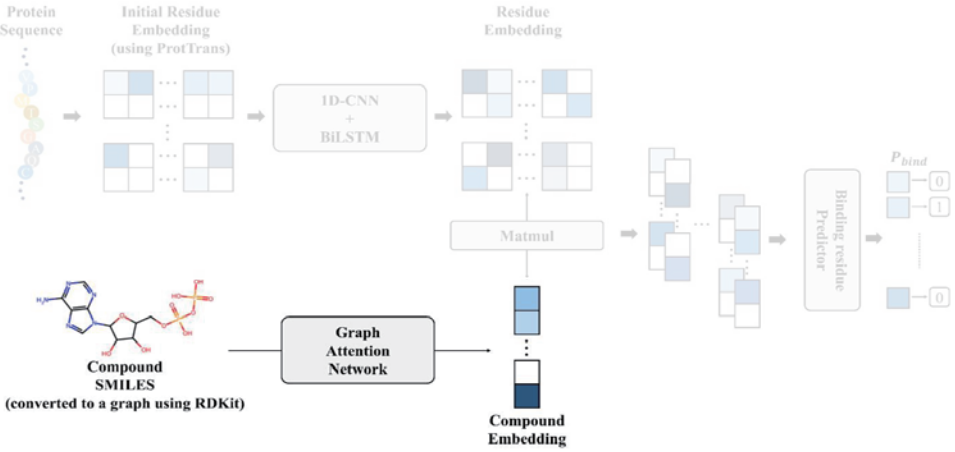
Residue embedding

- We use protein language model to learn residue embeddings, capturing each residue`s representation
- Apply 1D-CNN and BiLSTM to infuse each residue embedding with information about nearby and distant residues, enhancing the embeddings further



Ligand embedding

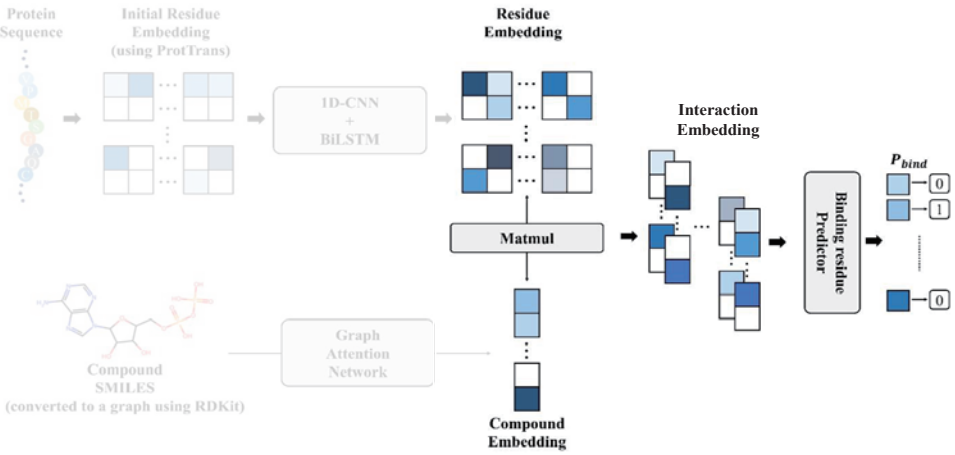
- We first use RDKit to convert the SMILES format of Ligand to a graph representation
- Then, we apply the GAT model from the BACPI[1] model to make Ligand embeddings



[1] Li, Min, *et al*. *Bioinformatics* 38.7 (2022)

Binding Prediction

- The residue embeddings and ligand embeddings are combined to construct interaction embeddings through element-wise multiplication
- The interaction embeddings are processed through a dense layer to obtain the binding score, which is used to predict whether binding or not



Datasets and Preprocessing

- Datasets
  - ✓ We use two databased, sc-PDB (v.2017) and PDBbind (v.2020), to collect protein-ligand binding residue data for model training
  - ✓ For model perfomance evaluation, we use the COACH420 and HOLO4K datasets

- Data preprocessing
  - ✓ Protein sequences exceeding 1000 residues were excluded from the dataset
  - ✓ Data with fewer than 10 binding residues were removed.
  - ✓ Highly similar sequences between the training and test sets were filtered out using the CD-HIT tool

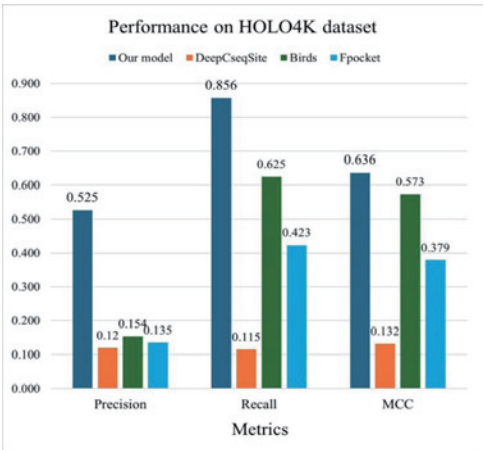
- Statistics training and test dataset

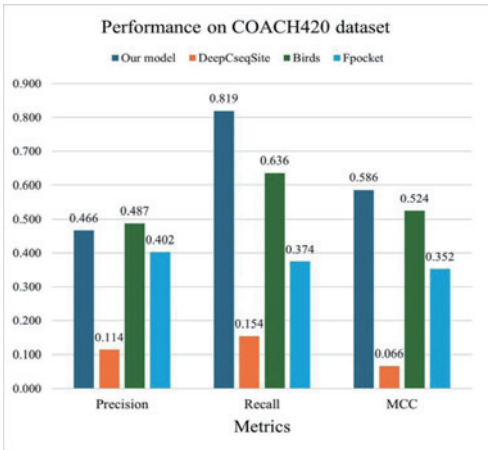
Dataset	$N_{complex}$	N_{BR}	N_{NBR}	$P_{BR}(\%)$
Training	27,486	1,123,317	8,632,471	13.01
COACH420	189	3,061	49,861	6.13
HOLO4K	1,238	31,053	387,435	8.01

$N_{complex}$: number of protein-ligand complex
N_{BR} : number of binding residues
N_{NBR} : number of non-binding residues
P_{BR}: proportion of binding residues

Performance comparison

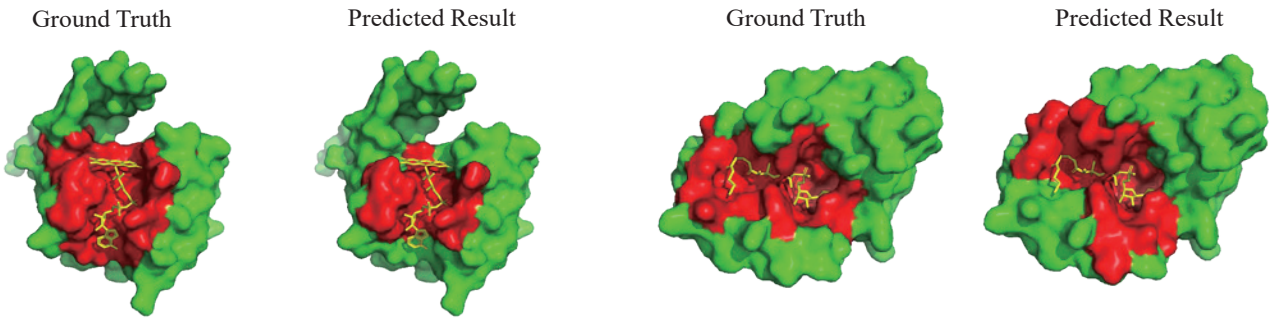
- Compare the proposed model with the sequence-based model DeepCseqSite and Birds, as well as the structure-based model Fpocket
  - ✓ The commonly used metric is MCC (Matthews correlation coefficient), Precision, Recall





Visualization Using PyMOL

- Comparison between ground truth and predicted results
  - ✓ with PDB IDs 2ed4 and 1b6t
  - ✓ the yellow structure represents the ligand
  - ✓ the red regions indicate the binding residues interacting with the ligand



PDB id : 2ed4_A

PDB id : 1b6t_B

Kang *et al.*, (In preparation) 2025