Lecture03. Deep Learning in Hit Discovery

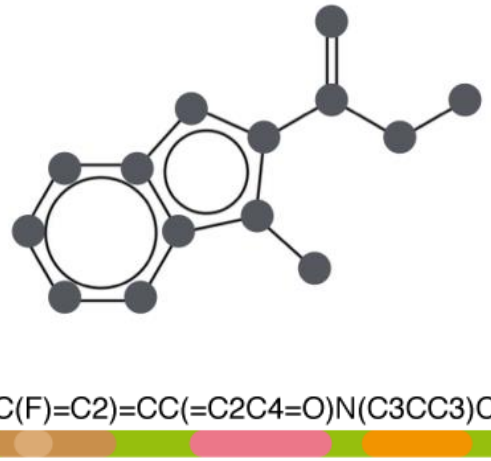
HITS 임 재 창

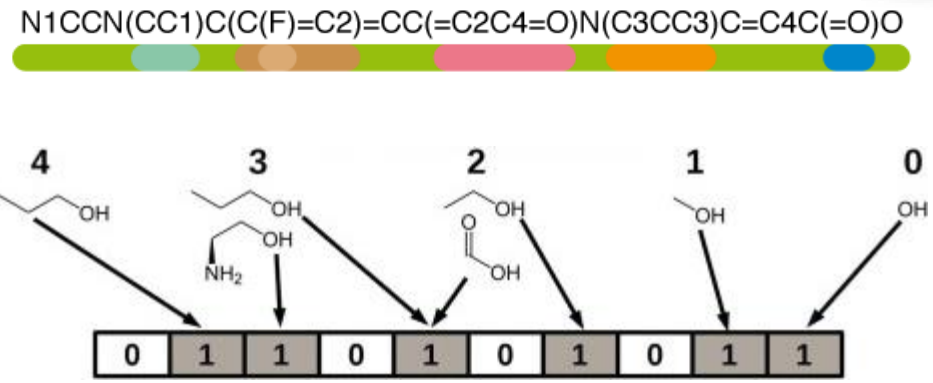
목차

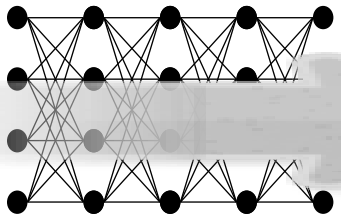
- Ligand based DTI prediction
- Sequence-based DTI prediction
- Structure-based DTI prediction (3DCNN and GNN)

Ligand based DTI prediction

- Ligand 2D structure (Graph, Smiles, fingerprint)를 이용해서 타겟 단백질에 대한 활성을 예측







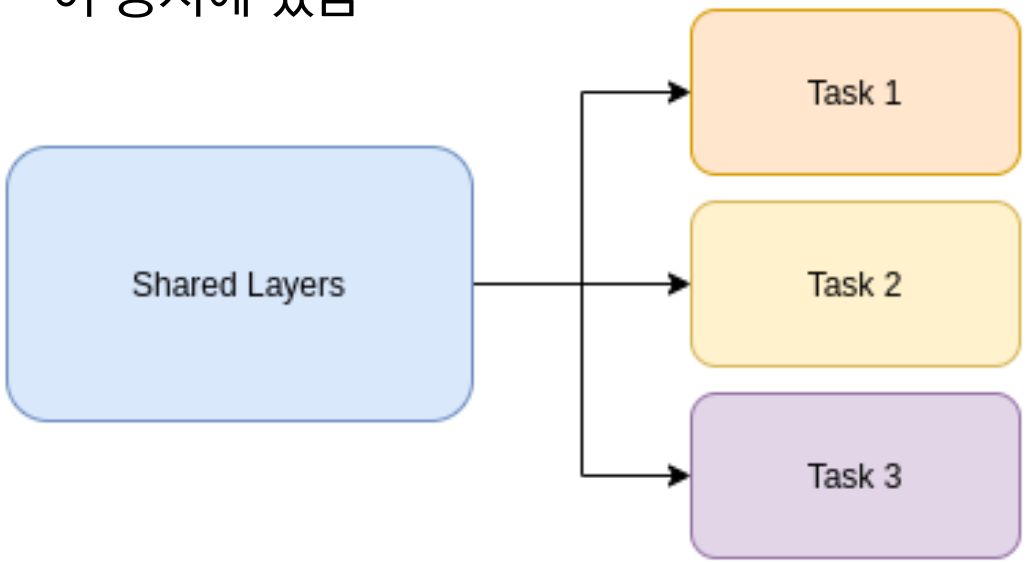
EGFR

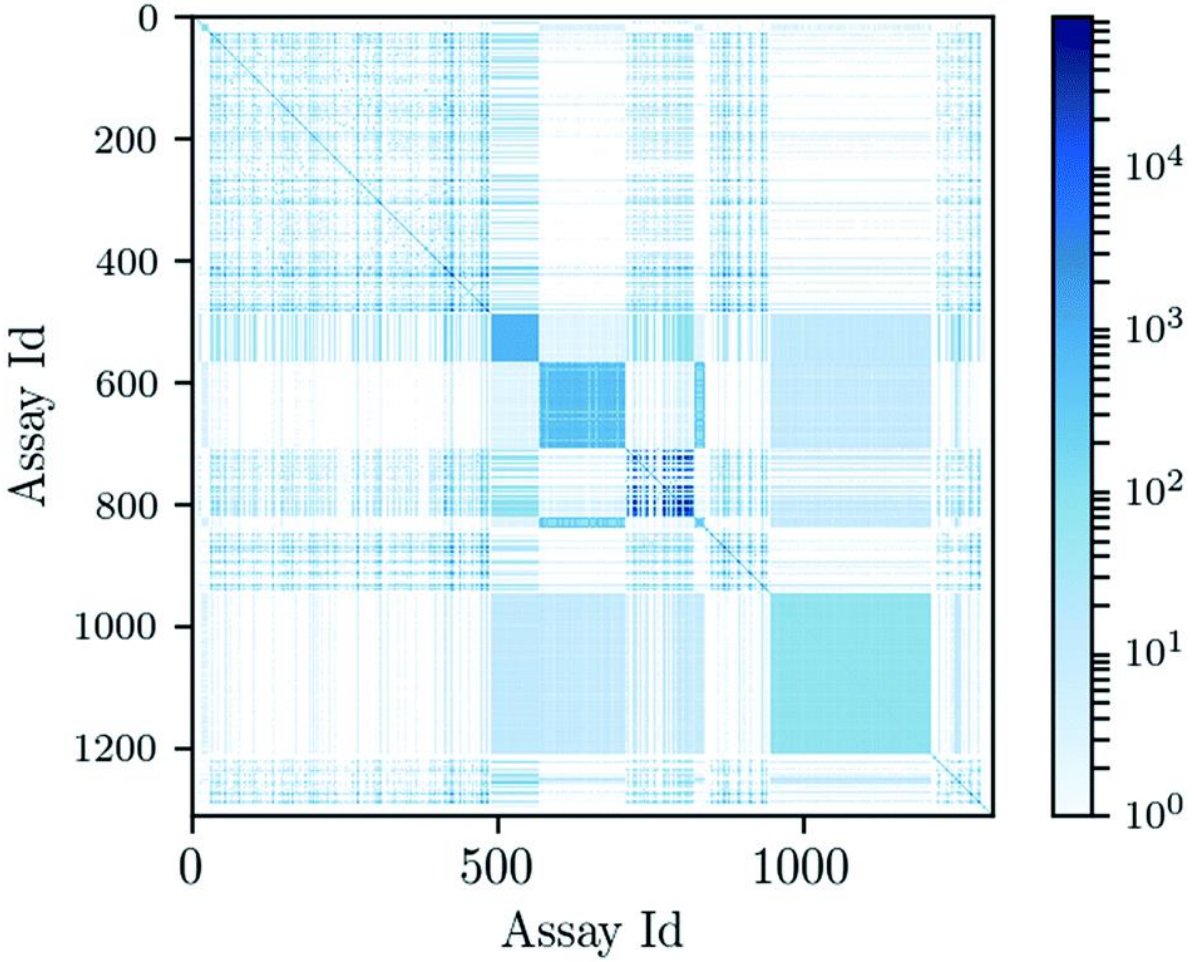
0: active or

1: inactive

Multitask learning

- 여러 타겟에 대한 활성 데이터를 동시에 학습.
- 공유 layers와 비공유 layers로 구성됨
- 데이터 개수가 늘어나므로, over-fitting의 위험을 줄일 수 있음
- 실제로 상당수의 분자들은 여러 타겟에 대한 활성값이 동시에 있음





Ligand based DTI prediction

	FNN	SVM	RF	KNN	NB	SEA	GC	Weave	SmilesLSTM
StaticF	0.687 ± 0.131	0.668 ± 0.128	0.665 ± 0.125	0.624 ± 0.120					
SemiF	0.743 ± 0.124	0.704 ± 0.128	0.701 ± 0.119	0.660 ± 0.119	0.630 ± 0.109				
ECFP6	0.724 ± 0.125	0.715 ± 0.127	0.679 ± 0.128	0.669 ± 0.121	0.661 ± 0.119	0.593 ± 0.096			
DFS8	0.707 ± 0.129	0.693 ± 0.128	0.689 ± 0.120	0.648 ± 0.120	0.637 ± 0.112				
ECFP6 + ToxF	0.731 ± 0.126	0.722 ± 0.126	0.711 ± 0.131	0.675 ± 0.122	0.668 ± 0.118				
Graph							0.692 ± 0.125	0.673 ± 0.127	
SMILES									0.698 ± 0.130

- 1000개이상의 타겟, 500,000개 이상의 활성 데이터에 대해서 DL과 기존 machine learning 방법의 성능을 비교
- 모든 표현형에 대해서 DL이 기존 방법론들에 비해서 유의미한 차이로 우수한 성능을 보여줌.

Ligand based DTI prediction의 장단점

장점

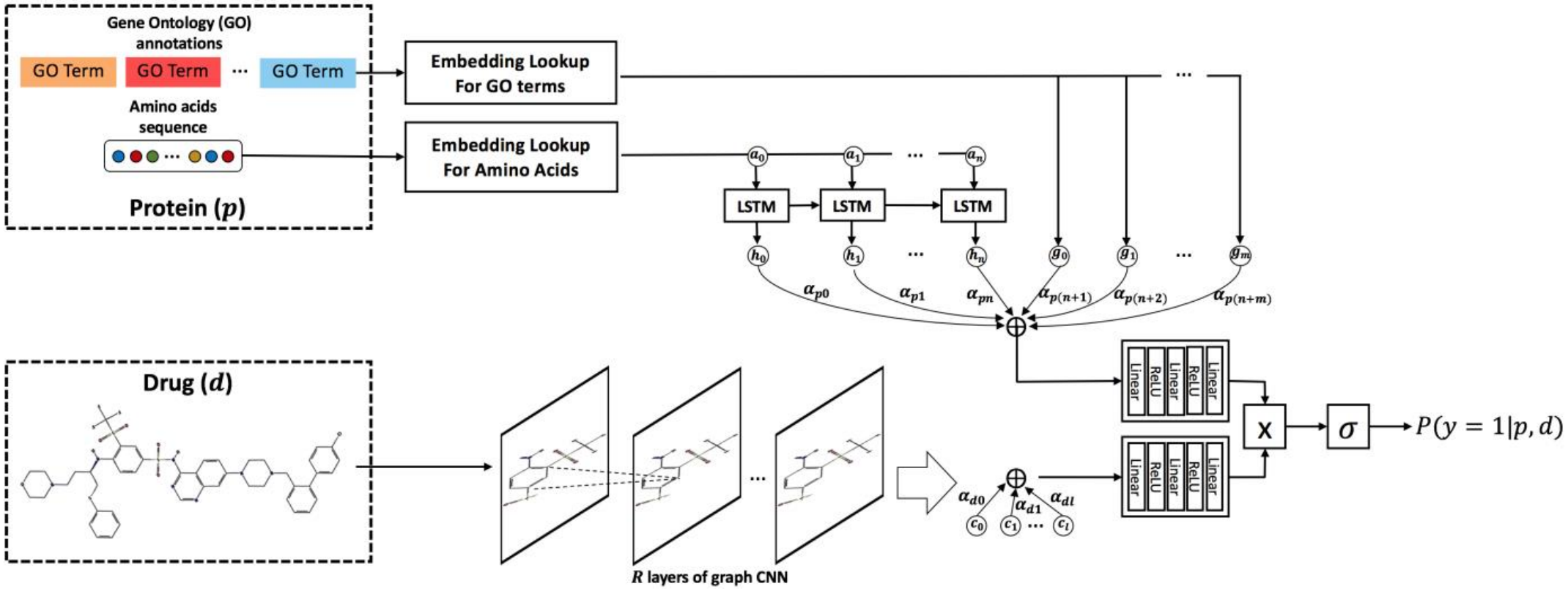
- Single modality이므로 학습이 쉬움.

단점

- Novel 타겟에 대해서 적용이 불가능.
- 신규타겟에 새로 적용시키기 위해서는 대량의 데이터가 필요함

Sequence based DTI prediction

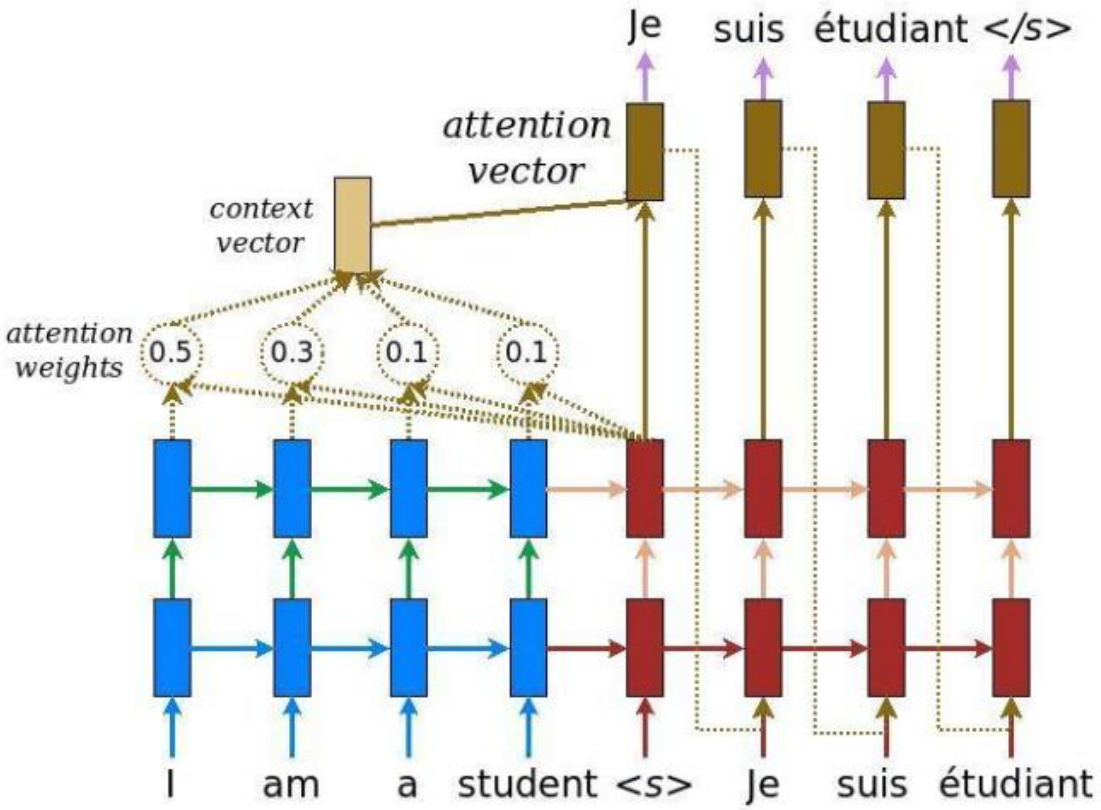
- Protein과 ligand의 결합구조 없이 각각의 2D or 3D구조를 이용

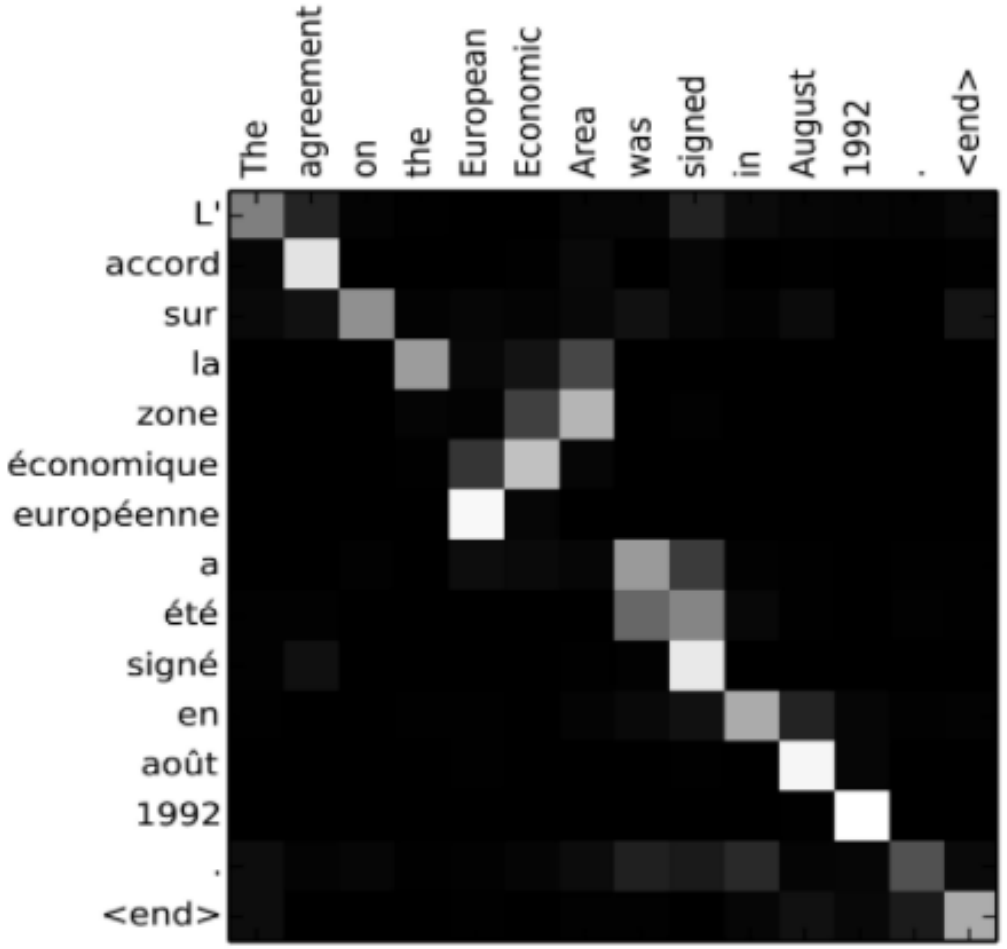


DOI:10.24963/ijcai.2018/468

Attention mechanism

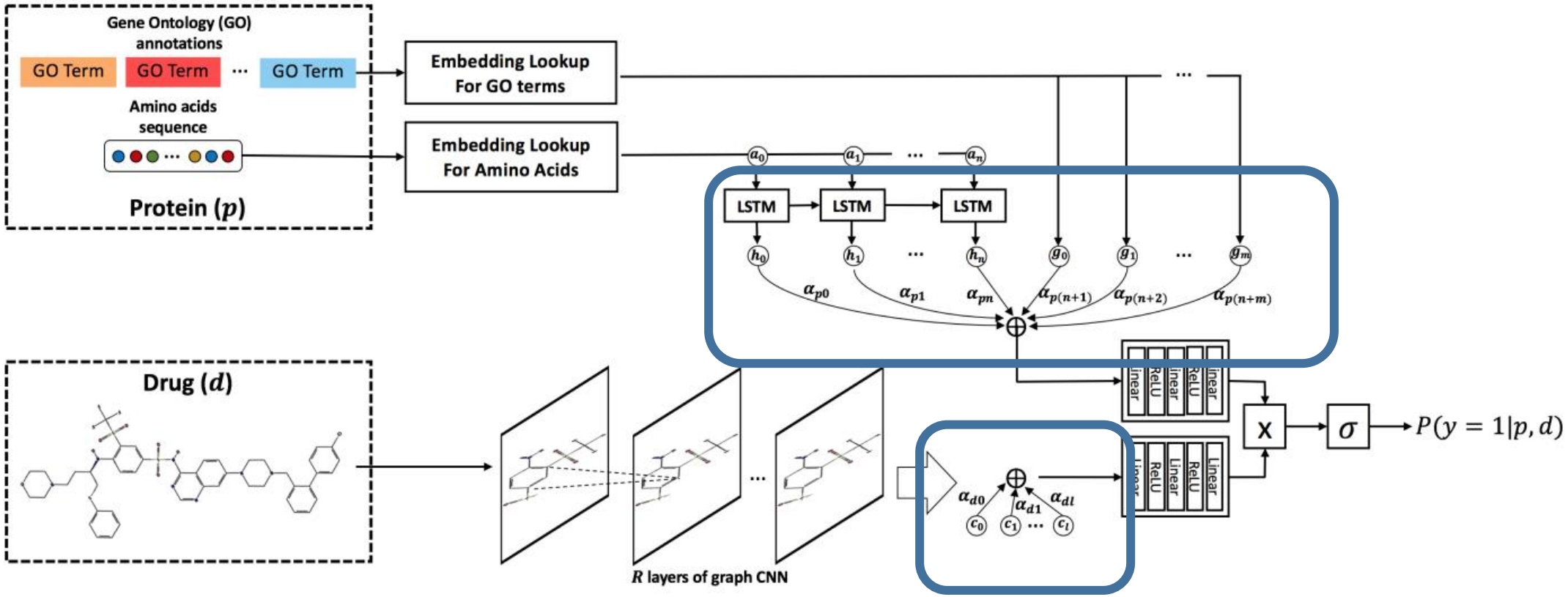
- 중요한 vector에 weight를 가함





Attention mechanism

- 중요한 vector에 weight를 가함



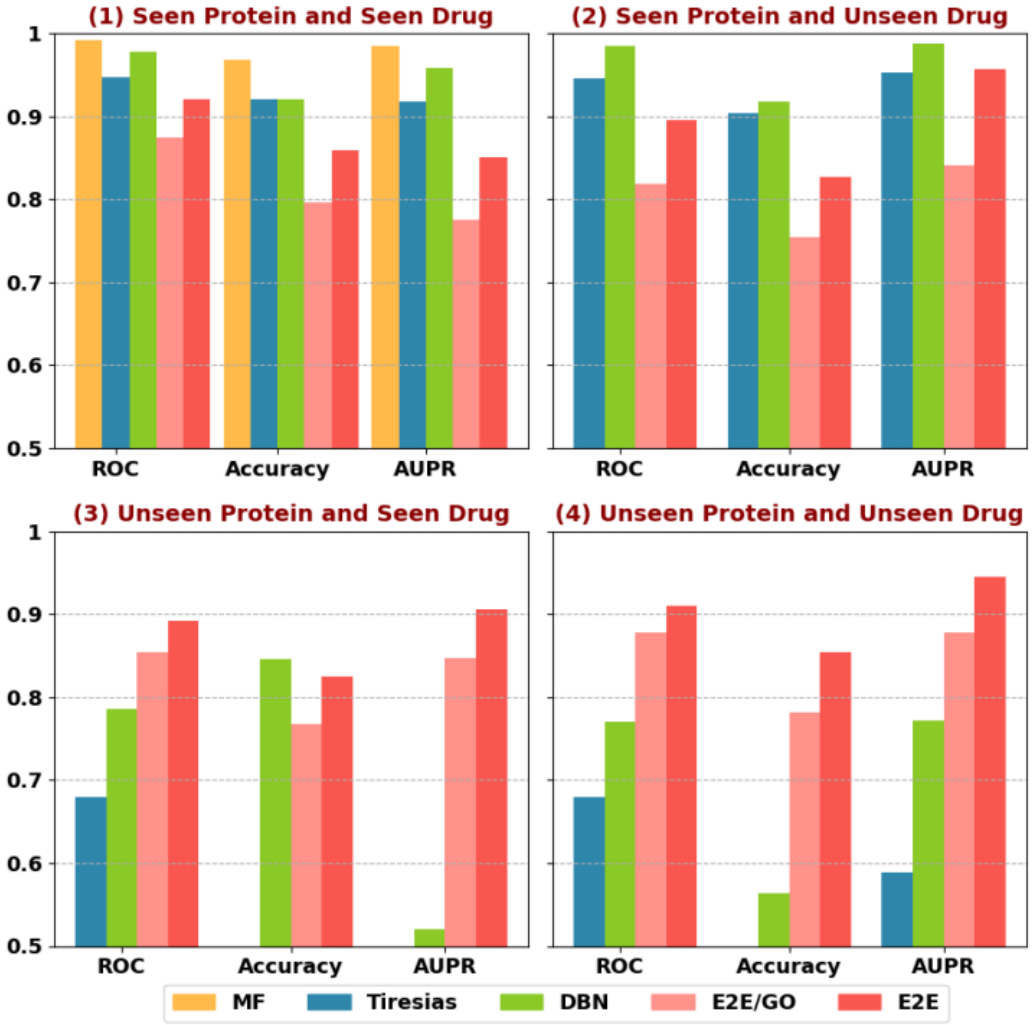
Dataset preparation

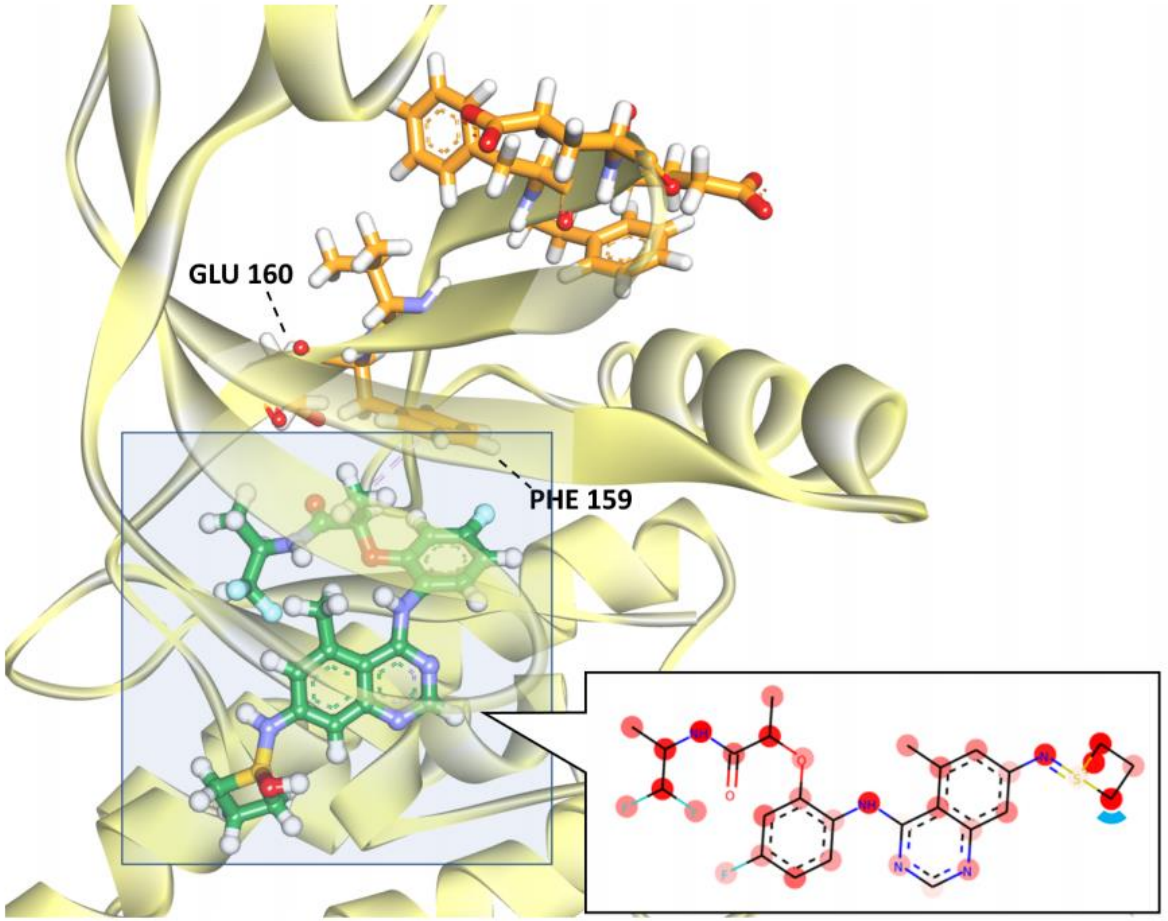
1. Record has chemical identifier (PubChem CID), and the small molecule has chemical structure represented by SMILES[3].
2. Record has protein identifier (Uniprot ID), and the protein has both sequence representation and Gene Ontology annotations [Ashburner *et al.*, 2000].
3. Record has IC50 value, a primary measure of binding effectiveness.
4. The chemical molecule weight is less than 1,000Da, due to our focus on small molecule drugs.
5. By following the activity threshold discussion in [Wang *et al.*, 2016], record is positive if its IC50 is less than 100nm, negative if IC50 greater than 10,000nm.

Dataset	Protein	Drug	Positive	Negative
Train	758	43,160	28,240	21,915
Dev	472	5,077	2,831	2,776
Test	466	5,016	2,706	2,802

Protein	Sequence	Embedding Size	16
		Hidden Dimension	16
		Embedding Dropout	0.1
	GO	Embedding Size	16
		Embedding Dropout	0.1
Drug	Graph CNN	Hidden Dimension	64
Siamese		Hidden Size	32
		Dropout	0.1
γ		0.0005	

Results


(1) Seen Protein and Seen Drug
(2) Seen Protein and Unseen Drug
(3) Unseen Protein and Seen Drug
(4) Unseen Protein and Unseen Drug
1
0.9
0.8
0.7
0.6
0.5
ROC
Accuracy
AUPR
MF
Tiresias
DBN
E2E/GO
E2E


GLU 160
PHE 159

Sequence based DTI prediction의 장단점

장점

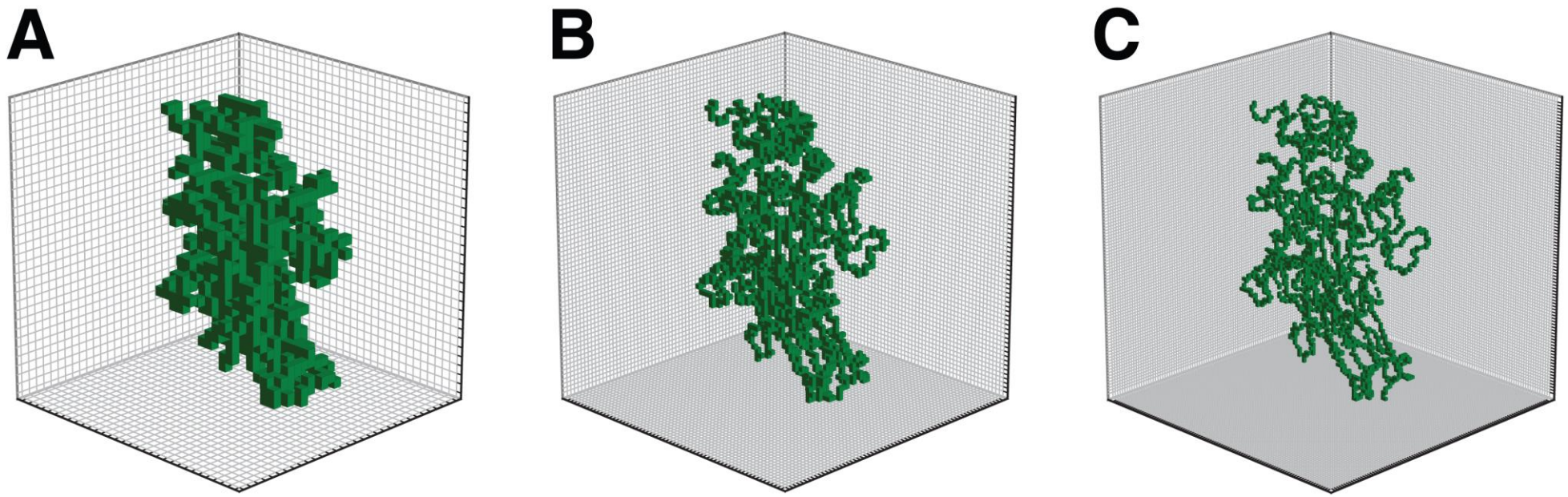
- Protein의 3차원 구조 및 ligand와의 결합구조를 필요로 하지 않으므로 적용할 수 있는 타겟의 범위가 넓음.

단점

- (장점이면서) 3차원 결합해석이 불가능
- Binding site가 여러 위치일 때 적용 불가능
- Attention기반 상호작용 해석은 cherry picking일 가능성이 높음

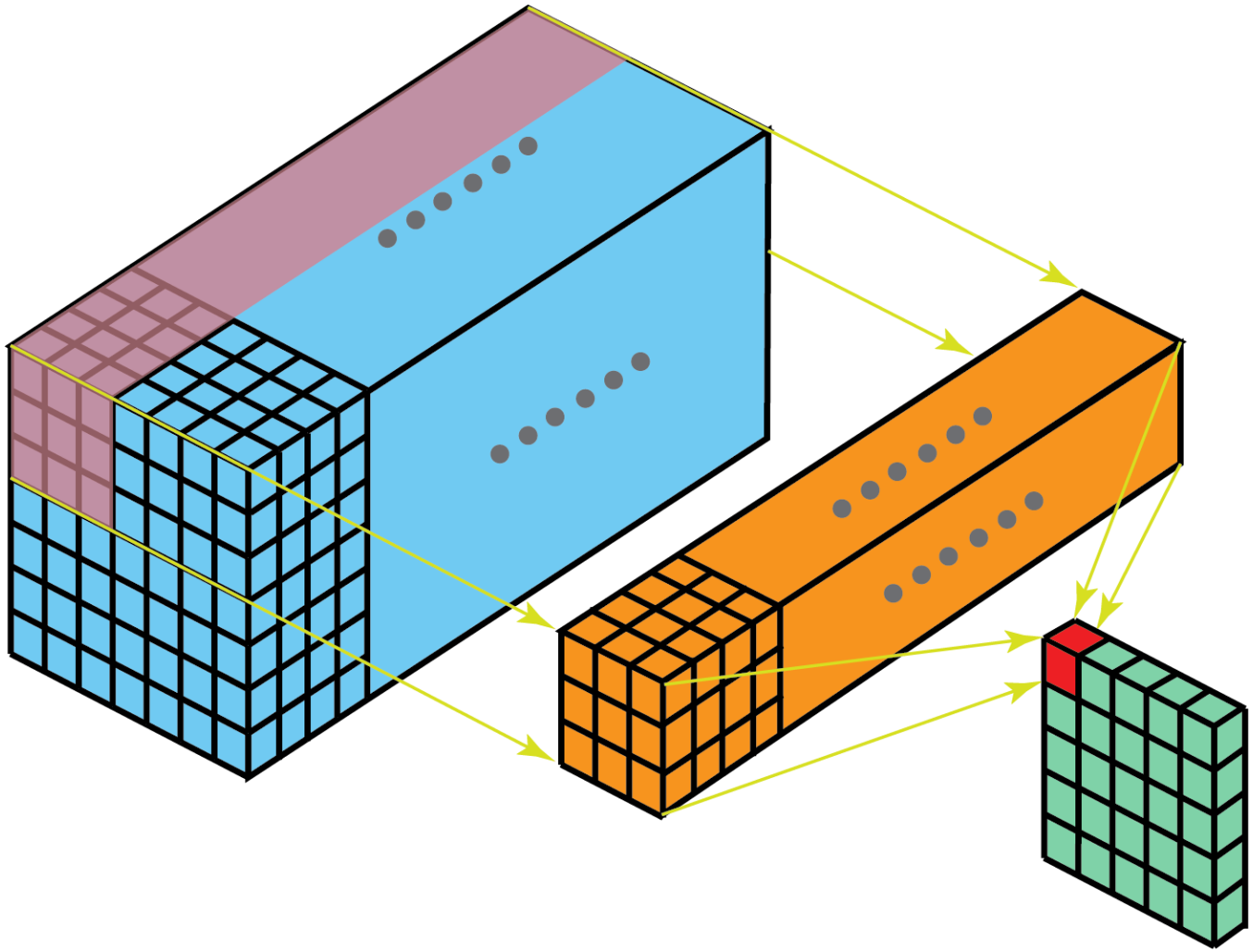
Structure based DTI prediction (3DCNN)

- Protein-ligand complex의 binding site 3차원 구조를 rectangular grid위에 표현한 뒤 3D convolutional neural network를 적용



Amidi A, Amidi S, Vlachakis D, Megalooikonomou V, Paragios N, Zacharaki EI. 2018. EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation. PeerJ 6:e4750 https://doi.org/10.7717/peerj.4750

3D convolutional neural network



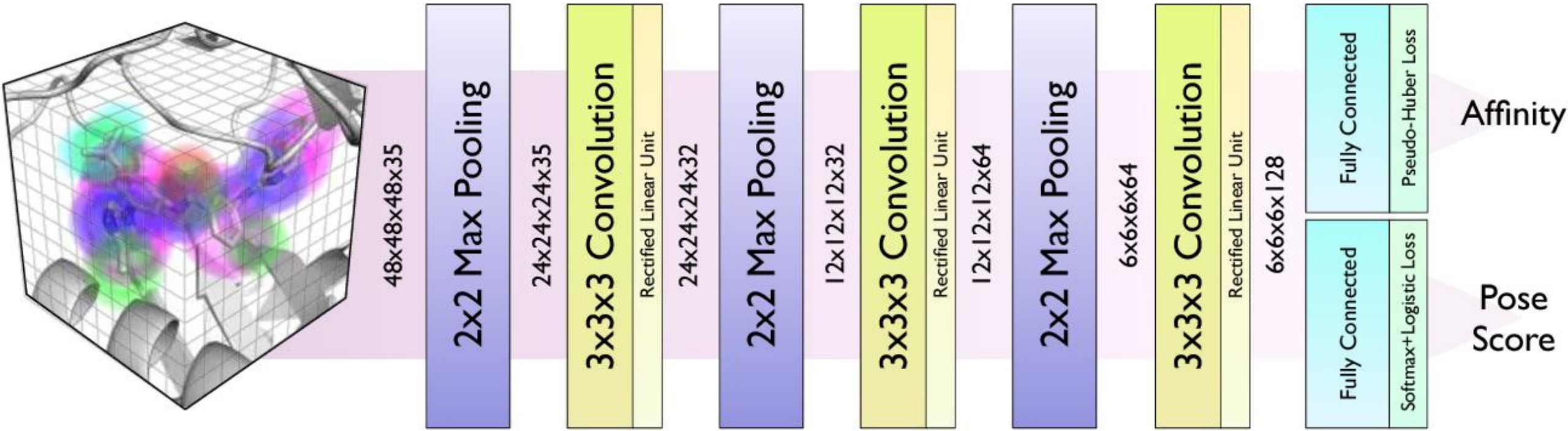
2D convolution :

$$v_{ij}^{xy} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right)$$

3D convolution :

$$v_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right)$$

3D convolutional neural network



J Mol Graph Model. 2018 September ; 84: 96–108

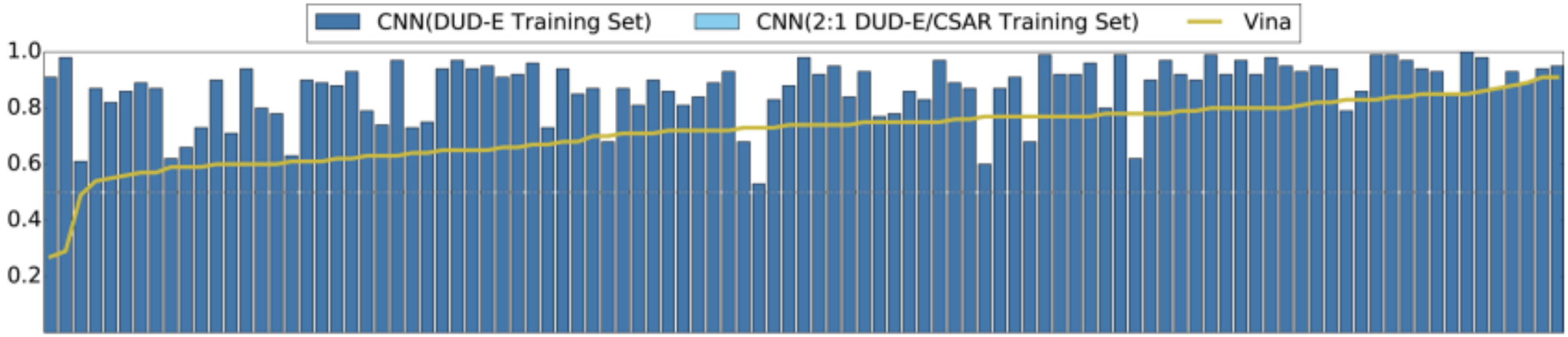
Results

AUC		> 0.5	> 0.6	> 0.7	> 0.8	> 0.9
ChEMBL-20 PMD	AtomNet	49	44	36	24	10
	Smina	38	10	4	1	0
DUDE-30	AtomNet	30	29	27	22	14
	Smina	29	25	14	5	1
DUDE-102	AtomNet	102	101	99	88	59
	Smina	96	84	53	17	1
ChEMBL-20 inactives	AtomNet	149	136	105	45	10
	Smina	129	81	31	4	0

Adjusted-LogAUC		> 0.0	> 0.1	> 0.2	> 0.3	> 0.4
ChEMBL-20 PMD	AtomNet	49	44	36	27	20
	Smina	35	8	2	1	0
DUDE-30	AtomNet	30	27	22	17	10
	Smina	29	19	8	2	1
DUDE-102	AtomNet	102	99	88	69	43
	Smina	94	65	28	5	1
ChEMBL-20 inactives	AtomNet	147	107	36	10	2
	Smina	123	35	5	0	0

arXiv:1510.02855

Results

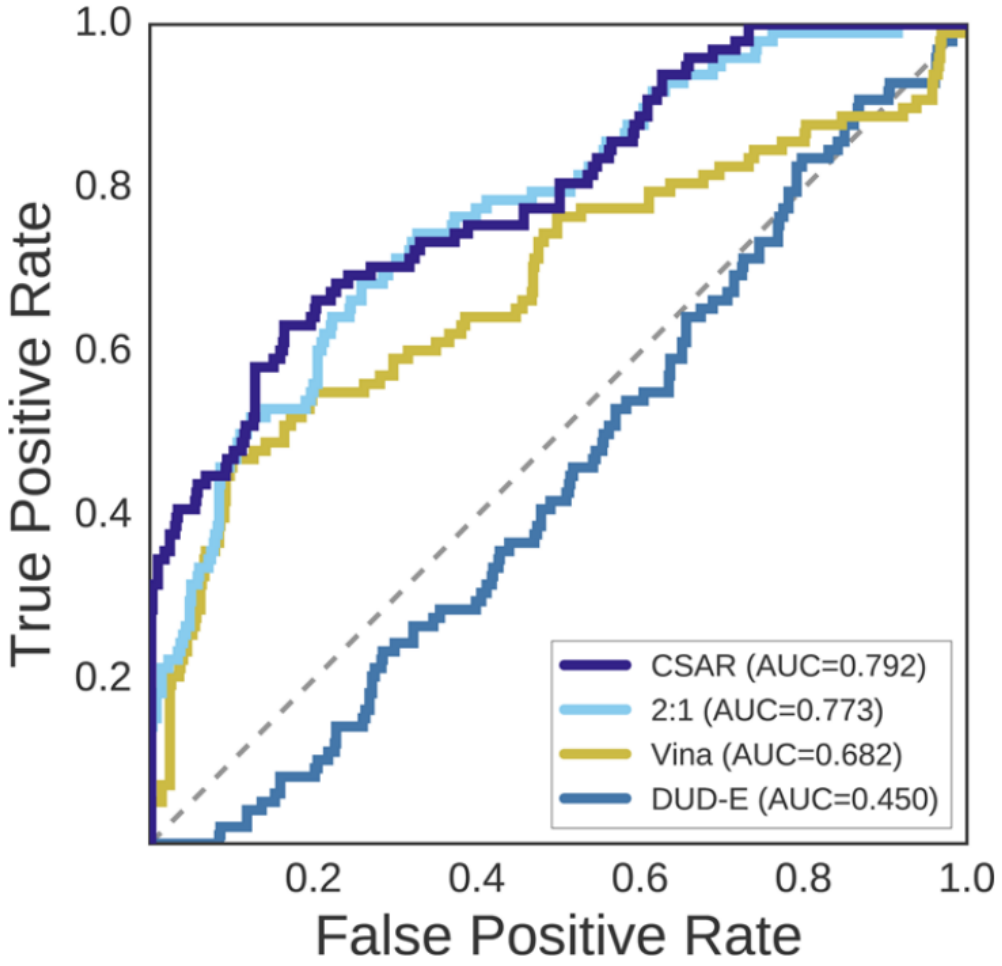


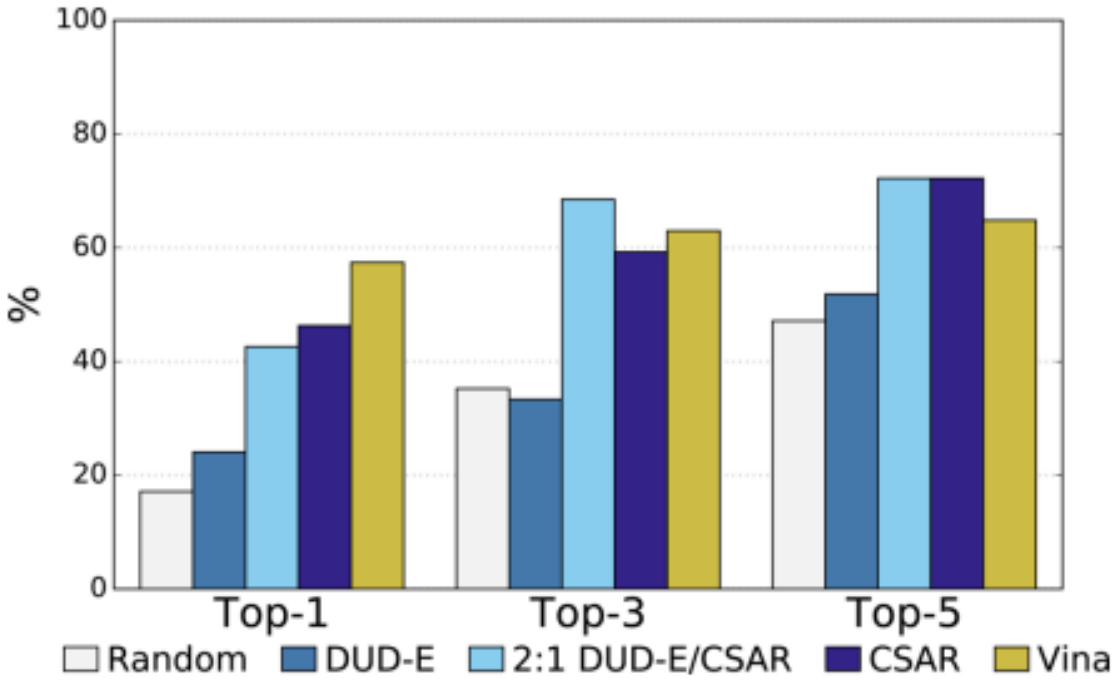
metric	DUD-E	2:1 D/C	Vina	RF-Score	NNScore
AUC	**0.868**	0.804	0.716	0.622	0.584
0.5% RE	**42.559**	22.366	9.139	5.628	4.166
1.0% RE	**29.654**	16.274	7.321	4.274	2.980
2.0% RE	**19.363**	11.888	5.881	3.499	2.460
5.0% RE	**10.710**	7.376	4.444	2.678	1.891

J. Chem. Inf. Model. 2017, 57, 942−957

Results

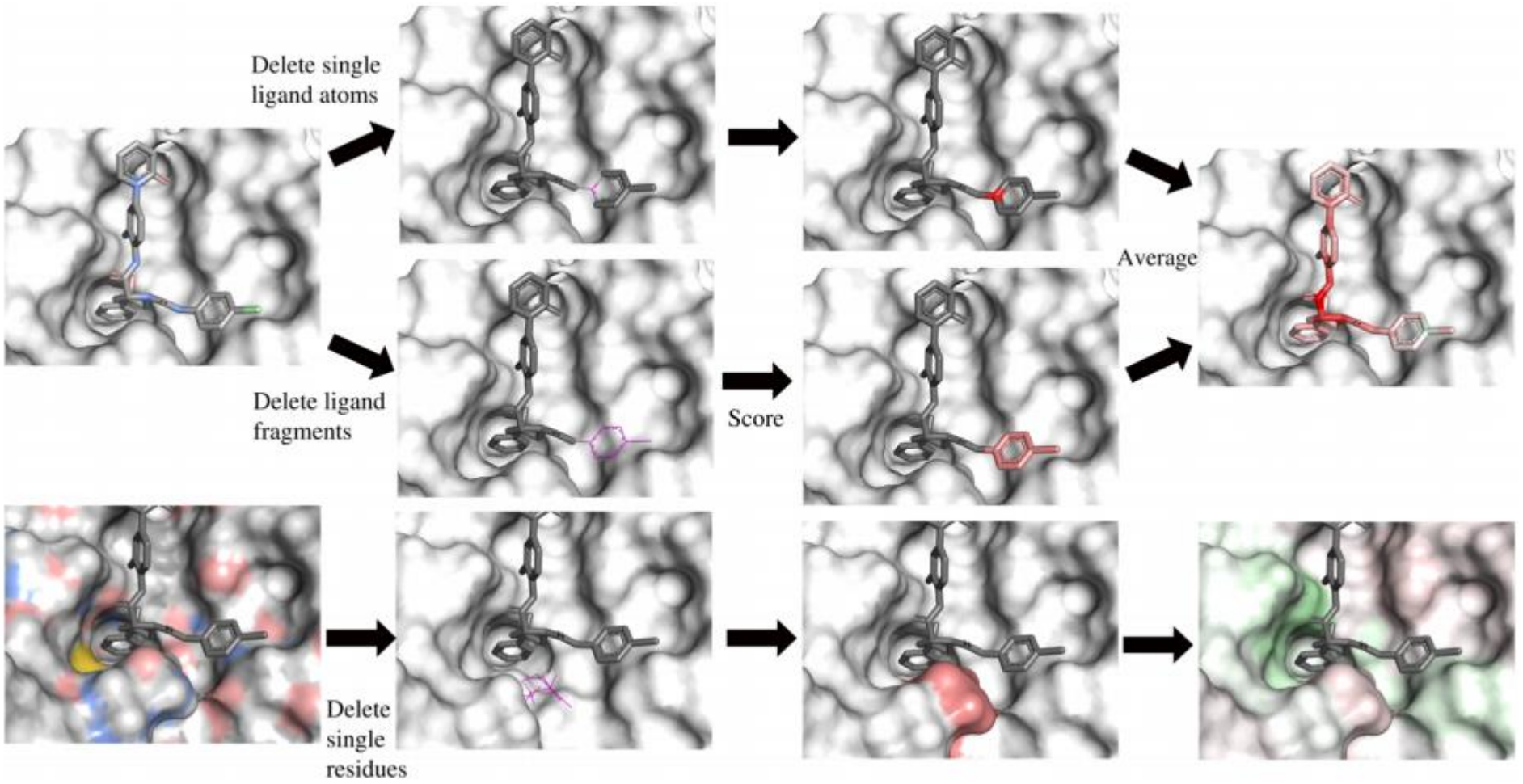
- 3차원 구조를 사용하기 때문에 pose prediction도 가능함





J. Chem. Inf. Model. 2017, 57, 942−957

Results



J. Chem. Inf. Model. 2017, 57, 942−957

Structure based DTI prediction (3DCNN)의 장단점

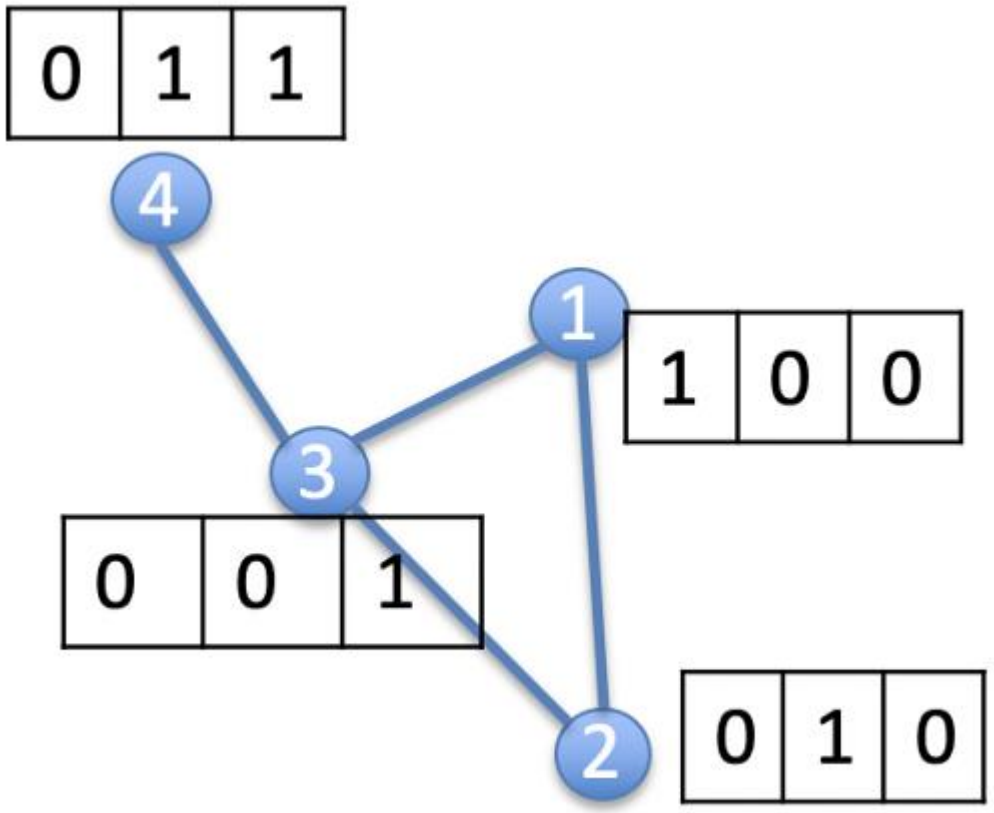
장점

- Protein-ligand 결합의 원천이 되는 3차원 결합구조를 반영할 수 있음.
- 3차원 구조를 반영하기 때문에 pose prediction이 가능함
- Torsion angle과 같은 복잡한 3차원적 특징을 스스로 학습가능

단점

- Grid를 사용하기 때문에, grid 간격에 따라 정보의 손실이 있음
- 3차원 구조를 사용하지만 protein residue와의 해석에 한계가 있음
- Rotational 및 translational invariant 하지 않음

Structure based DTI prediction (GNN)



1	1	1	0
1	1	1	0
1	1	1	1
0	0	1	1

Adjacency matrix (A)

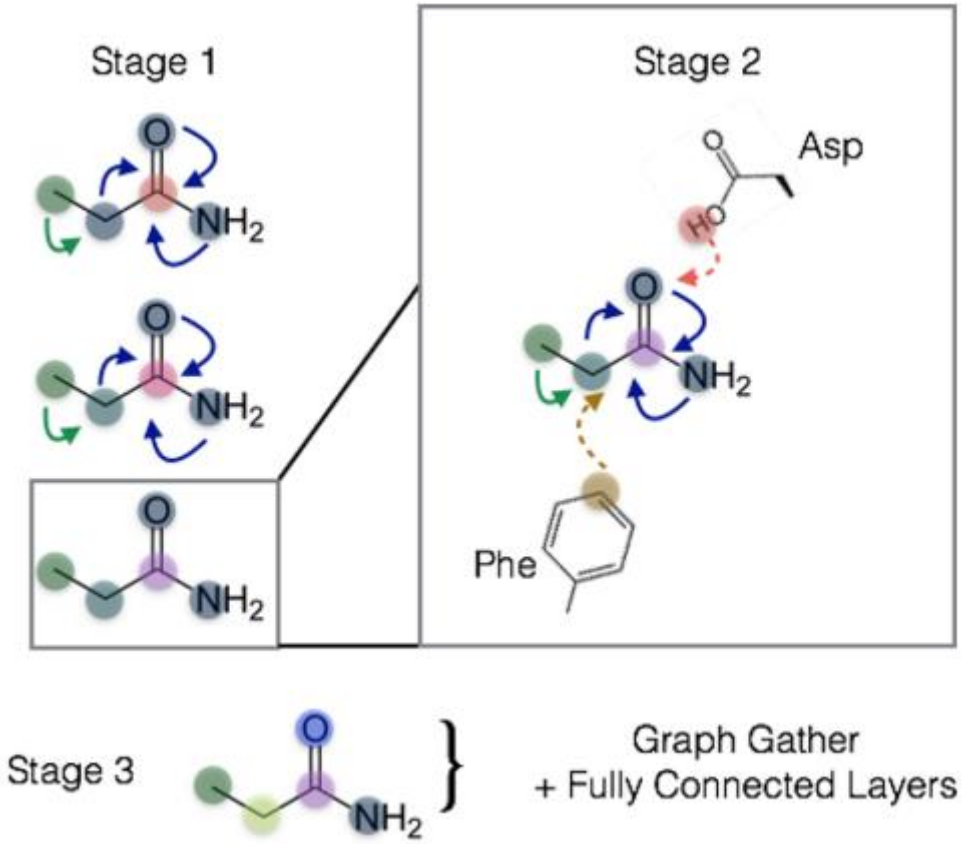
1	0	0
0	1	0
0	0	1
0	1	1

Feature matrix (X)

- A Massage Passing function: $m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_u^t, e_{uv})$
- A Node Update function: $h_v^{l+1} = U_t(h_v^t, m_v^{t+1})$
- A Readout function (for graph classification): $\hat{y} = R(\{h_v^T | v \in G\})$

Structure based DTI prediction (GNN)

- Protein-ligand 3차원 결합구조를 2D 그래프에 어떻게 표현할 것인가?



- Interaction을 하고 있는 Protein과 ligand 원자들을 edge로 연결
- Edge에 distance, interaction type (π–π stacking, hydrogen bonds, hydrophobic
- contact) 정보를 표시
- Intramolecular interaction 학습, intermolecular interaction 학습, 2가지 stage로 학습이 진행됨.

ACS Cent. Sci. 2018, 4, 11, 1520–1530

PotentialNet

PotentialNet, stage 1

$$h_i^{(b_1)} = \mathrm{GRU}\left(x_i, \sum_{\epsilon}^{N_{et}} \sum_{j \in N^{(\epsilon)}(v_i)} NN^{(\epsilon)}(x_j)\right)$$

$$\vdots$$

$$h_i^{(b_K)} = \mathrm{GRU}\left(h_i^{(b_{K-1})}, \sum_{\epsilon}^{N_{et}} \sum_{j \in N^{(\epsilon)}(v_i)} NN^{(\epsilon)}(h_j^{(b_{K-1})})\right)$$

$$h^{(b)} = \sigma(i^{(b)}(h^{(b_K)}, x)) \odot (j^{(b)}(h^{(b_K)})) \in \mathbb{R}^{(N \times f_b)}$$

PotentialNet, stage 2

$$h_i^{(sp_1)} = \mathrm{GRU}\left(h_i^{(b)}, \sum_{\epsilon}^{N_{et}} \sum_{j \in N^{(\epsilon)}(v_i)} NN^{(\epsilon)}(h_j^{(b)})\right)$$

$$\vdots$$

$$h_i^{(sp_K)} = \mathrm{GRU}\left(h_i^{(sp_{K-1})}, \sum_{\epsilon}^{N_{et}} \sum_{j \in N^{(\epsilon)}(v_i)} NN^{(\epsilon)}(h_j^{(sp_{K-1})})\right)$$

$$h^{(sp)} = \sigma(i^{(sp)}(h^{(sp_K)}, h^{(b)})) \odot (j^{(sp)}(h^{(sp_K)})) \in \mathbb{R}^{(N \times f_b)}$$

PotentialNet, stage 3

$$h^{(FC_0)} = \sum_{j=1}^{N_{lig}} h_j^{(sp)}$$

$$h^{(FC_1)} = ReLU(W^{(FC_1)} h^{(FC_0)})$$

$$\vdots$$

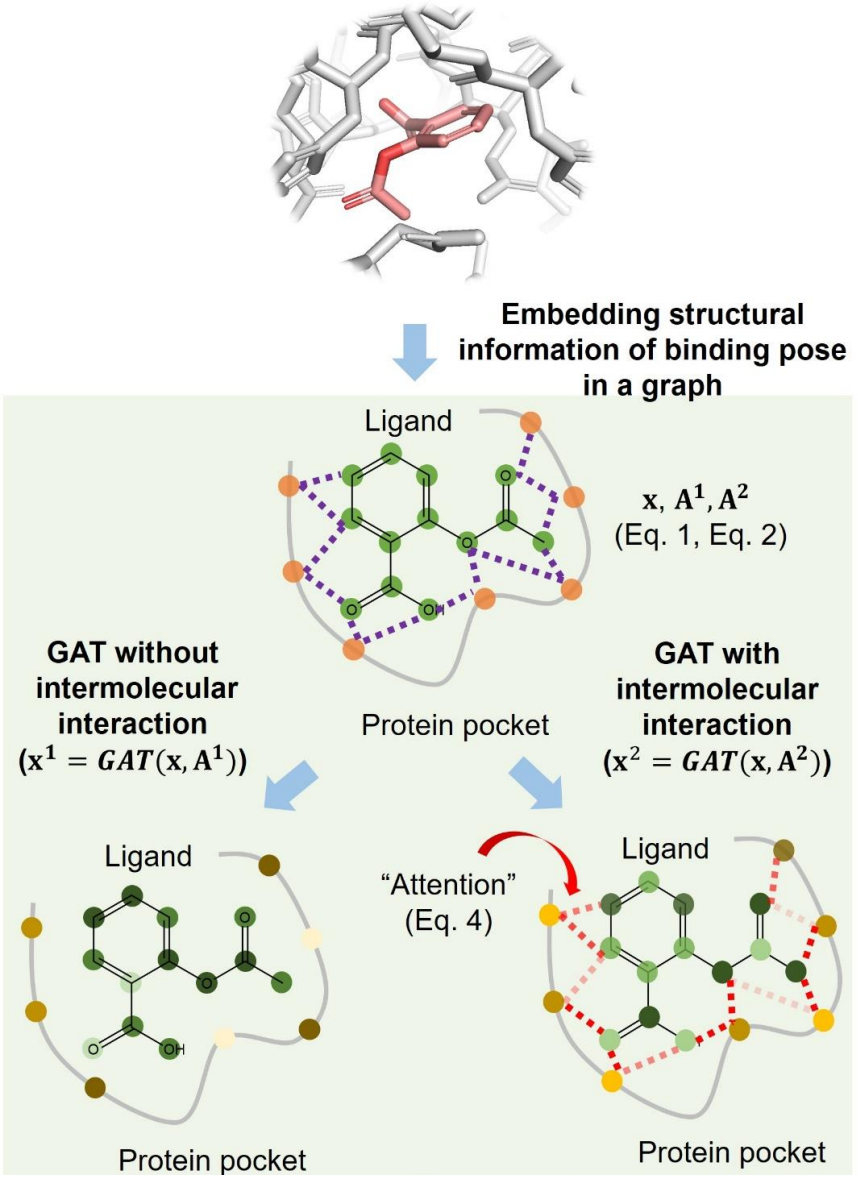
$$h^{(FC_K)} = W^{(FC_K)} h^{(FC_{K-1})}$$

model	Test R^2	Test $EF_{\chi}^{(R)}$	Test Pearson	Test Spearman	Test MUE
PotentialNet	**0.629** (0.044)	**1.576** (0.053)	**0.823** (0.023)	**0.805** (0.019)	**1.553** (0.125)
ligand-only PotentialNet	0.500 (0.010)	1.498 (0.411)	0.733 (0.007)	0.726 (0.005)	1.700 (0.067)
RF-Score	0.594 (0.005)	0.869 (0.090)	0.779 (0.003)	0.757 (0.005)	1.542 (0.046)
X-Score	0.517	0.891	0.730	0.751	1.751

ACS Cent. Sci. 2018, 4, 11, 1520–1530

Distance aware graph attention neural network

- Deep learning의 가장 큰 장점은 hand-crafted feature없이 raw data에서 key representation을 학습하는 것
- Edge feature를 사람이 정하는 것이 아니라, raw data로 부터 학습하는 것이 더 좋은 성능을 보여줄 수 있음



J. Chem. Inf. Model. 2019, 59, 9, 3981–3988J. Chem. Inf. Model. 2019, 59, 9, 3981–3988

Distance aware graph attention neural network

$$\mathbf{A}^1_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected by covalent bond or } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{A}^2_{ij} = \begin{cases} \mathbf{A}^1_{ij} & \text{if } i, j \in \text{protein atoms or } i, j \in \text{ligand atoms} \\ e^{-(d_{ij}-\mu)^2/\sigma} & \text{if } d_{ij} < 5 \text{ Å and } i \in \text{ligand atoms and } j \in \text{protein atoms, or if } d_{ij} < 5 \text{ Å and } i \in \text{protein atoms and } j \in \text{ligand atoms} \\ 0 & \text{otherwise} \end{cases}$$

J. Chem. Inf. Model. 2019, 59, 9, 3981–3988

Distance aware graph attention neural network

- Node states can be represented as follow:

$$\mathbf{x}^{in} = \{x_1^{in}, x_2^{in}, ..., x_N^{in}\} \text{ with } x \in \mathbb{R}^F$$

- The node states are transformed by a neural network

$$x_i' = \mathbf{W} x_i^{in}$$

- Attention coefficient is calculated from two node states

$$e_{ij} = x_i'^T \mathbf{E} x_j' + x_j'^T \mathbf{E} x_i',$$

- Attention coefficient is normalized with consideration of adjacency matrix

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{j \in N_i} \exp(e_{ij})} \mathbf{A}_{ij}.$$

Distance aware graph attention neural network

- New node states are computed by weighted linear combination of current node states

$$x_i'' = \sum_{j \in N_i} a_{ij} x_j'.$$

- Final node states are computed by linear combination of current and new node states

$$x_i^{out} = z_i x_i' + (1 - z_i) x_i''$$

$$z_i = \sigma(\mathbf{U}(x_i \| x_i') + b),$$

Results

	AUROC	adjusted LogAUC	PRAUC	sensitivity	specificity	balanced accuracy
ours	**0.968**	**0.633**	**0.697**	**0.826**	0.967	**0.909**
ours w/o attention	0.936	0.577	0.623	0.758	**0.970**	0.888
docking	0.689	0.153	0.016			
Atomnet[19]	0.855	0.321				
Ragoza et al.[22]	0.868					
Torng et al.[40]	0.886					
Gonczarek et al.[17]	0.904					

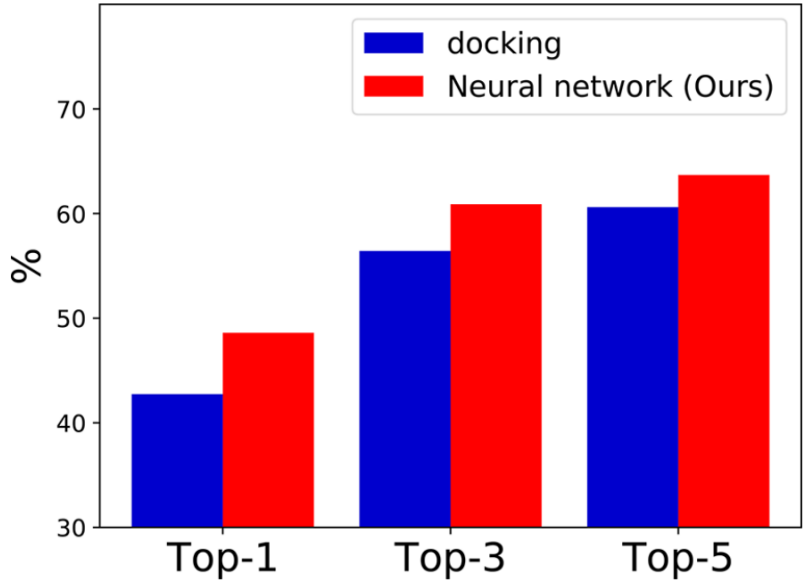
[a]We note that the division of the training and test sets may be different for each model.

	0.5%	1.0%	2.0%	5.0%
ours	**124.031**	**69.037**	**38.027**	**16.910**
ours w/o attention	107.734	61.346	34.326	16.029
docking	11.538	9.749	6.153	3.789
Ragoza et al.[22]	42.559	29.654	19.363	10.710
Torng et al.[40]	44.406	29.748	19.408	10.735

[a]The RE score indicates the ratio of the true positive rate (TPR) to the false positive rate (FPR) at a certain FPR value.

Results

	AUROC	PRAUC
ours	**0.935**	**0.772**
ours w/o attention	0.927	0.698
docking	0.825	0.509



Structure based DTI prediction (GNN)의 장단점

장점

- Protein과 ligand atom과 bond를 explicit하게 표현 할 수 있음
- Protein과 ligand를 compact하게 표현할 수 있음
- Rotational 및 translational invariant함

단점

- 3차원 구조를 직관적으로 표현하기 어려움

Thank you