Lecture 06.
Better generalization of deep learning models in drug discovery

HITS 임 재 창

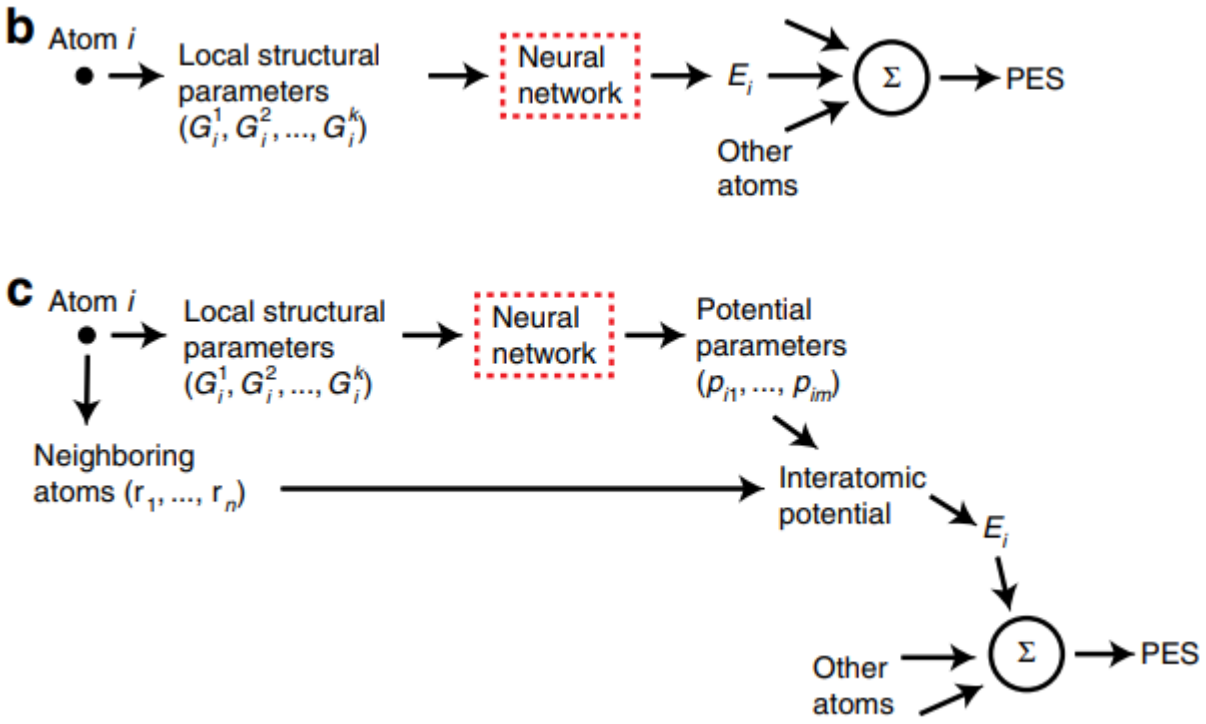
목차

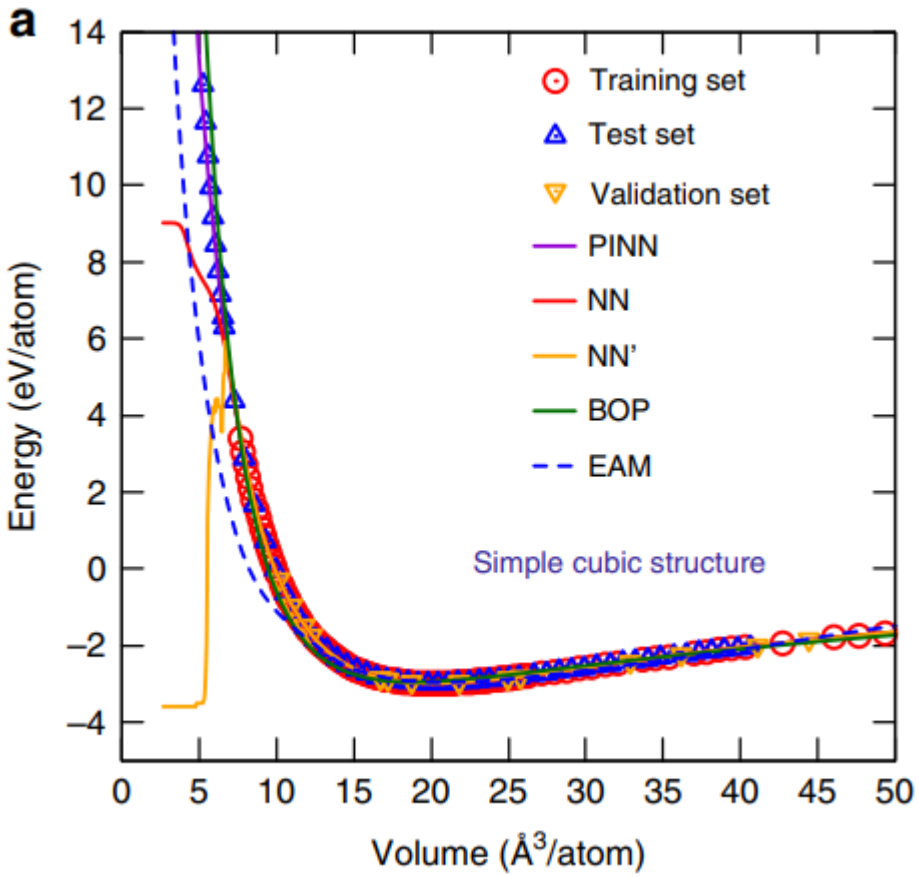
- Physics informed neural network
- Data augmentation
- Uncertainty quantification for reliable prediction
- Semi-supervised learning
- Transfer learning

Physics informed neural network

- 대부분의 경우 분자의 성질을 예측하는 딥러닝 모델을 개발하기에 데이터가 부족함
- 숫자가 부족하고 dataset들이 많은 intrinsic bias를 내포하고 있음
- High quality의 추가 데이터를 생성하기 어려움
- 분자의 물리화학 성질을 학습하는 경우 관련된 물리법칙 혹은 모델들이 존재함
- 이러한 물리법칙과 모델들은 principle에 기반하기 때문에 높은 일반화 성능을 보여줌

Physics informed neural network





Nature Communications volume 10, Article number: 2339 (2019)

Physics informed DTI prediction

$$E^{total} = \frac{E^{vdw} + E^{hbond} + E^{metal} + E^{hydrophobic}}{T^{rotor}}$$

$$d'_{ij} = r_i + r_j + c \cdot b_{ij}$$

- Van der waal interaction: LJ potential을 이용하여 계산

$$E^{vdw} = \sum_{i,j} c_{ij} \left[\left(\frac{d'_{ij}}{d_{ij}} \right)^{12} - 2 \left(\frac{d'_{ij}}{d_{ij}} \right)^{6} \right]$$

Physics informed DTI prediction

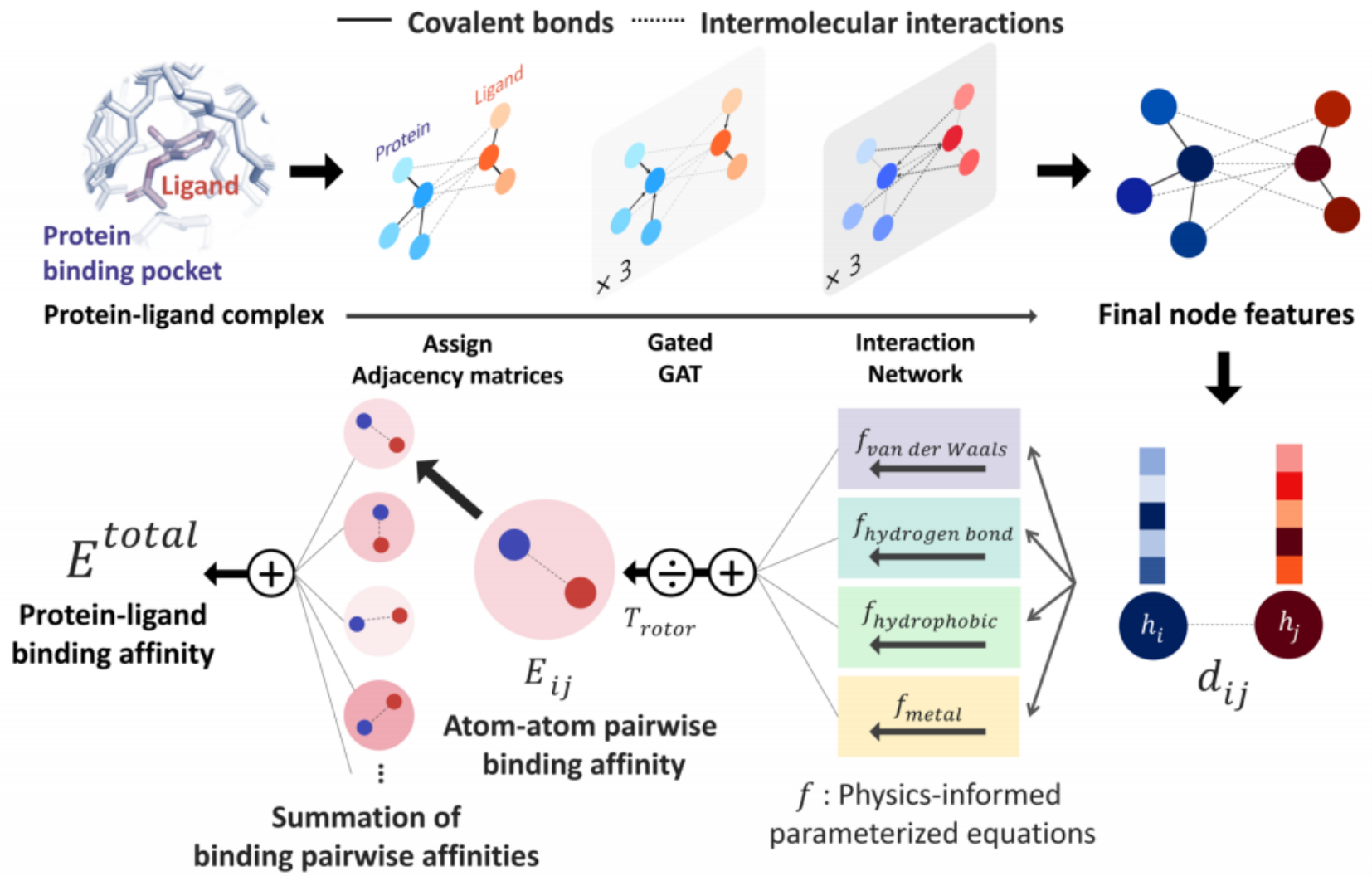
- Hydrogen bond, metal interaction, and hydrophobic interaction

$$e_{ij} = \begin{cases} w & \text{if} \quad d_{ij} - d'_{ij} < c_1 \\ w\left(\frac{d_{ij} - d'_{ij} - c_2}{c_1 - c_2}\right) & \text{if} \quad c_1 < d_{ij} - d'_{ij} < c_2 \\ 0 & \text{if} \quad d_{ij} - d'_{ij} > c_2 \end{cases}$$

$$E = \sum_{i,j} e_{ij}$$

$$T^{rotor} = 1 + C_{rotor} \times N_{rotor}$$

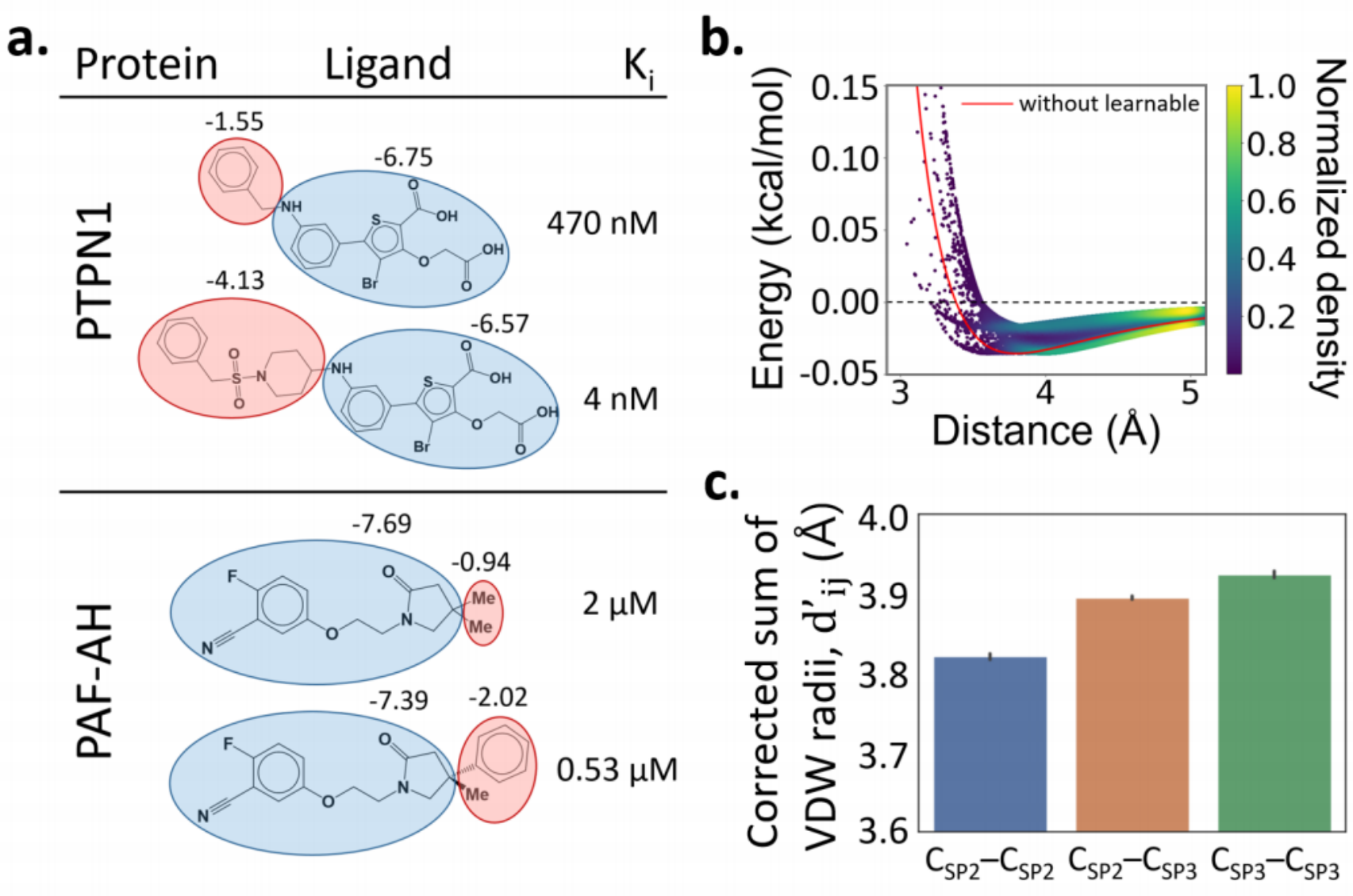
Physics informed DTI prediction



Results

	CASF2016 Benchmark					CSAR	
	Scoring	Ranking	Docking	Screening		NRC-HiQ set1	NRC-HiQ set2
	R	ρ	Success Rate	Average EF	Success Rate	R	R
X-Score[10]	0.631	0.604	63.5%	2.7%	7.0%	0.6	0.65
AutoDock Vina[8]	0.604	0.528	84.6%	7.7%	29.8%	-	-
GlideScore-SP[13]	0.513	0.419	84.6%	11.4%	36.8%	-	-
GlideScore-XP[13]	0.467	0.257	81.8%	8.8%	26.3%	-	-
ChemPLP@GOLD[15]	0.614	0.633	83.2%	11.9%	35.1%	-	-
KDEEP[33]	-	-	-	-	-	0.72	0.65
3D CNN based model	0.652	0.611	42.5%	1.4%	3.5%	0.692	**0.787**
GNN based model	0.723	0.583	67.7%	7.0%	26.3%	0.635	0.786
PIGNet	**0.761**	**0.64**	**85.6%**	**15.1%**	**49.1%**	**0.736**	0.763

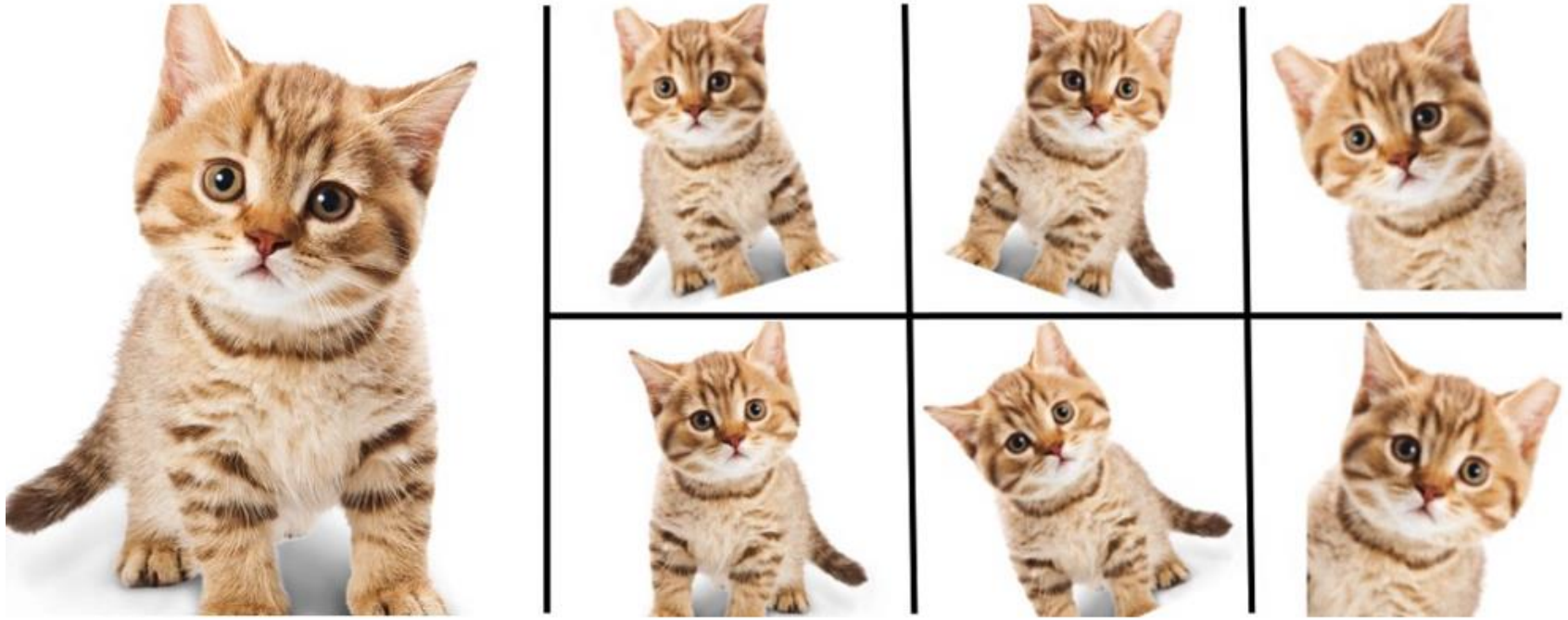
Results


a.
Protein
Ligand
K_i
PTPN1
-1.55
-6.75
470 nM
-4.13
-6.57
4 nM
PAF-AH
-7.69
-0.94
2 μM
-7.39
-2.02
0.53 μM
b.
Energy (kcal/mol)
0.15
0.10
0.05
0.00
-0.05
without learnable
3
4
5
Distance (Å)
1.0
0.8
0.6
0.4
0.2
Normalized density
c.
Corrected sum of VDW radii, d'ij (Å)
4.0
3.9
3.8
3.7
3.6
C_{SP2}–C_{SP2}
C_{SP2}–C_{SP3}
C_{SP3}–C_{SP3}

Data augmentation

- Data augmentation 이란?

  데이터가 부족한 경우, 데이터를 변형, 가공, 생성하여 데이터의 양을 늘리는 기술



Data augmentation in DTI prediction

- Docking power

  True binding pose를 실험으로 측정하는 것은 어려움. 그러나 하나의 true binding pose가 있으면, 무수히 많은 false binding pose를 생산할 수 있음

- Screening power

  대부분의 compound들은 타겟 단백질에 대해서 inactive함. Active한 compound를 확보하는 것은 어려우나, inactive한 compound는 쉽게 생산할 수 있음 (negative set에 대한 labeling 오류를 동반하게 됨)

- Global and local minimum

  X-ray구조들은 가장 안정한 구조. 따라서 potential surface상에서 global minimum이자 local minimum. 이러한 constraint를 적용하여 model의 overfitting risk를 줄일 수 있음

Data augmentation in DTI prediction

- Docking augmentation

$$L_{docking} = \sum_i \max\left(y_{exp,i} - y_{decoy,i},\ 0\right)$$

- Screening augmentation

$$L_{random\_screening} = \sum_i \max\left(-y_{random,i} - 6.8, 0\right)$$

$$L_{cross\_screening} = \sum_i \max\left(-y_{cross,i} - 6.8, 0\right)$$

Data augmentation in DTI prediction

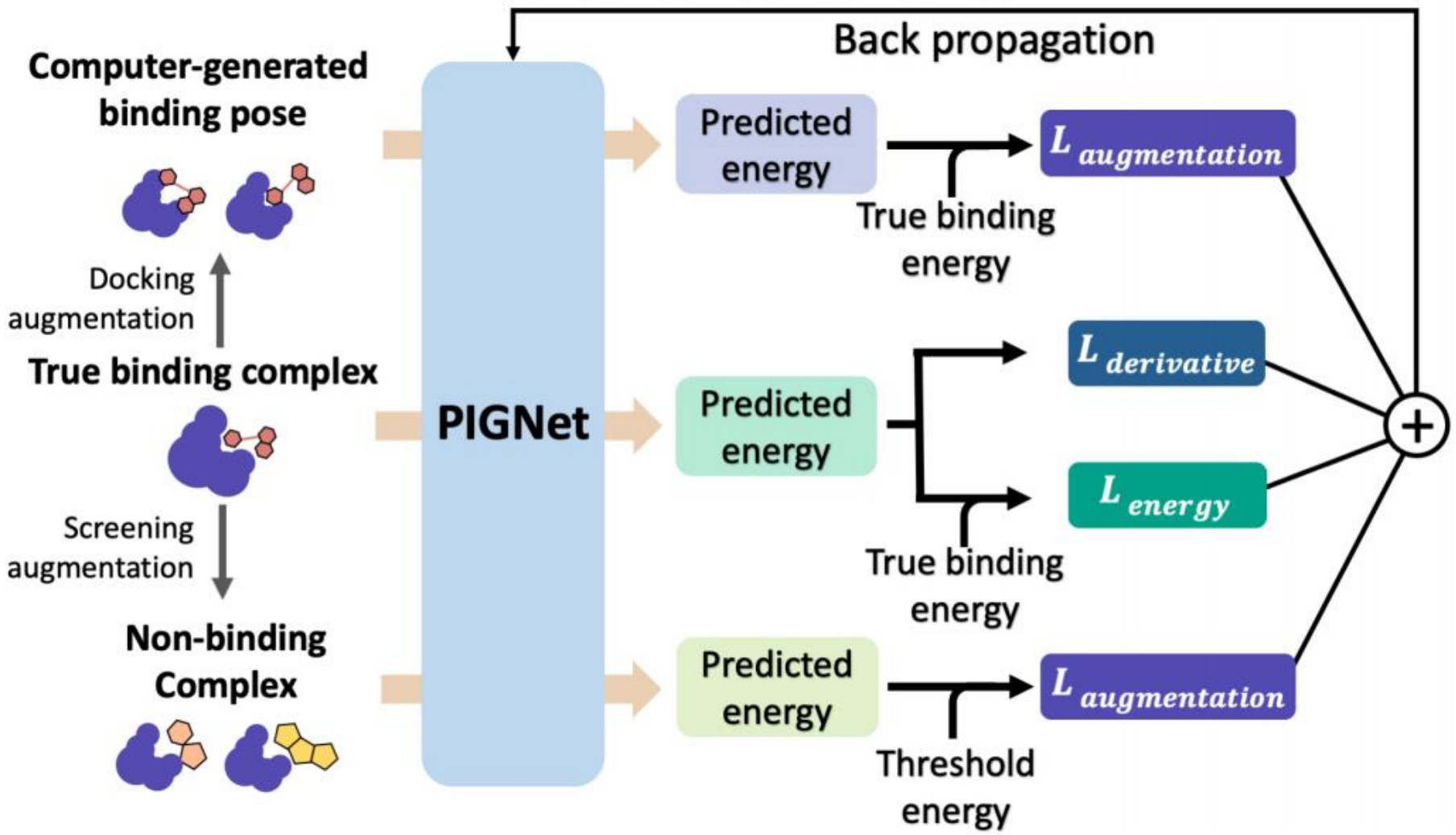
- Global and local minimum constraint

$$L_{derivative} = \sum_i \left(\frac{\partial E^{total}}{\partial q_i} \right)^2 - min\left(\left(\frac{\partial^2 E^{total}}{\partial q_i^2} \right), C_{der2} \right)$$

- Total loss function

$$\begin{aligned} L_{total} = & L_{energy} \\ & + c_{derivative} L_{derivative} \\ & + c_{docking} L_{docking} \\ & + c_{random\_screening} L_{random\_screening} \\ & + c_{cross\_screening} L_{cross\_screening}, \end{aligned}$$

Data augmentation in DTI prediction

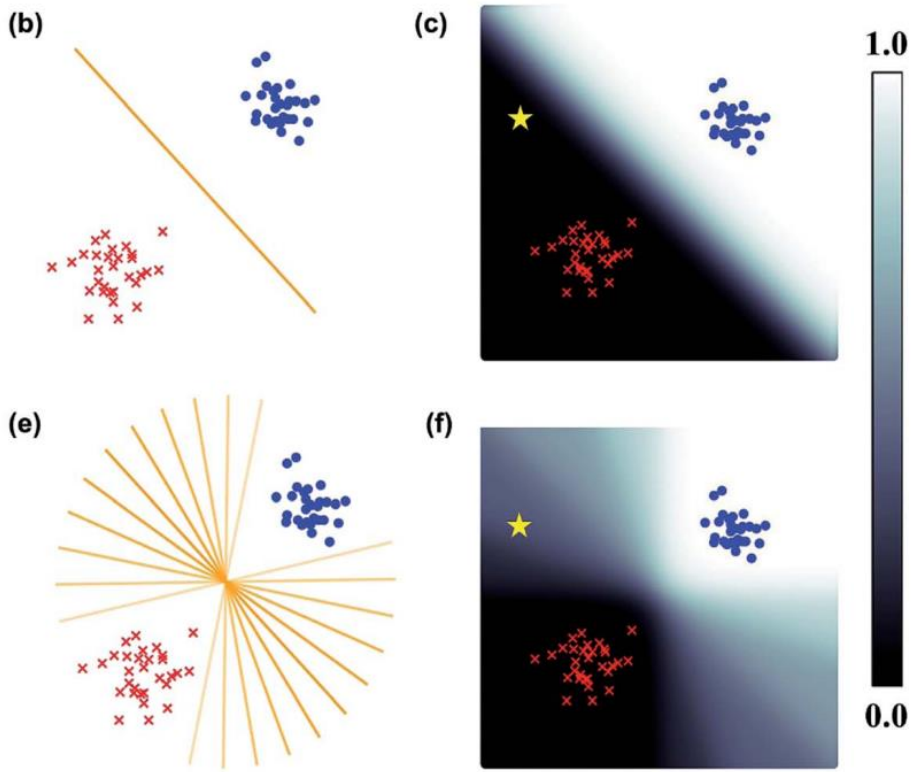


Results

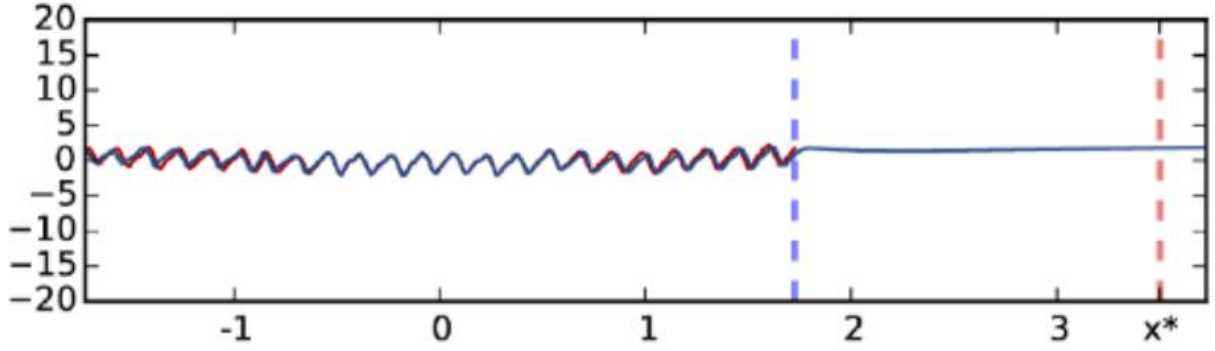
	CASF2016 Benchmark					CSAR	
	Scoring	Ranking	Docking	Screening		NRC-HiQ set1	NRC-HiQ set2
	R	ρ	Success Rate	Average EF	Success Rate	R	R
3D CNN based model W/O data augmentation	**0.695**	0.589	20.4%	0.7%	1.8%	**0.786**	**0.785**
3D CNN based model with data augmentation	0.652	**0.611**	**42.5%**	**1.4%**	**3.5%**	0.692	0.787
GNN based model W/O data augmentation	**0.773**	**0.617**	28.1%	1.4%	5.3%	**0.792**	**0.787**
GNN based model with data augmentation	0.723	0.583	**67.7%**	**7.0%**	**26.3%**	0.635	0.786
PIGNet W/O data augmentation	0.703	0.606	77.9%	6.0%	26.3%	0.72	**0.789**
PIGNet with data augmentation	**0.761**	**0.64**	**85.6%**	**15.1%**	**49.1%**	**0.736**	0.763

Uncertainty quantification

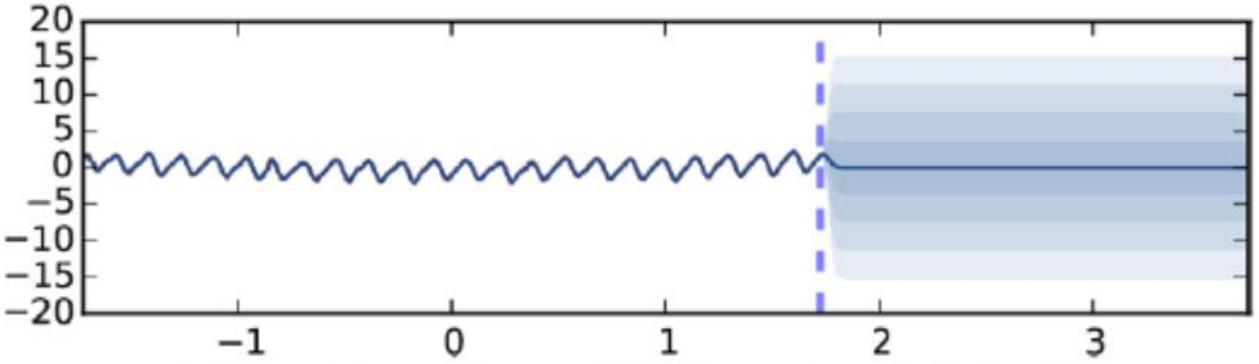
- Deep learning model은 항상 어떤 값을 제공해줌. 과연 이 값을 항상 신뢰할 수 있는가?
- Prediction의 신뢰도 범위를 정량적으로 측정할 수 있다면 모델의 결과를 받아드릴지 말지 결정하는데 있어 큰 도움이 됨. ex) 이 그림이 강아지일 확률이 90±5 %이다.



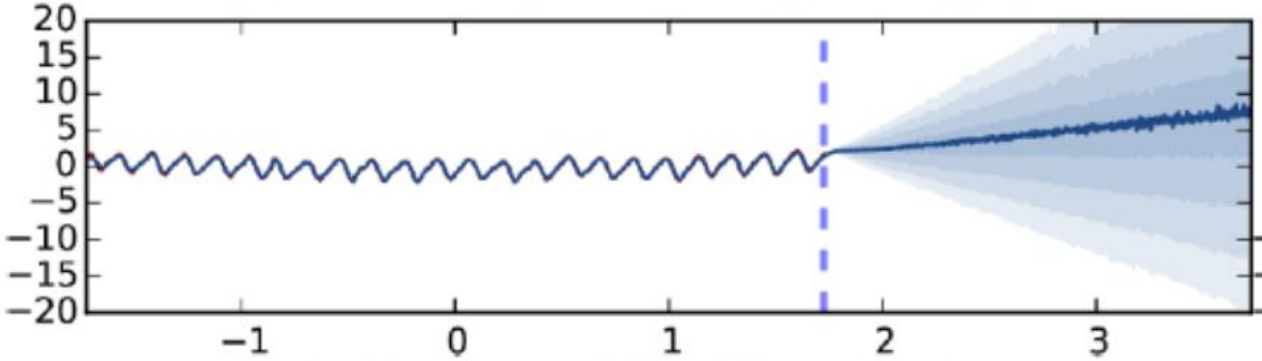
Uncertainty quantification



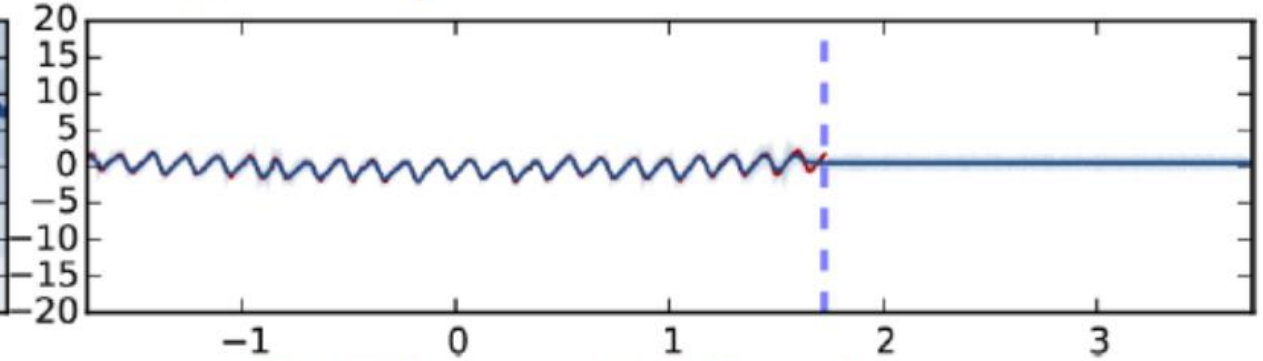
(a) Standard dropout with weight averaging



(b) Gaussian process with SE covariance function

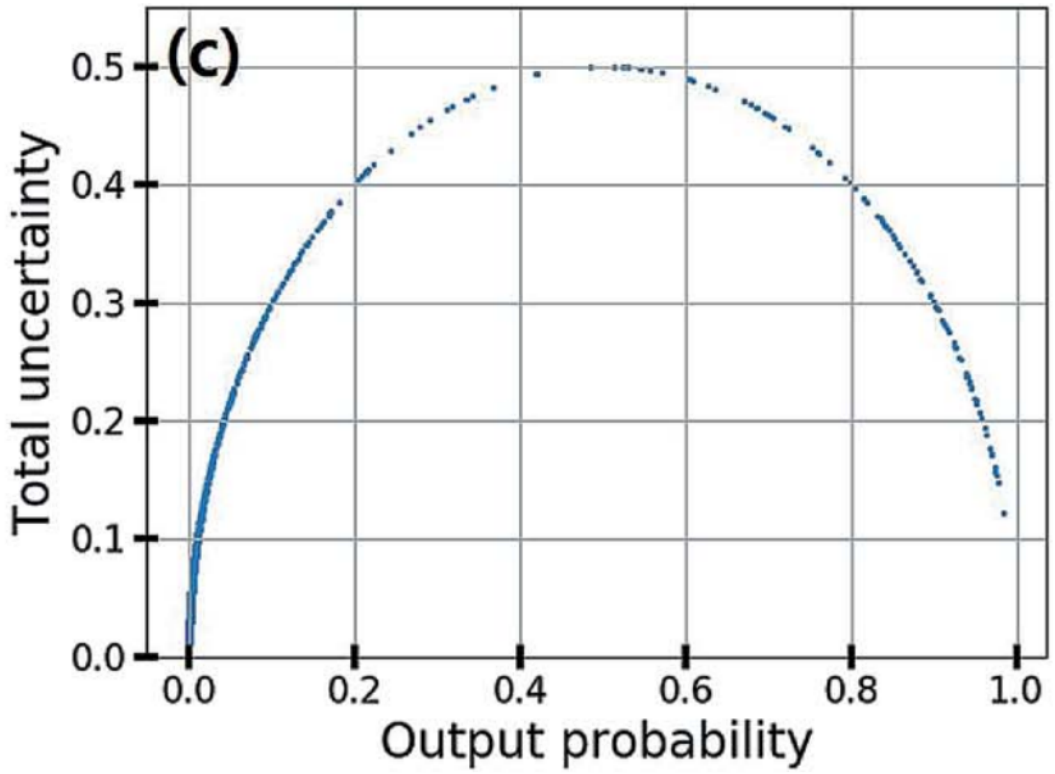


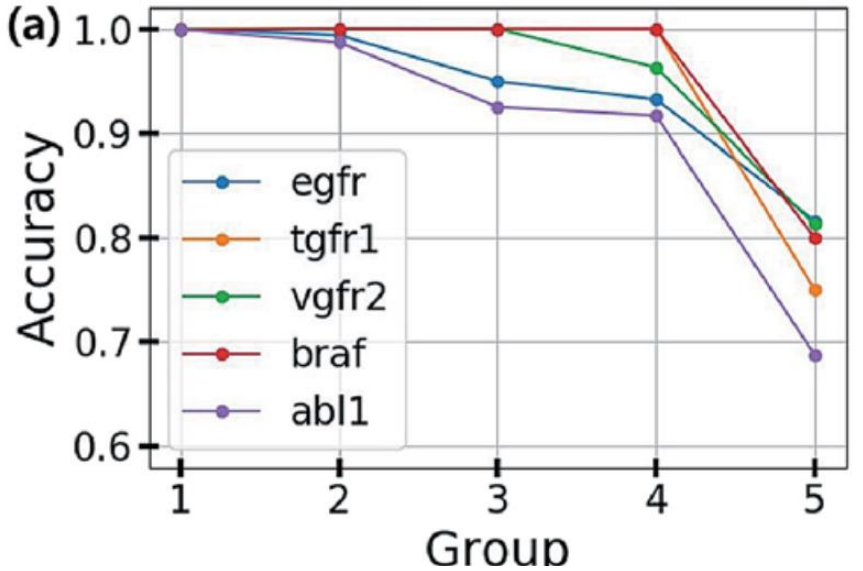
(c) MC dropout with ReLU non-linearities

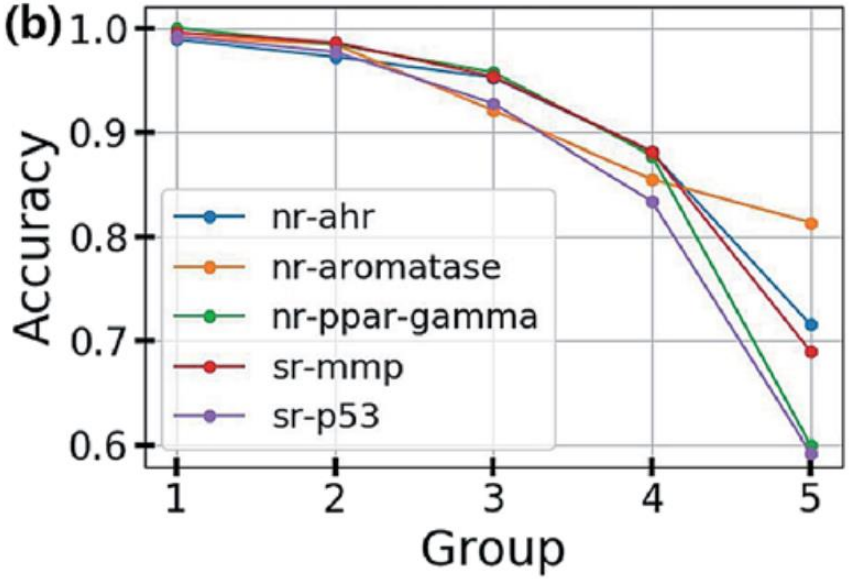


(d) MC dropout with TanH non-linearities

Uncertainty quantification







Chem. Sci., 2019, 10, 8438

Uncertainty quantification in DTI prediction

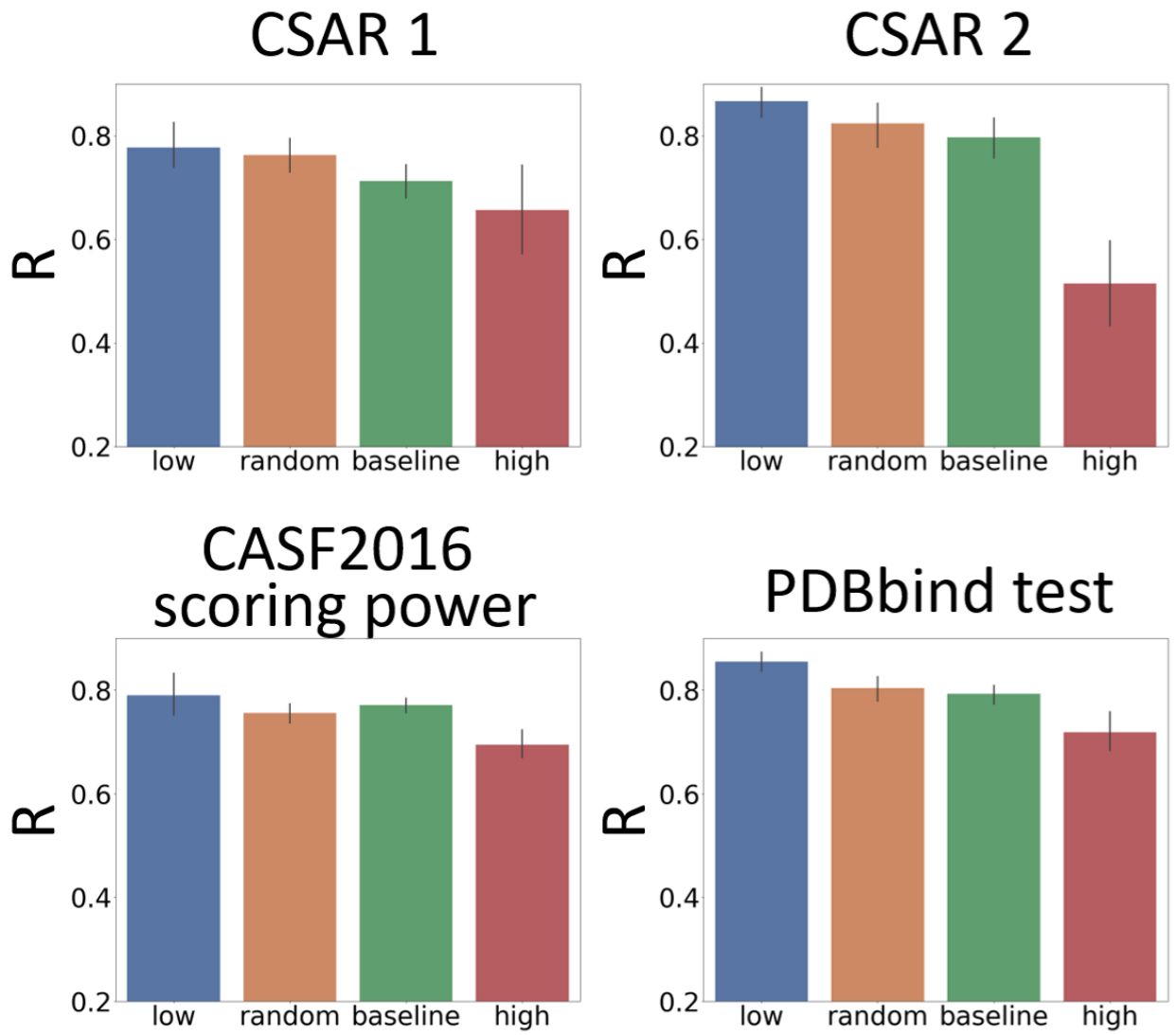
- Aleatoric uncertainty: data에 포함된 intrinsic noise

$$L_{aleatoric}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2\sigma(x_i)^2}\|y_i - f(x_i)\|^2 + \frac{1}{2}\log\sigma(x_i)^2.$$

- Epistemic uncertainty: model parameter에 내제된 uncertainty (MC dropout)
- atom-atom pair uncertainty and distance dependency uncertainty

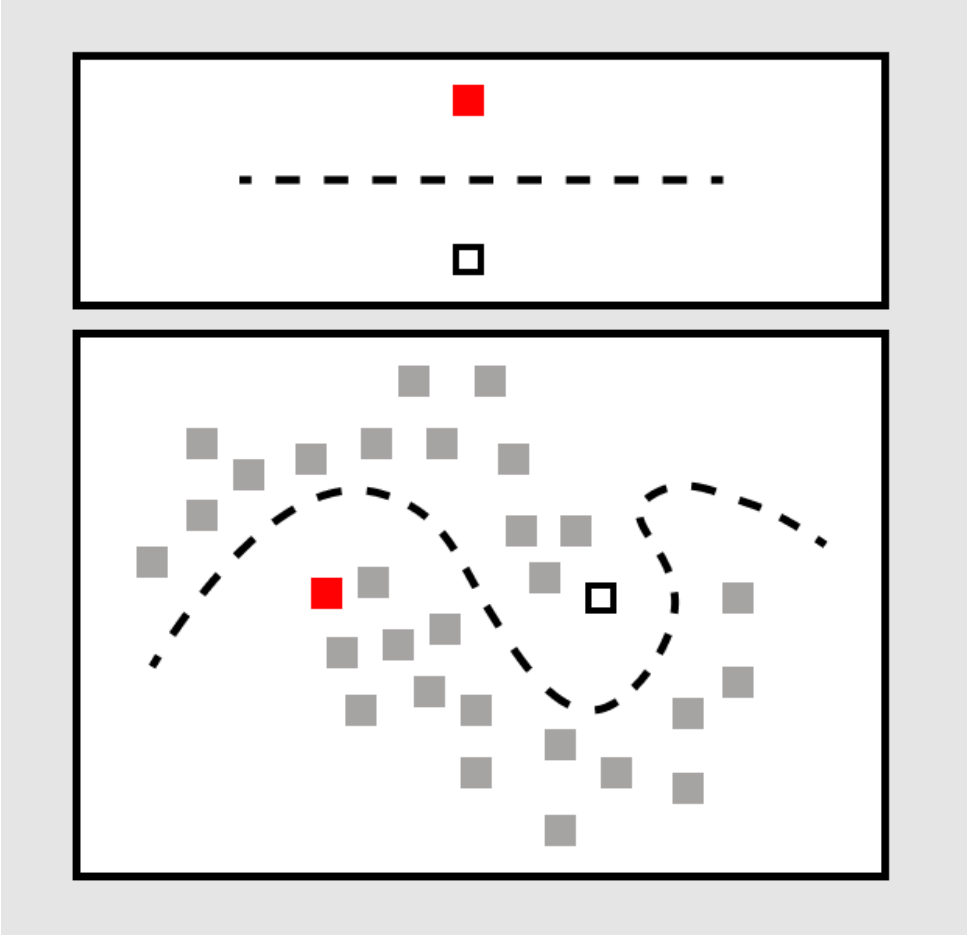
$$\sigma^2 = \prod_{i,j}\sigma_{ij}^2 = \prod_{i,j}|W_2^{var}(ReLU(W_1^{var}(h_{ij}^{concat}))) \times a\exp(-bd_{ij})|,$$

Uncertainty quantification



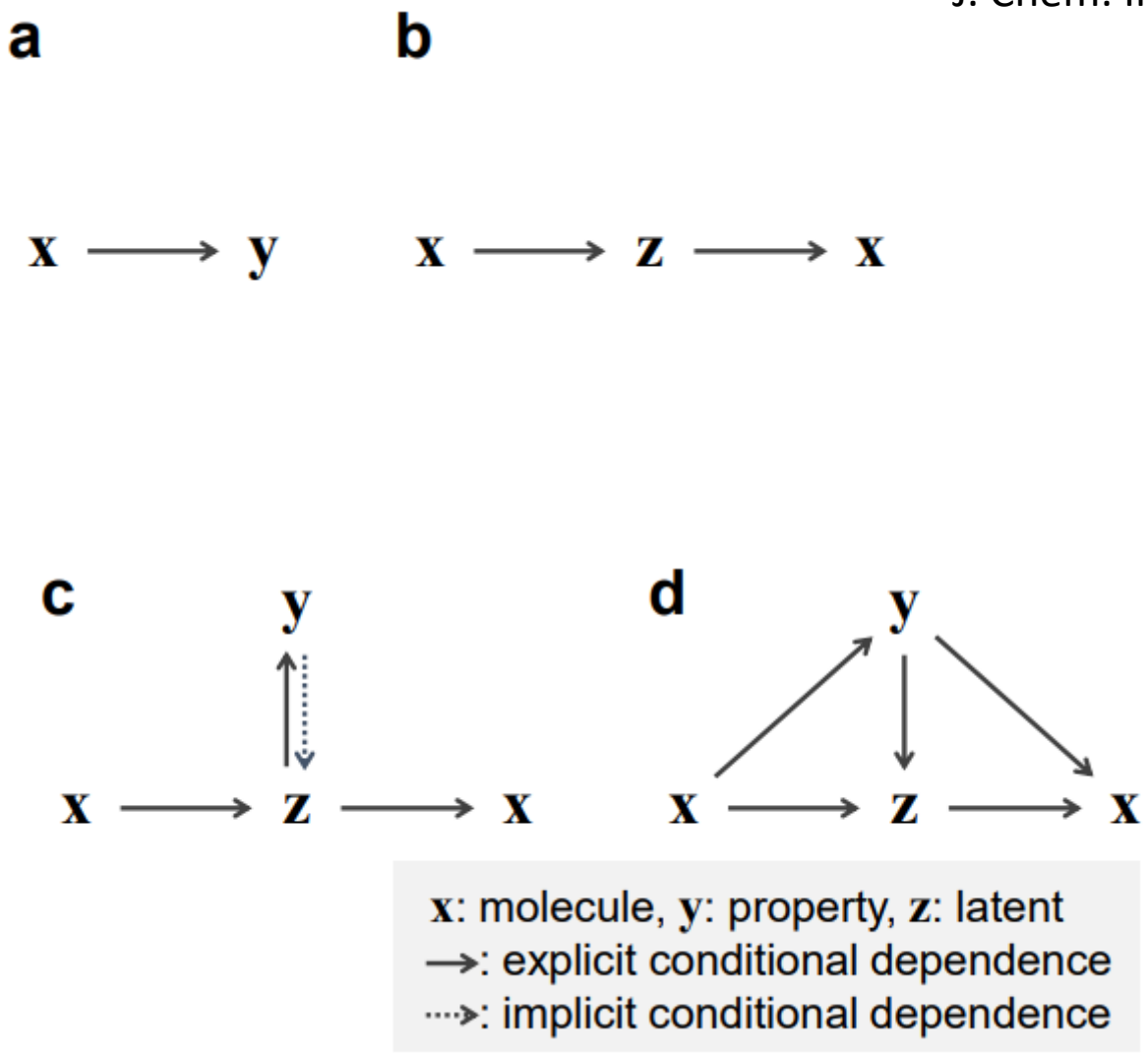
Semi-supervised learning

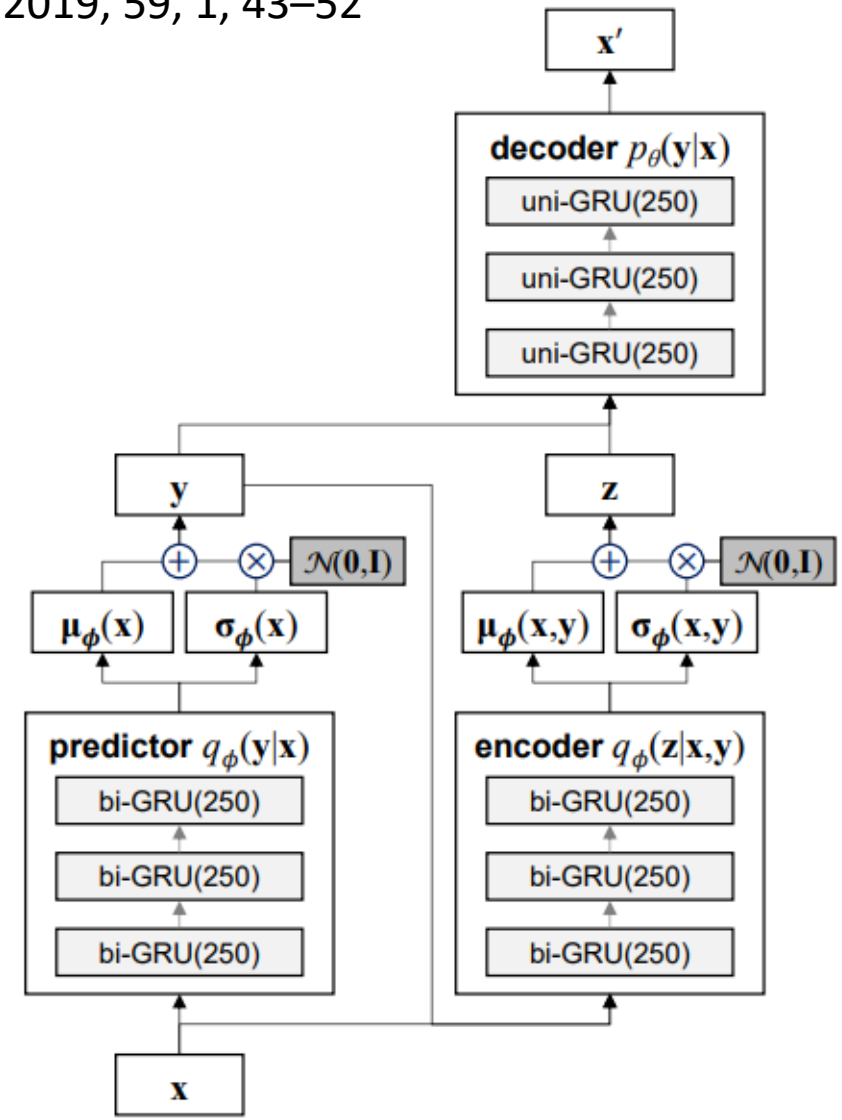
- 소수의 labeled 데이터와 다수의 unlabeled 데이터를 동시에 사용하는 방법



Semi-supervised learning

J. Chem. Inf. Model. 2019, 59, 1, 43–52





Results

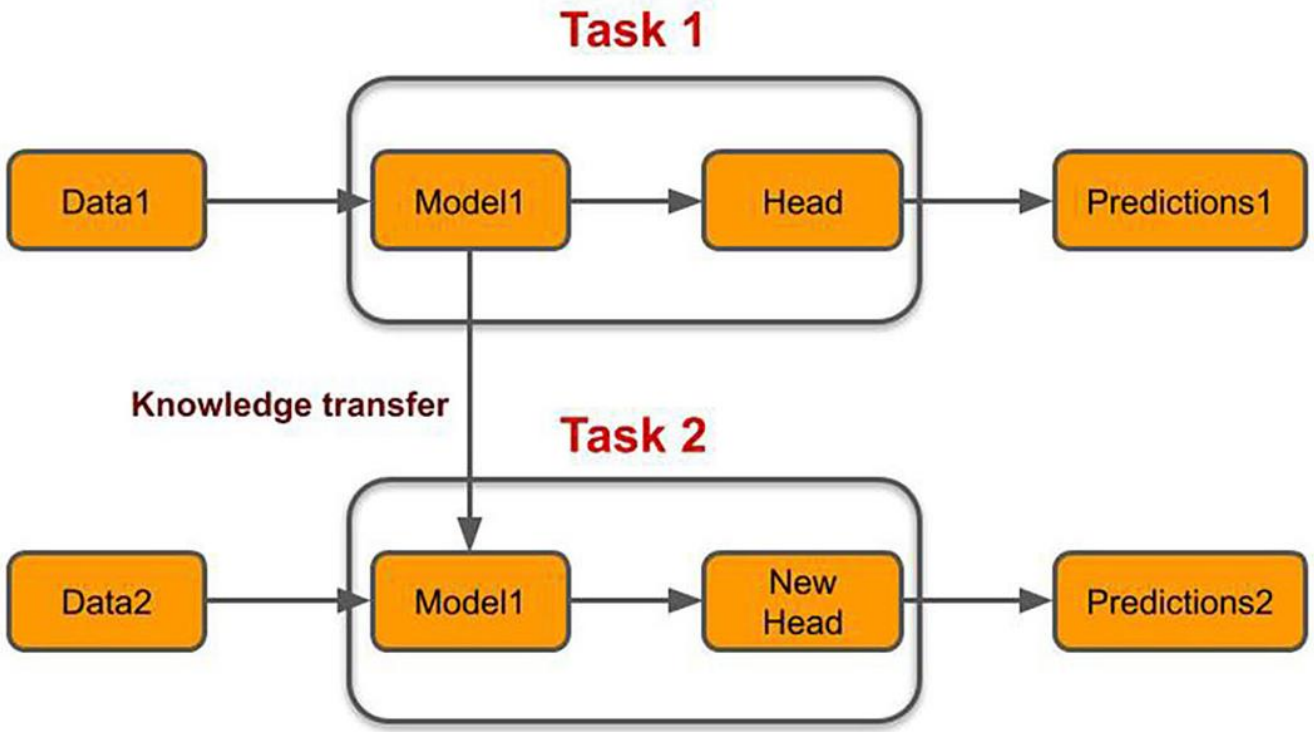
frac. labeled	property	ECFP	GraphConv	predictor network	$VAE_{property}$	SSVAE
5%	MolWt	17.713±0.396	6.723±2.116	2.582±0.288	3.463±0.971	1.639±0.577
	LogP	0.380±0.009	0.187±0.015	0.162±0.006	0.125±0.013	0.120±0.006
	QED	0.053±0.001	0.034±0.004	0.037±0.002	0.029±0.002	0.028±0.001
10%	MolWt	15.057±0.358	5.255±0.767	1.986±0.470	2.464±0.581	1.444±0.618
	LogP	0.335±0.005	0.148±0.016	0.116±0.006	0.097±0.008	0.090±0.004
	QED	0.045±0.001	0.028±0.003	0.027±0.002	0.021±0.002	0.021±0.001
20%	MolWt	12.047±0.168	4.597±0.419	1.228±0.229	1.748±0.266	1.008±0.370
	LogP	0.249±0.004	0.112±0.015	0.070±0.007	0.074±0.006	0.071±0.007
	QED	0.033±0.001	0.021±0.002	0.017±0.002	0.015±0.001	0.016±0.001
50%	MolWt	9.012±0.184	4.506±0.279	1.010±0.250	1.350±0.319	1.050±0.164
	LogP	0.180±0.003	0.086±0.012	0.045±0.005	0.049±0.008	0.047±0.003
	QED	0.023±0.000	0.018±0.001	0.011±0.001	0.009±0.002	0.010±0.001

J. Chem. Inf. Model. 2019, 59, 1, 43–52

Transfer learning

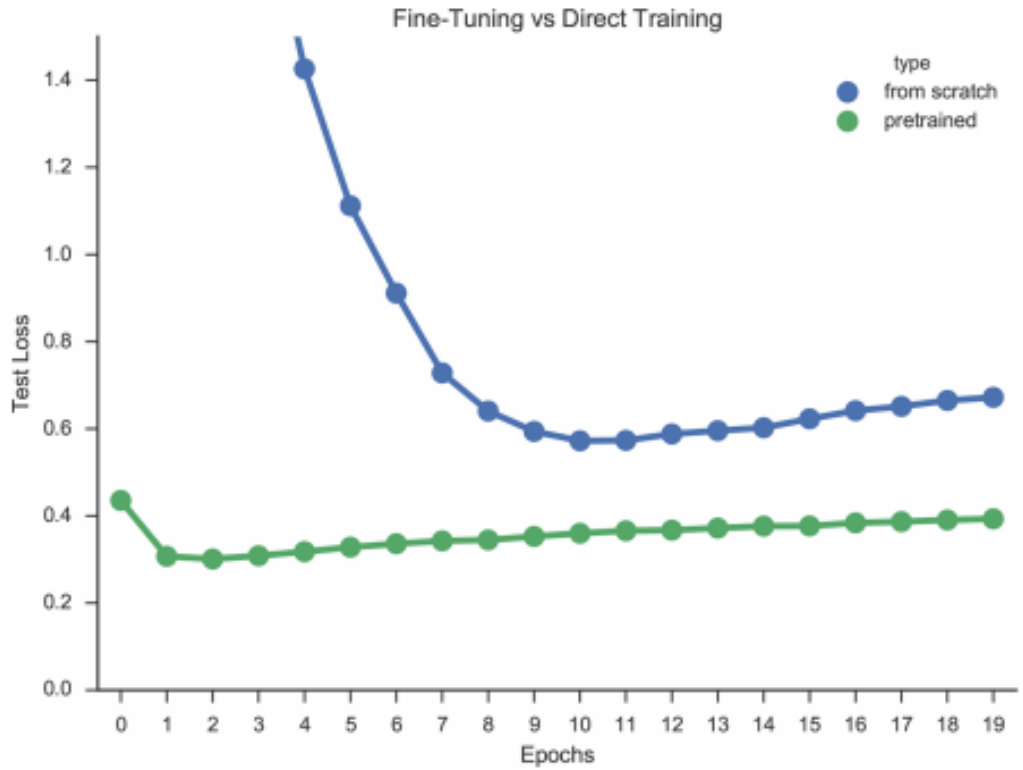
- Transfer learning이란?

  어떤 학습과정에서 얻어진 knowledge를 다른 유사 학습 문제에 적용하는 방법. 이를 통해 적은 데이터를 이용하지만 빠른 속도로 학습 가능하며, overfitting의 위험성을 줄일 수 있다.

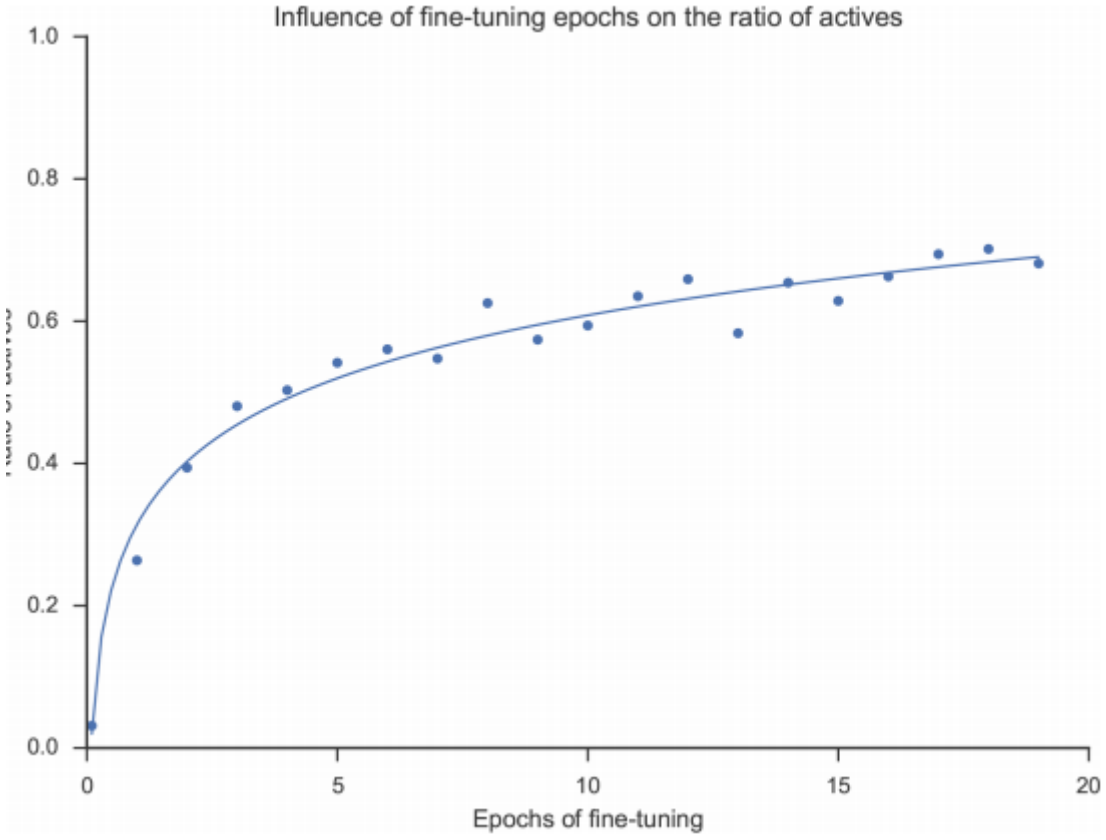


Transfer learning

- 대규모 데이터 (Zinc, pubchem 등 100만개 이상)을 이용해서 생성모델을 1차로 학습시킴
- 학습된 모델을 초기 모델로 해서 소규모 데이터셋 (5-HT_{2A} active compounds, 732 molecules)
- Target prediction model (TPM))을 이용해서 생성된 분자의 활성 검증



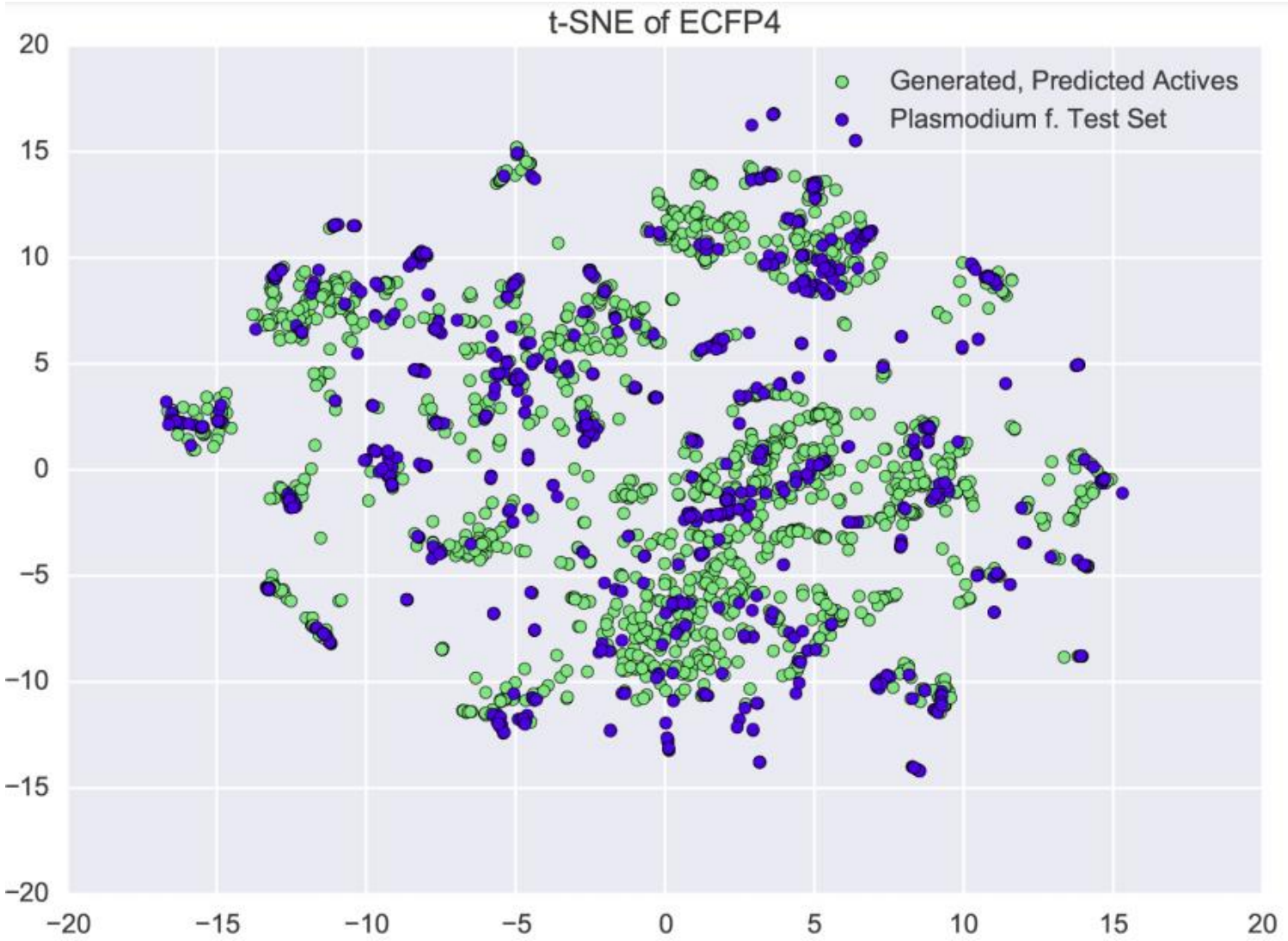
Transfer learning



no.	pIC_{50}	training	test	gen mols	reprod (%)	EOR[a]
1	>8	1239	1240	128,256	28	66.9
2	>8	100	1240	93,721	7	19.0
3	>9	100	1022	91,034	11	35.7

[a]EOR: Enrichment over random.

Transfer learning



Thank you