Lecture05. Generalization issue of deep learning methods in drug discovery

HITS 임 재 창

목차

- Generalization ability
- Performance of DTI model across different datasets
- Performance of DTI model depending on different metrics
- Problem of DTI model in terms of learning intrinsic bias of dataset
- Molecules with high similarity from molecular generative model
- Origin of low generalization ability

Generalization 이란?

- 일반화 (generalization) 란?

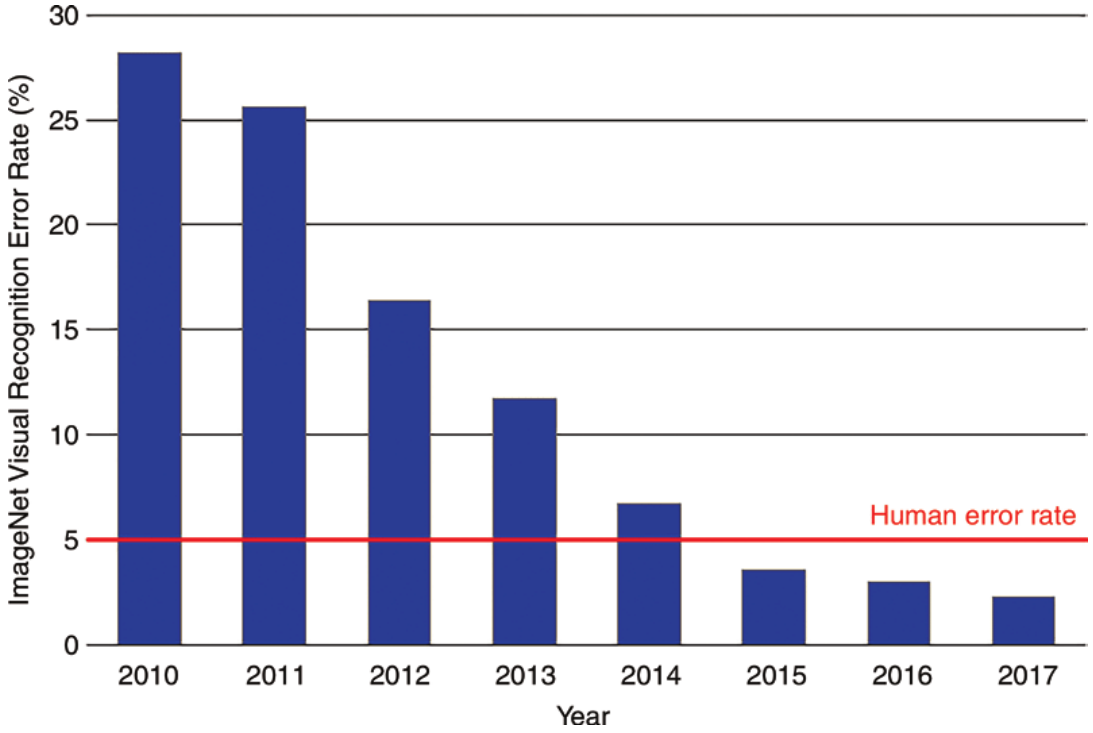
  Training set뿐 아니라 다양한 범위의 input과 application들에 대해서도 성능을 유지

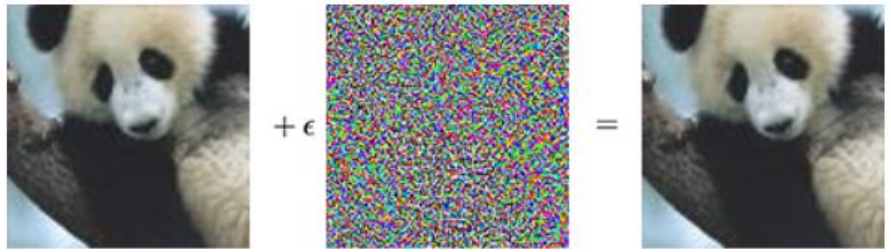
AUC		> 0.5	> 0.6	> 0.7	> 0.8	> 0.9
ChEMBL-20 PMD	AtomNet	49	44	36	24	10
	Smina	38	10	4	1	0
DUDE-30	AtomNet	30	29	27	22	14
	Smina	29	25	14	5	1
DUDE-102	AtomNet	102	101	99	88	59
	Smina	96	84	53	17	1
ChEMBL-20 inactives	AtomNet	149	136	105	45	10
	Smina	129	81	31	4	0

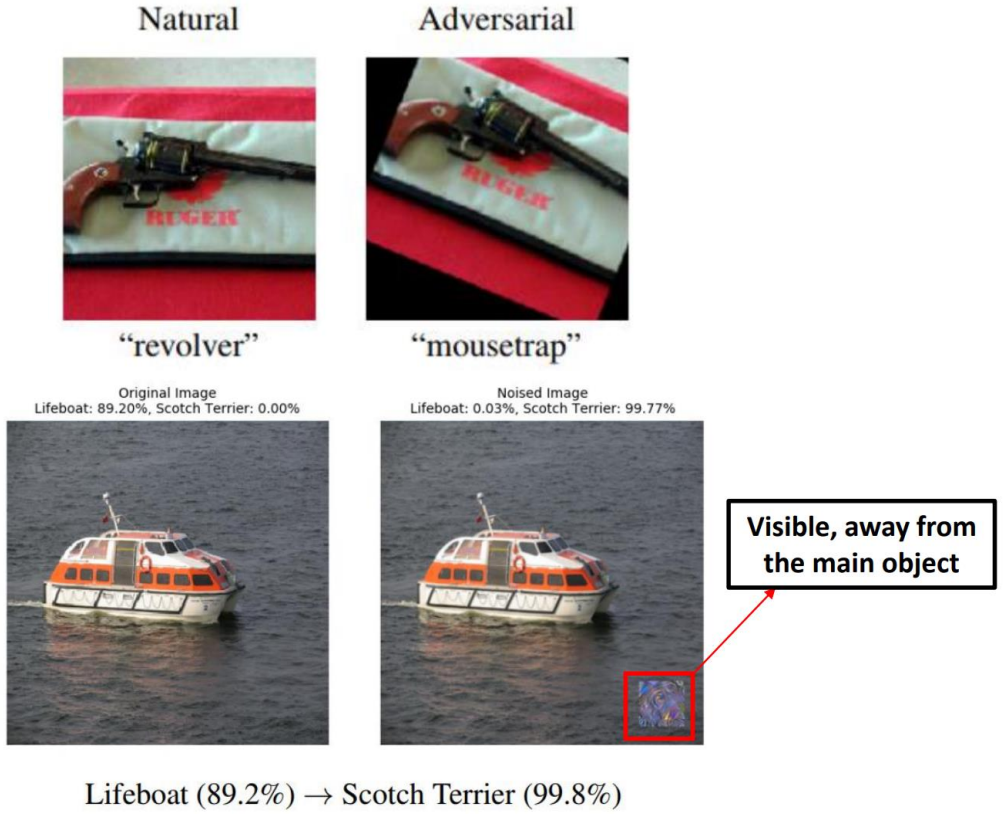
- Test set에서도 좋은 성능을 보여줌. 이 모델은 generalization을 달성한 것인가?
  - ✓진정한 generalization을 달성했다면, 유사 데이터 셋에서도 좋은 성능을 보여줘야함
  - ✓또한 유사 task들에 대해서도 성능이 유지 되어야함
  - ✓기존까지는 이러한 것들에 대한 discussion이 부족했음

Issue of generalization in image classification

ImageNet challenge







Performance of DTI model across different datasets

- Deep DTI 모델들이 generalization을 달성했다면, 특정 dataset뿐 아니라 다양한 dataset에 대해서 일관된 성능을 보여야 한다.

	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					

	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

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

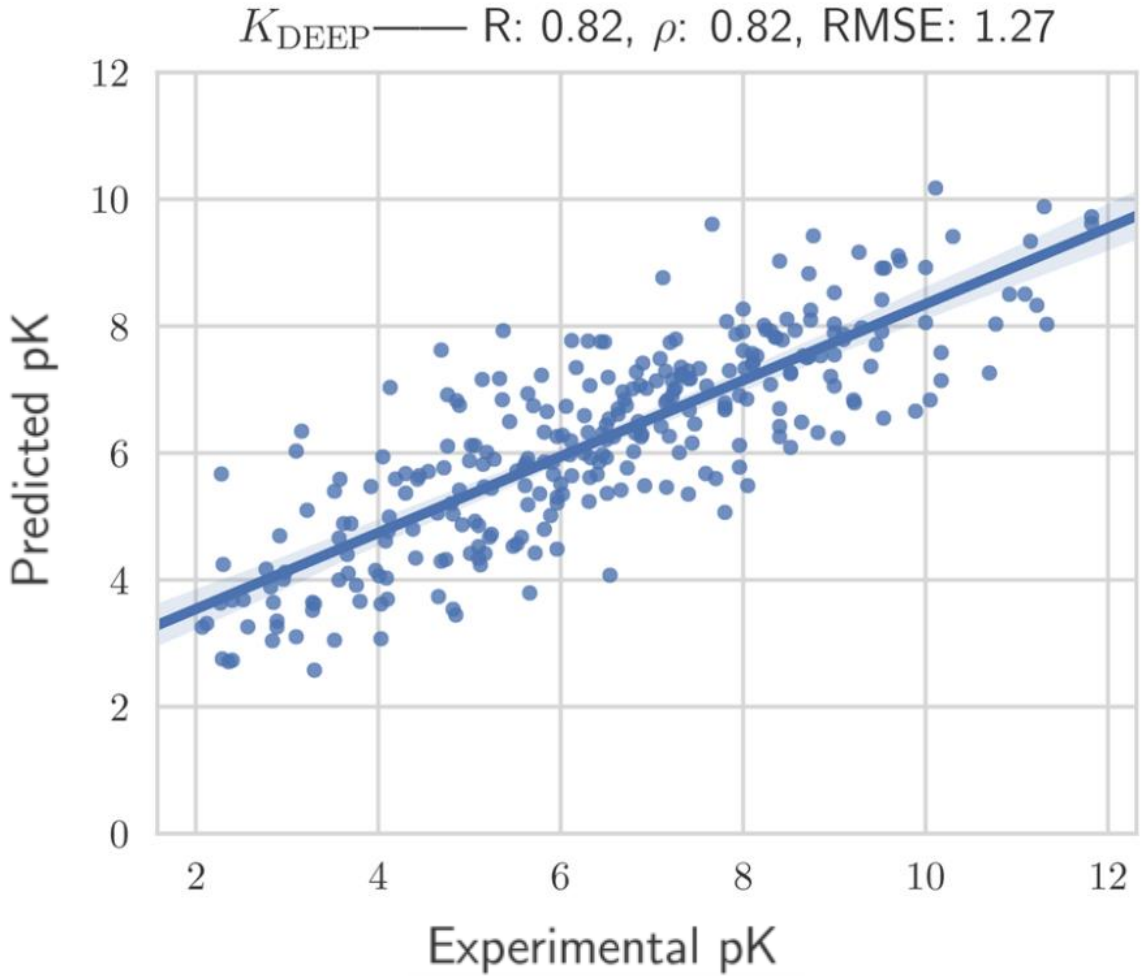
Performance of DTI model across different datasets

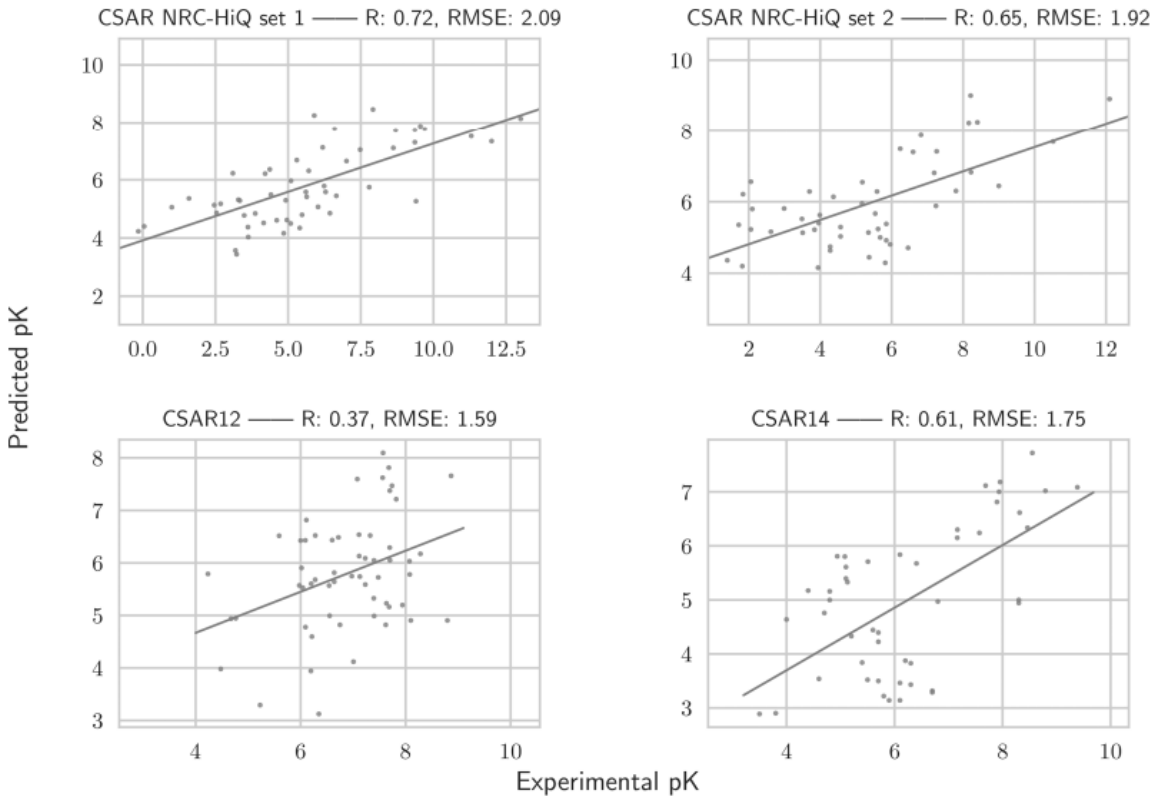
- DUD-E set에서 높게 유지되었던 성능이 ChEMBL과 MUV set에 대해서는 큰 폭으로 감소함.
- Docking도 줄어들기는 하지만, Deep DTI와 docking의 차이가 큰 폭으로 줄어듬

	ChEMBL			
	AUROC	sensitivity	specificity	balanced accuracy
ours	**0.633**	0.813	0.325	0.569
docking	0.572			
	MUV			
	AUROC	sensitivity	specificity	balanced accuracy
ours	0.536	0.286	0.752	0.519
docking	0.533			
Ragoza et al.[22]	0.518			
Torng et al.[40]	**0.563**			

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

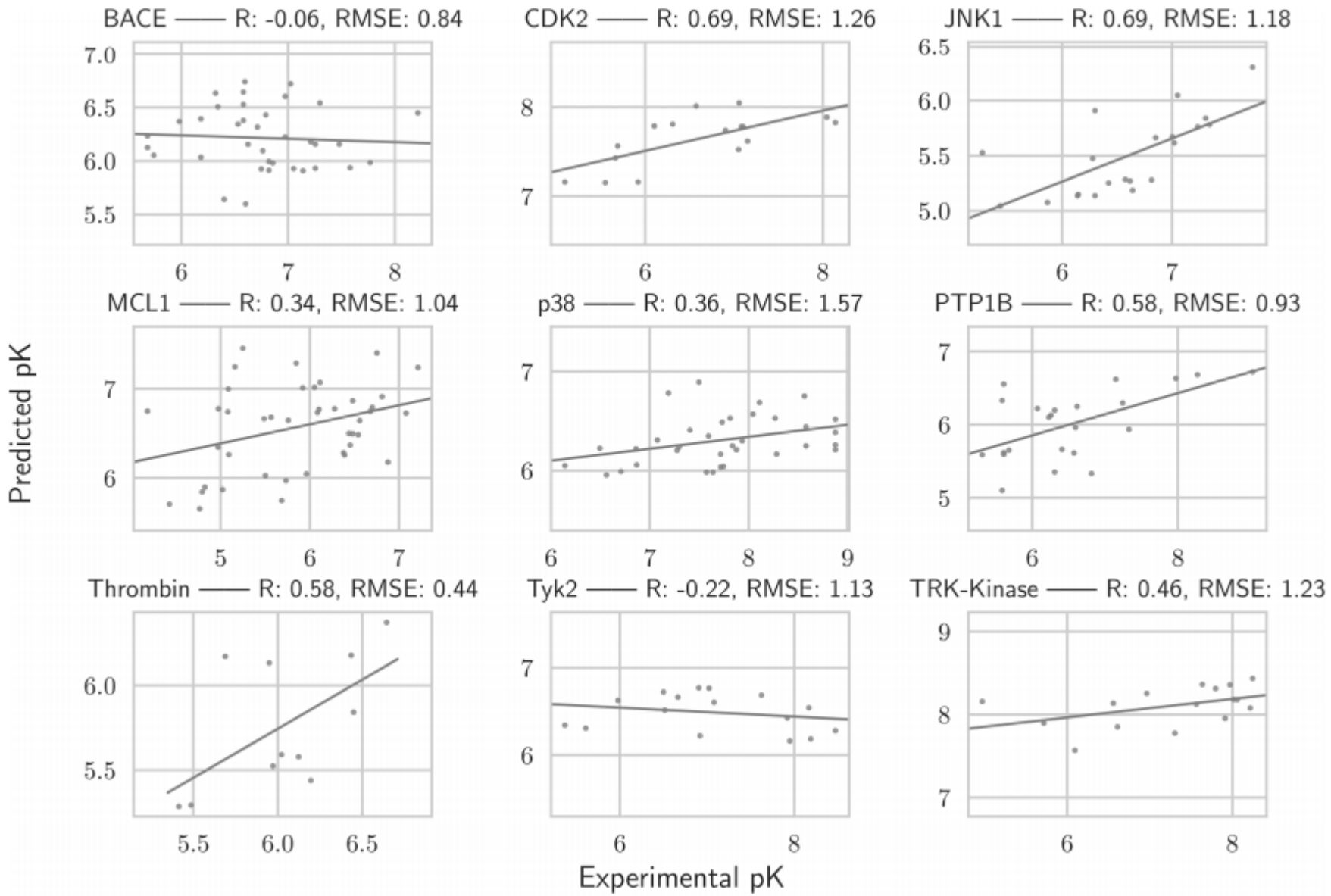
Performance of DTI model across different datasets





J. Chem. Inf. Model. 2018, 58, 2, 287–296

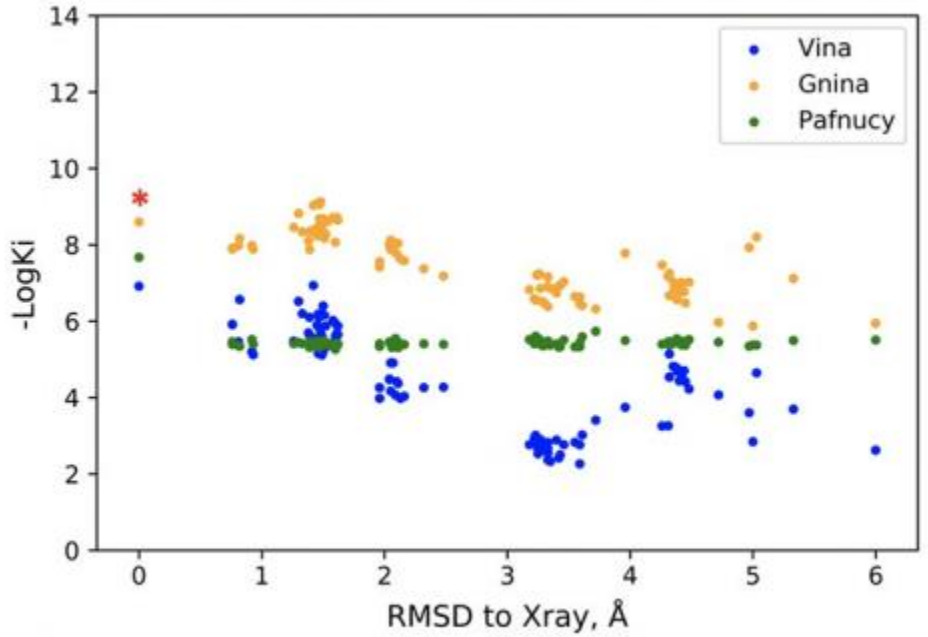
Performance of DTI model across different datasets



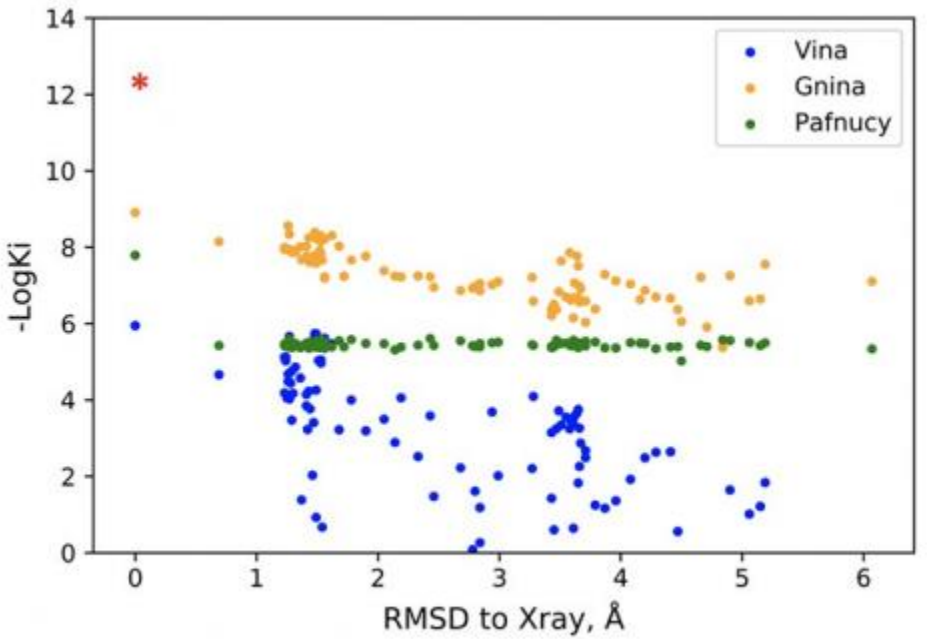
J. Chem. Inf. Model. 2018, 58, 2, 287–296

Performance of DTI model across different datasets

	Average AUC	Frequency (AUC>0.8)	Frequency (AUC>0.9)
Vina	0.725	24	3
Gnina	0.709	28	10
Pafnucy	0.632	12	0



(A) XLC

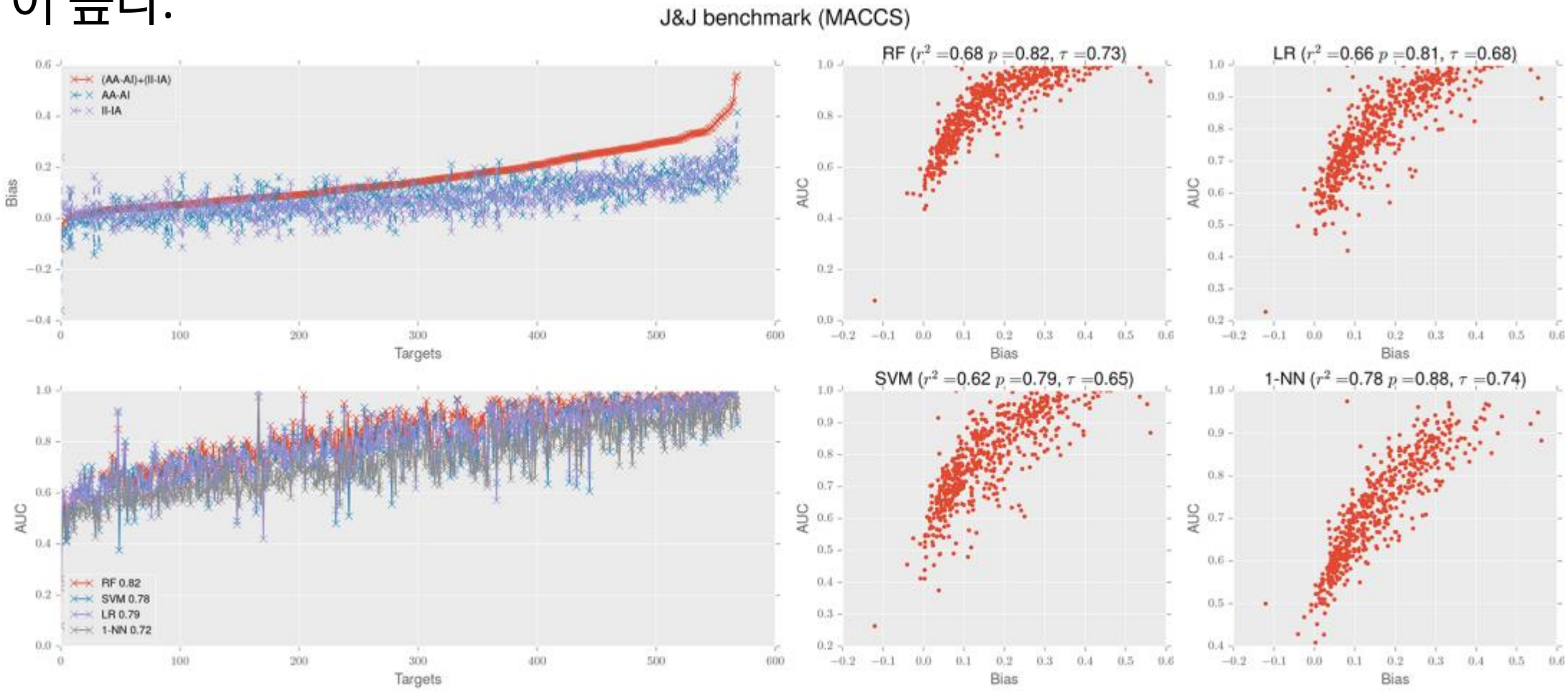


(B) XLD

https://doi.org/10.1371/journal.pone.0220113

Performance of DTI model across different datasets

- Machine learning에서 좋게 측정된 성능들의 상당 부분이 over-fitting일 가능성이 높다.



J. Chem. Inf. Model. 2018, 58, 5, 916–932

Performance of DTI model depending on different metrics

- 대부분의 deep learning 방법들이 scoring만을 기준으로 모델의 성능을 평가함.
- 하지만 robust한 model을 개발하기 위해서는 보다 여러 측면에서 성능평가가 필요함.

Metric 이름	내용	지표
Scoring power	protein-ligand complex 구조에 대해서 experimental binding affinity 예측 성능	Pearson correlation (R)
Ranking power	같은 cluster에 속한 protein-ligand complex x-ray구조의 experimental binding affinity 순서 예측 성능	Pearson ranking correlation
Docking power	True x-ray binding structure와 False binding structure 구분 정확도	True binding pose 구분 성공률 (%)
Screening power	Virtual screening 성능 (가상 library에서 True binder 발견 확률)	Enrichment factor

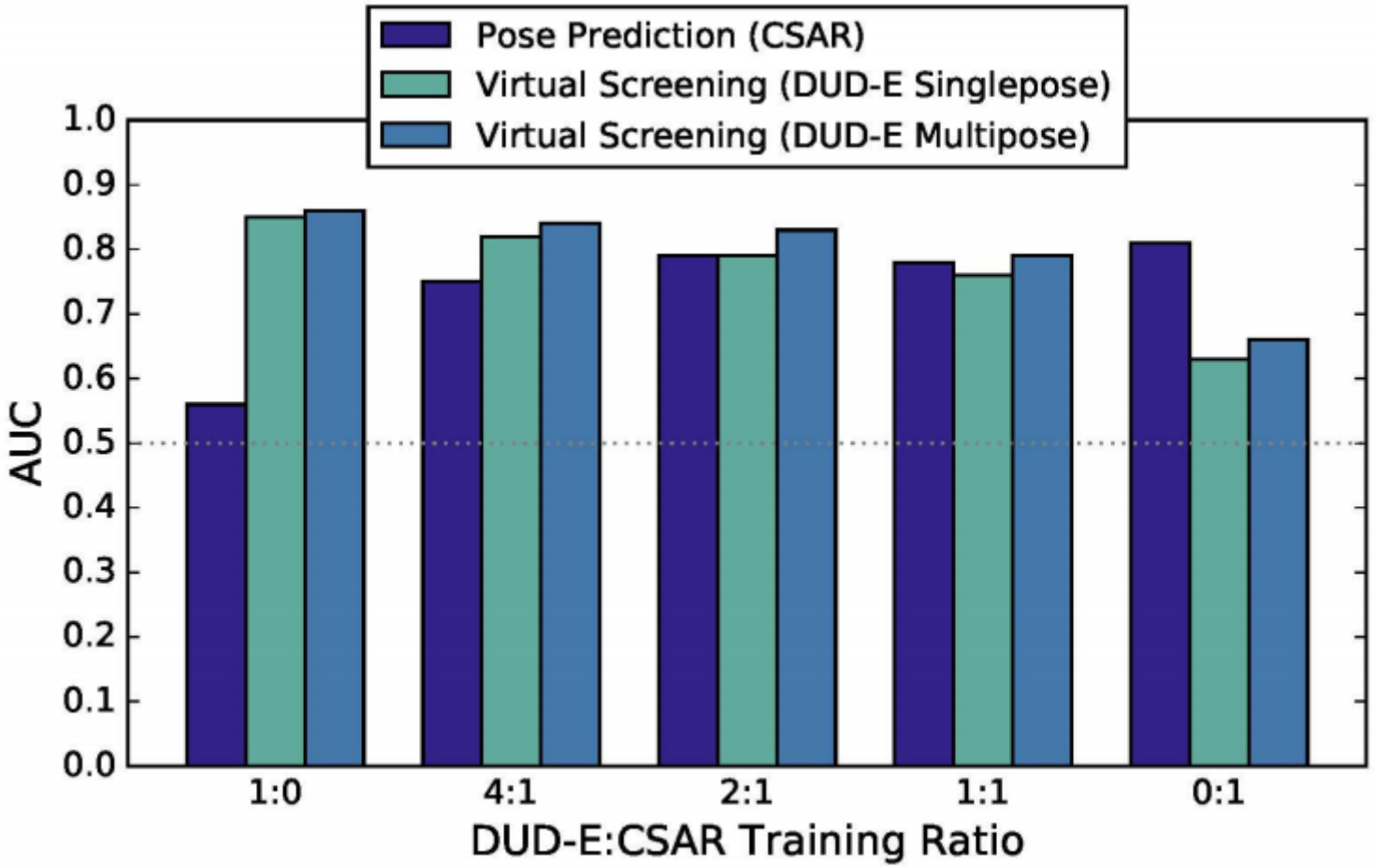
Performance of DTI model depending on different metrics

- Deep DTI model들이 scoring power에 대해서는 좋은 성능을 보여주지만, 다른 지표 특히 docking power에 대해서 낮은 성능을 보여줌

	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

arXiv:2008.12249

Performance of DTI model depending on different metrics

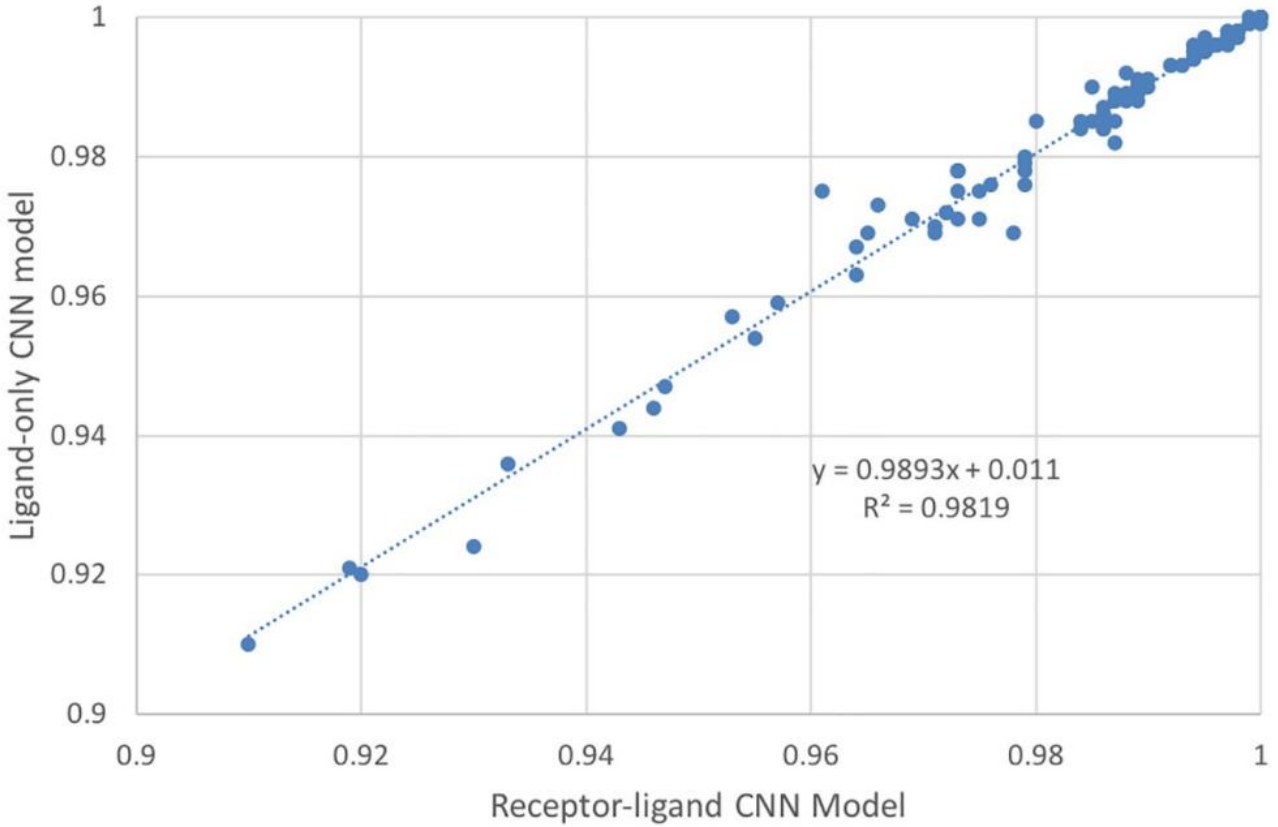


J. Chem. Inf. Model. 2017, 57, 4, 942–957

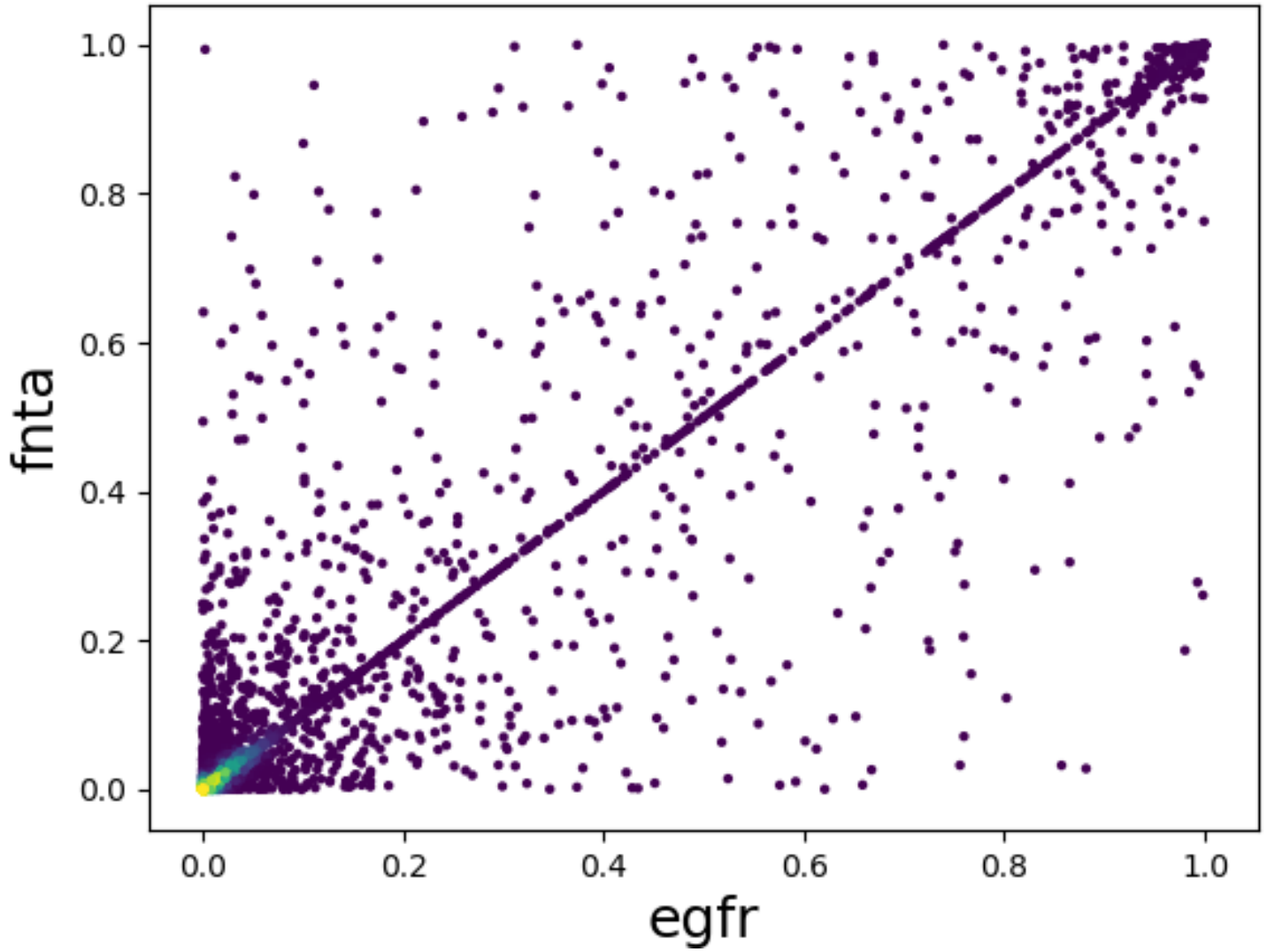
Problem of DTI model in terms of learning intrinsic bias of dataset

- DUD-E set에 대해서 ligand only CNN model과 receptor-ligand CNN model을 학습시켰을 때, 두 모델간의 성능차이가 거의 없음

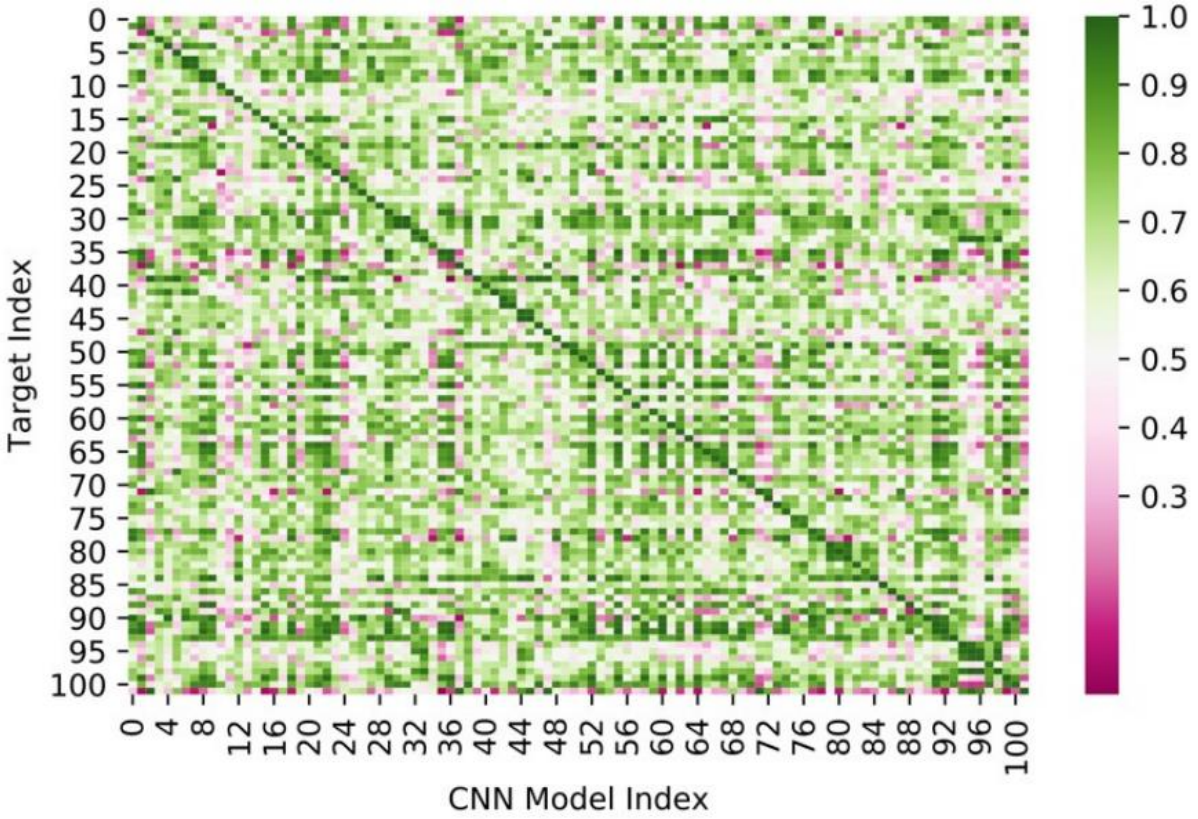
→ protein-ligand 상호작용을 학습하는 것이 아니라, dataset에 있는 intrinsic한 bias를 학습함

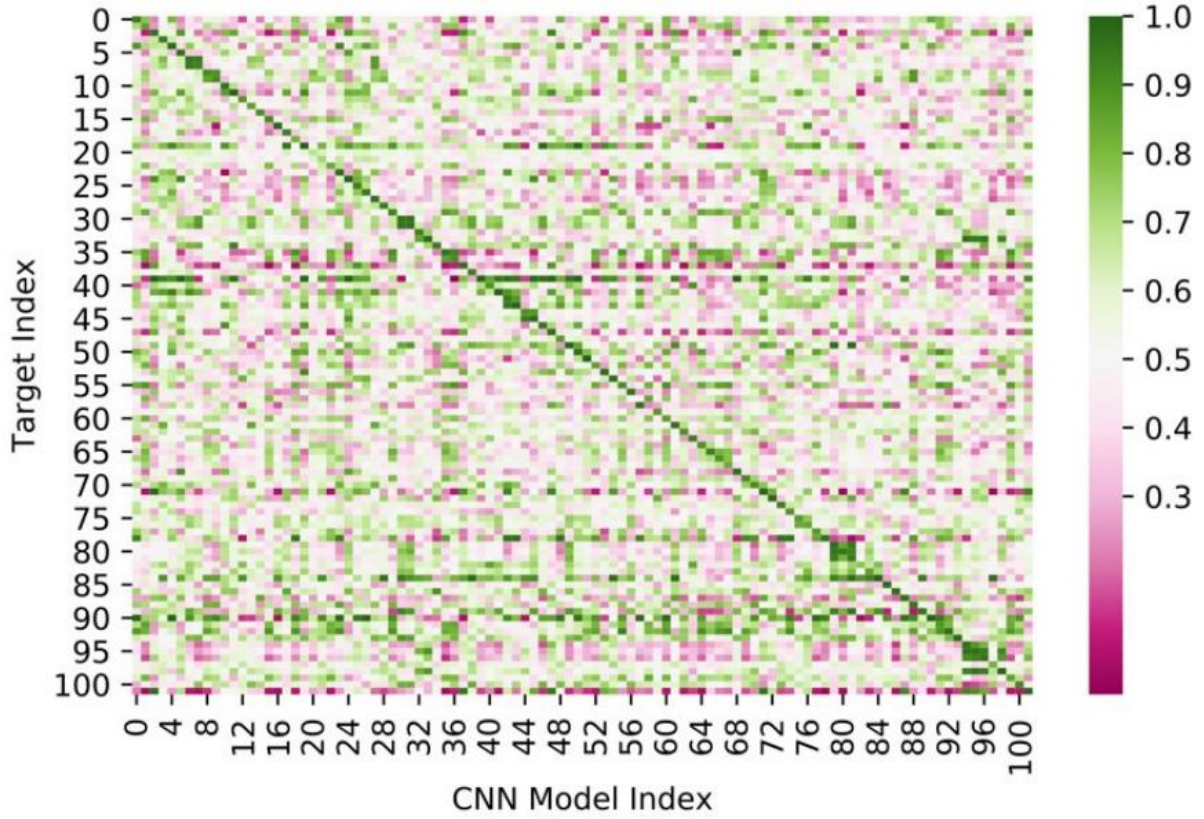


Problem of DTI model in terms of learning intrinsic bias of dataset



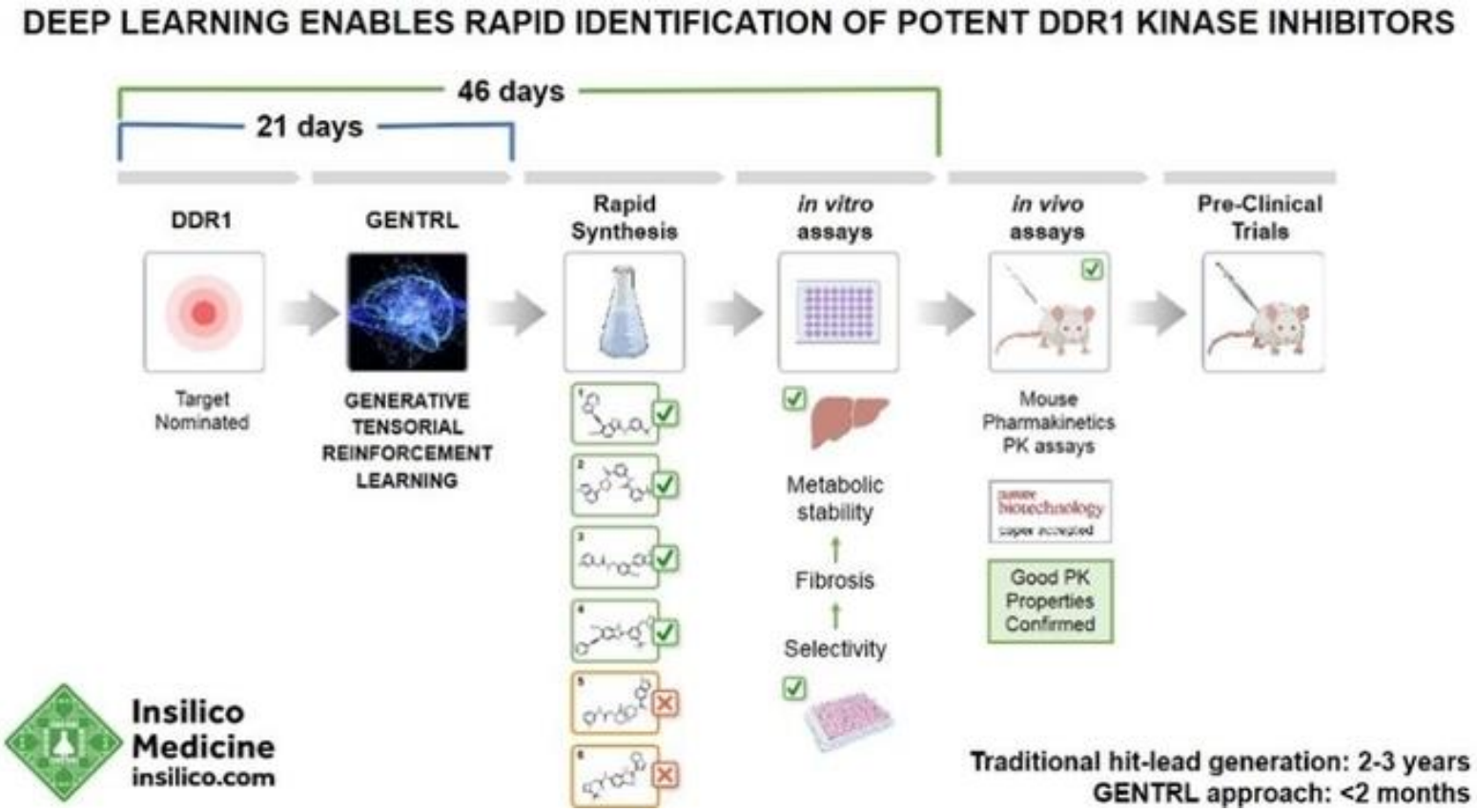
Problem of DTI model in terms of learning intrinsic bias of dataset

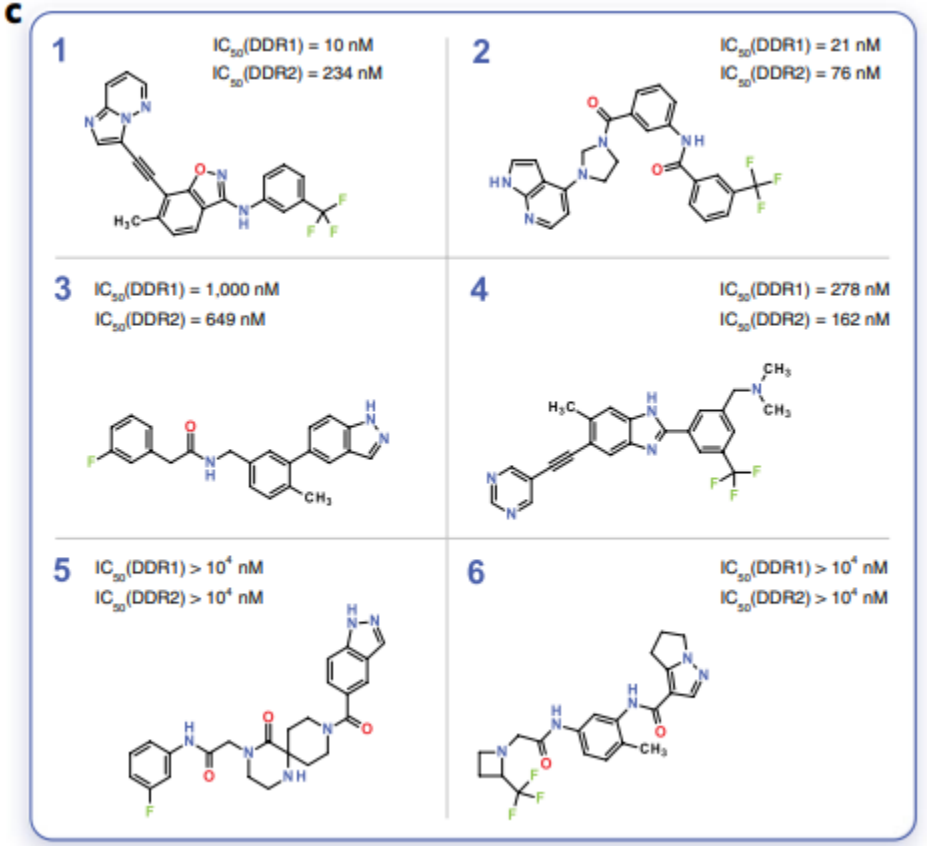




J. Chem. Inf. Model. 2017, 57, 4, 942–957

Molecules with high similarity from molecular generative model



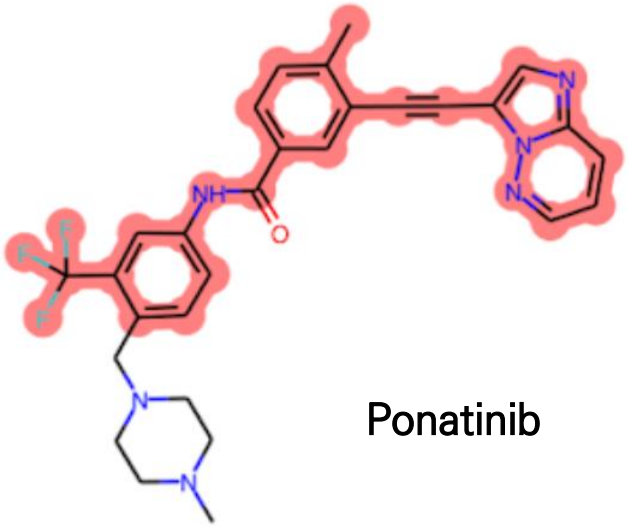


Nature Biotechnology volume 37, pages1038–1040(2019)

Molecules with high similarity from molecular generative model

- 인공지능을 통해 도출한 화합물이 학습에 사용한 약물(Ponatinib)와 구조적으로 매우 유사
- Novel target에 대한 신규약물 발굴의 어려움
- 학습한 데이터가 많지 않기 때문에 유사한 구조가 생성되는 문제가 있음
- 순수 데이터 방법만으로는 해결하기 쉽지 않은 문제
- (biological test는 학습단계에서 고려되지 않음, 해당 과정 통과는 AI로 인한 효과가 아님)

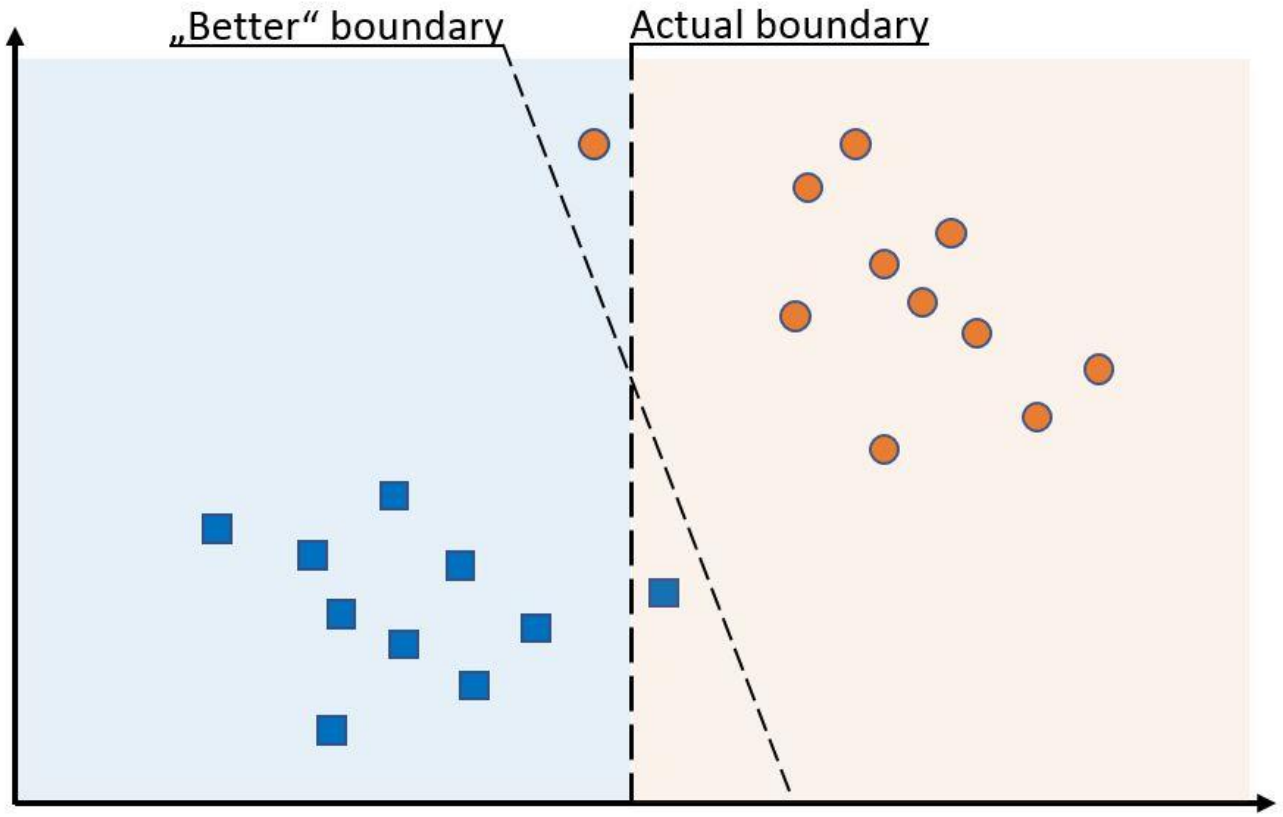
AI 생성 분자



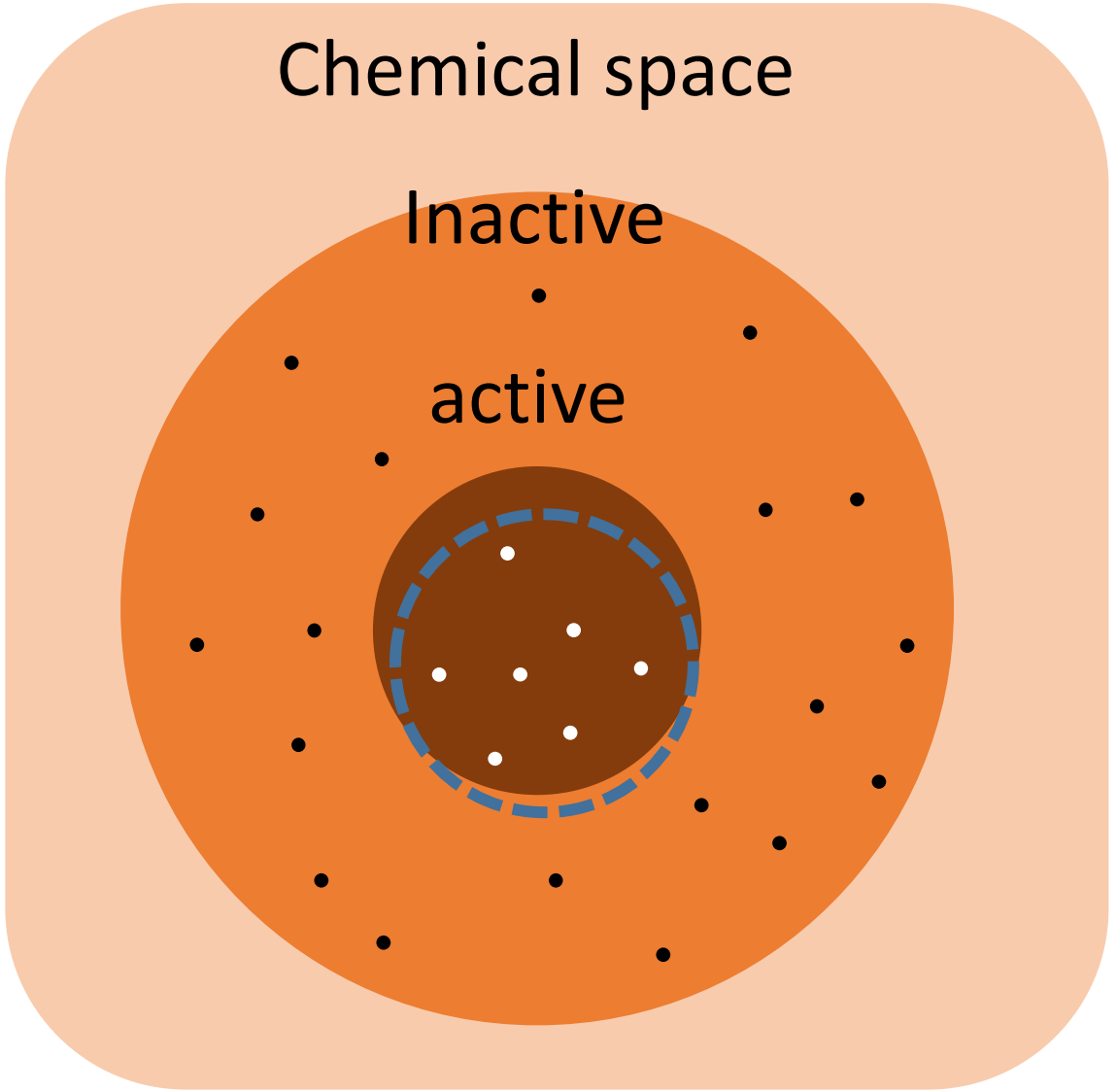
Ponatinib

Origin of low generalization ability

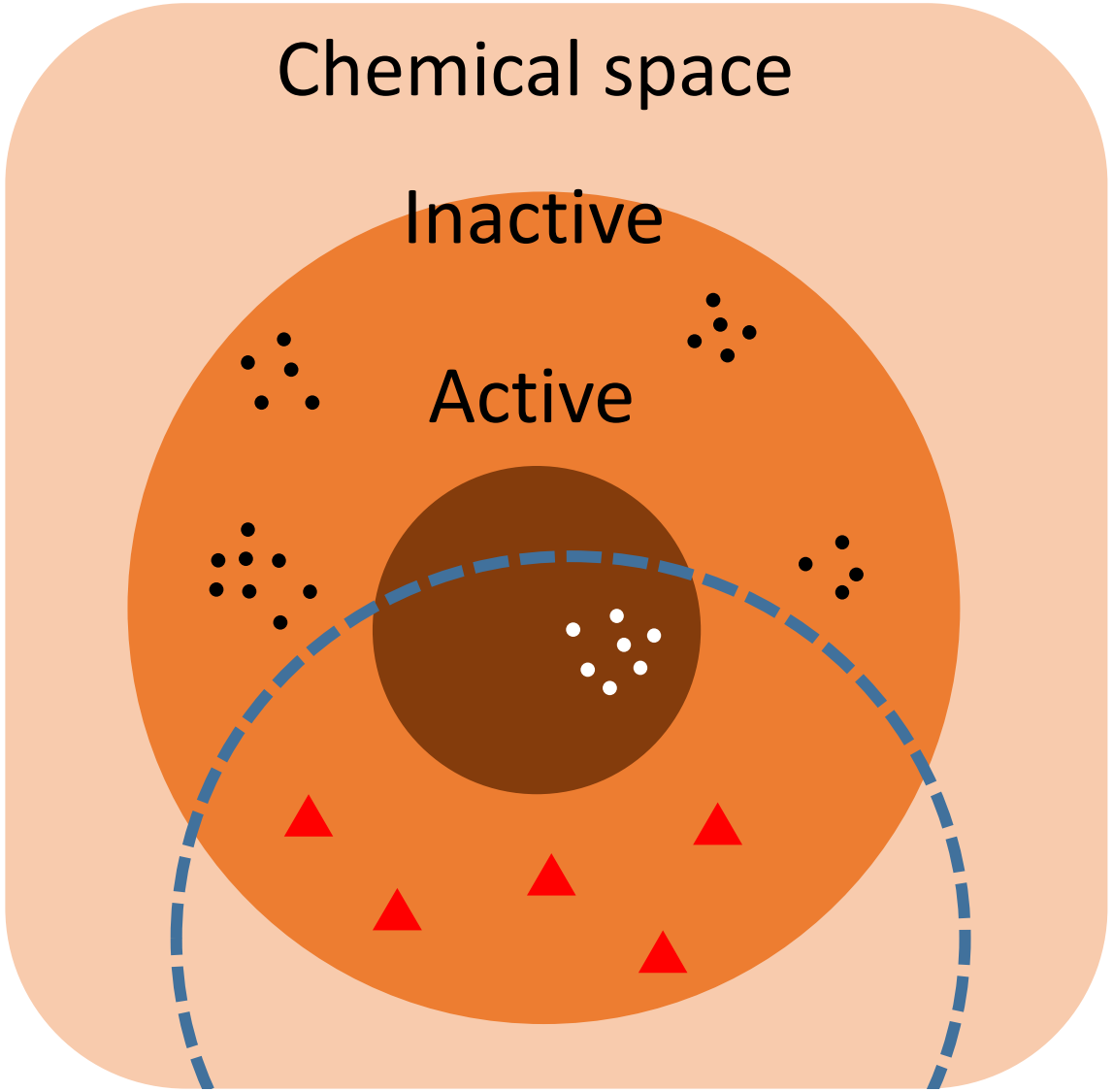
- 이러한 over fitting은 왜 발생하는가?
- 가장 큰 이유는 chemical space 크기에 비해서 데이터가 적기 때문….



Origin of low generalization ability



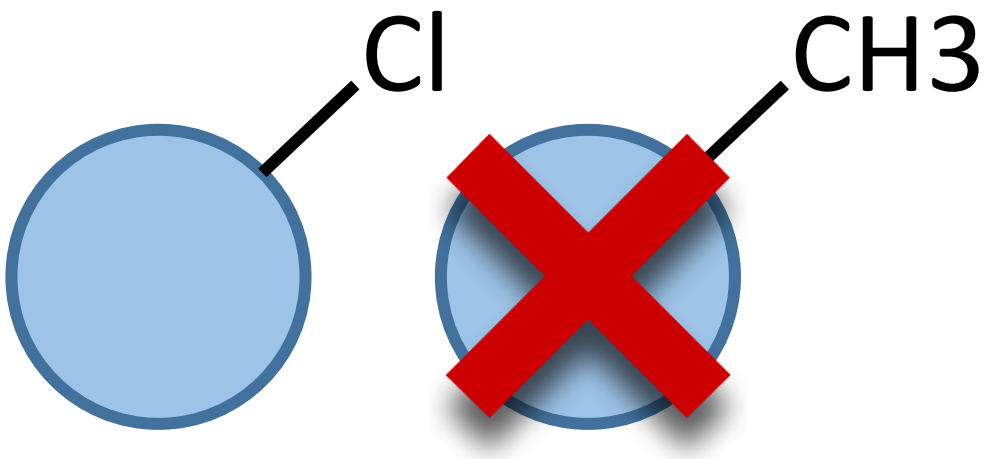
Origin of low generalization ability

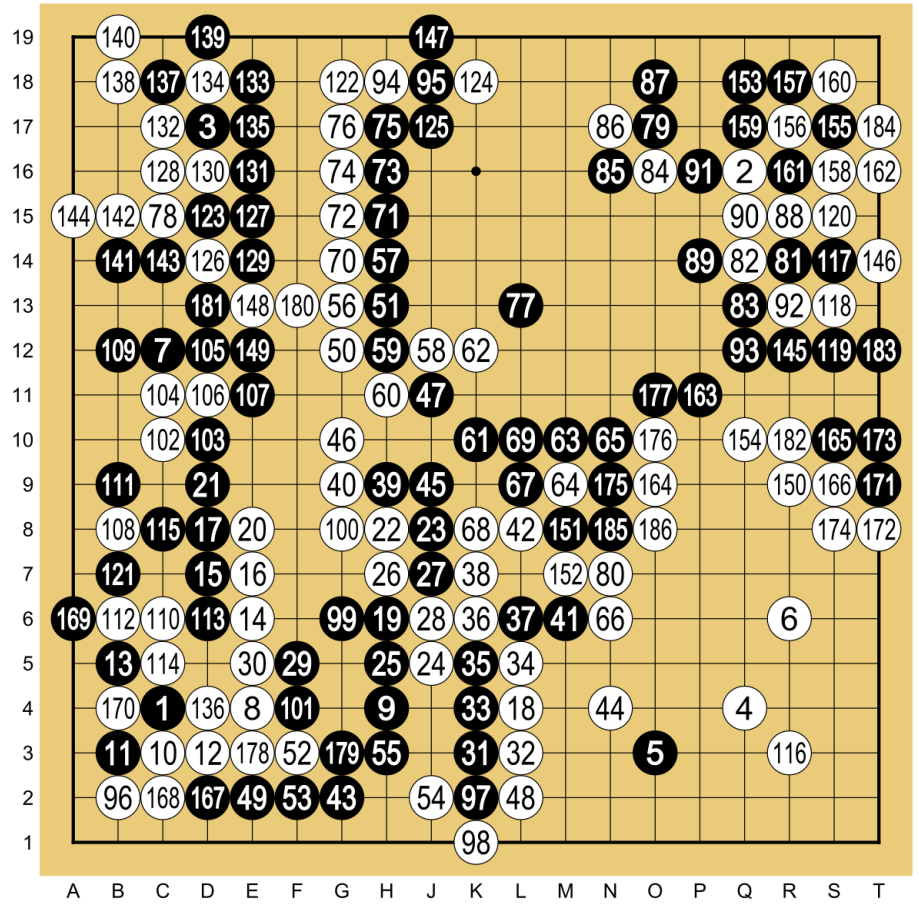


Origin of low generalization ability

- 분자 구조의 작은 변화에도, 활성에서 큰 차이를 나타냄

→ decision boundary가 매우 sensitive해야 하고, 그렇기 위해서는 데이터가 많이 있어야함





Lee Sedol (B) vs AlphaGo (W) - Game 1

Thank you