AI in Drug Discovery

Molecular Representation Learning

2020. 11. 25

Korean AI Center For Drug Discovery and Development

Sooheon Kim

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

REPRESENTATION LEARNING

Tipping Points in Deep Learning

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

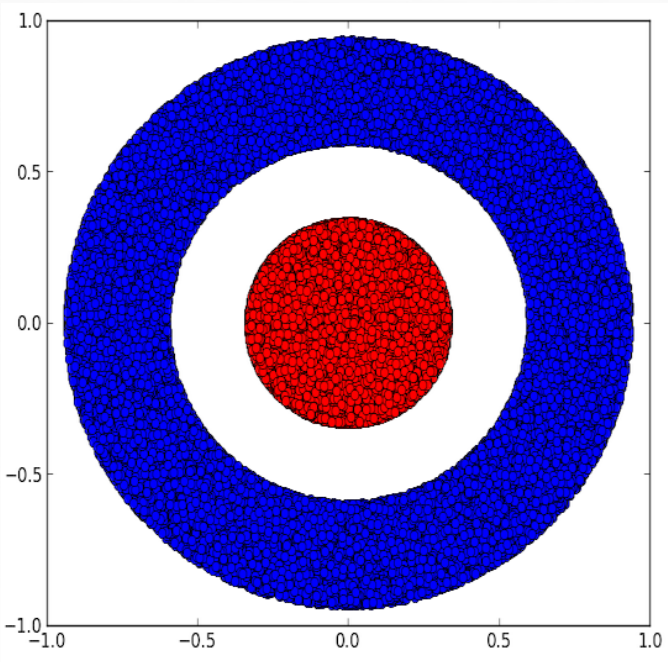
Mode	Dataset	Size	Model
Image	ImageNet	10^8 images	ResNet
Text	CommonCrawl	10^{11} tokens	GPT-3
Compound	PubChem, ChEMBL	10^7 compounds	???
Protein	UniRef100, BFD	10^9 proteins, 10^{11} amino acids	???

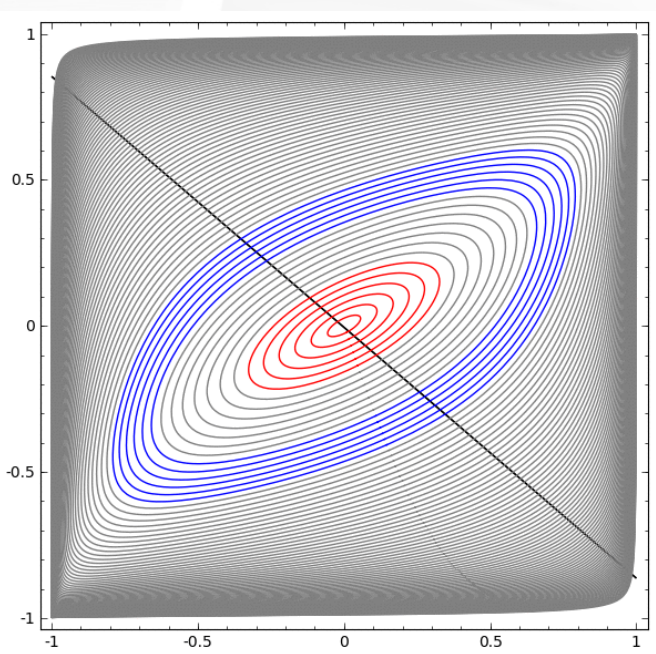
Tipping Points in Deep Learning

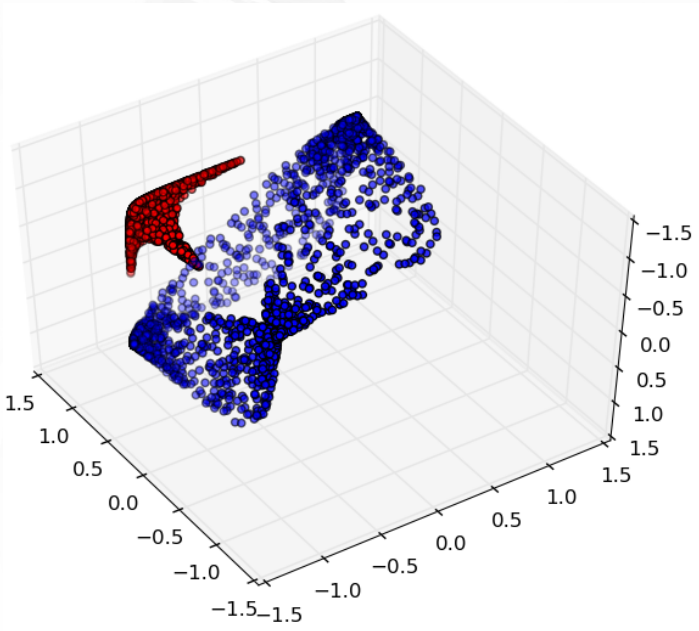
- How big is image space?
  - Pixels: 512x512
  - Color channels: 3
  - Color values: 256 (for 8 bit)
  - $256^{(3*512*512)} = 10^{1893916}$
- How big is language space?
  - 50k token vocabulary ^ 2048 context length = 10^{9623}
- How big is chemical space?
  - 10^{60}?
- Protein space?
  - 21 amino acids ^ 4096 protein length = $6 * 10^{5415}$

Representation Learning

- 데이터를 변형하여 추후 Task (분류, 회귀 등)에 용이하게 만드는 것
  - 표현이 적합할수록 모델의 복잡도는 내려갈 수 있다 (Ex. 딥러닝 모델의 Head로 얕은 Linear Model 활용)





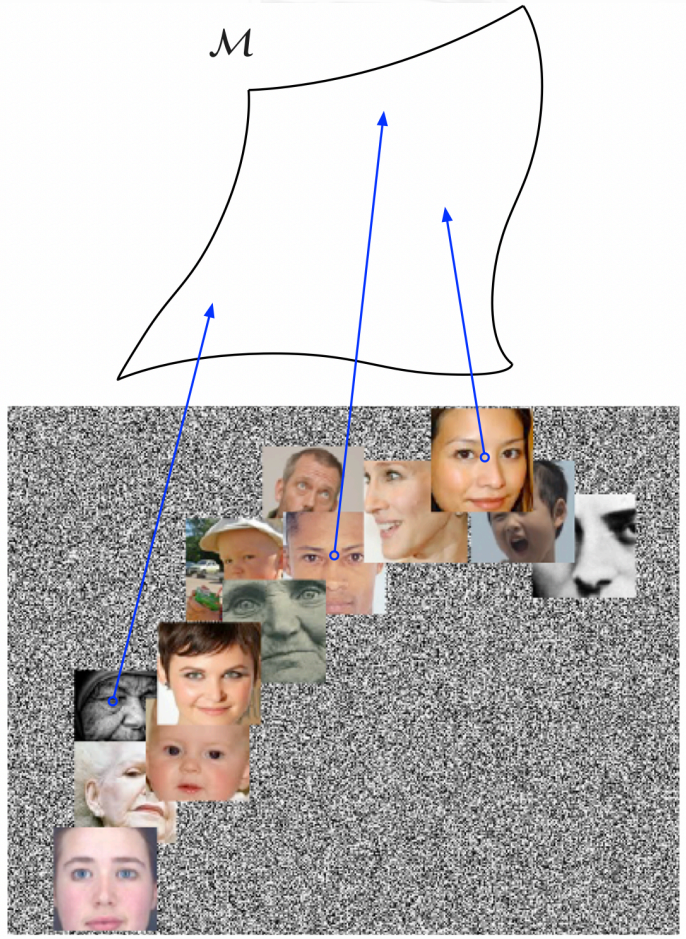


https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

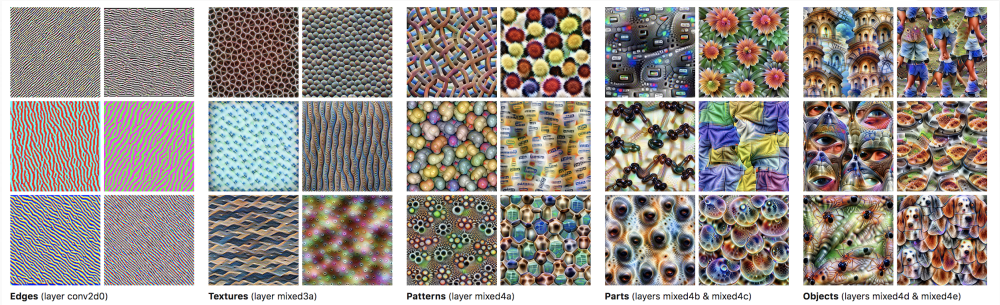
Representation Learning

Manifold Hypothesis

- 고차원으로 표현되는 현상데이터의 실제 **유용한** 정보는 저차원 manifold M에 몰려있다
  - Random image (256^(512^2)) vs. images of human face
  - Chemical space (10^60) vs. hits for given target
- 이 Manifold는 구조상 많이 "구겨지고 꼬여"있을 가능성이 높으니, 좋은 representation learning의 목적 중 하나는 이를
  - 찾아내고
  - "펴"주는 것



Representation Learning



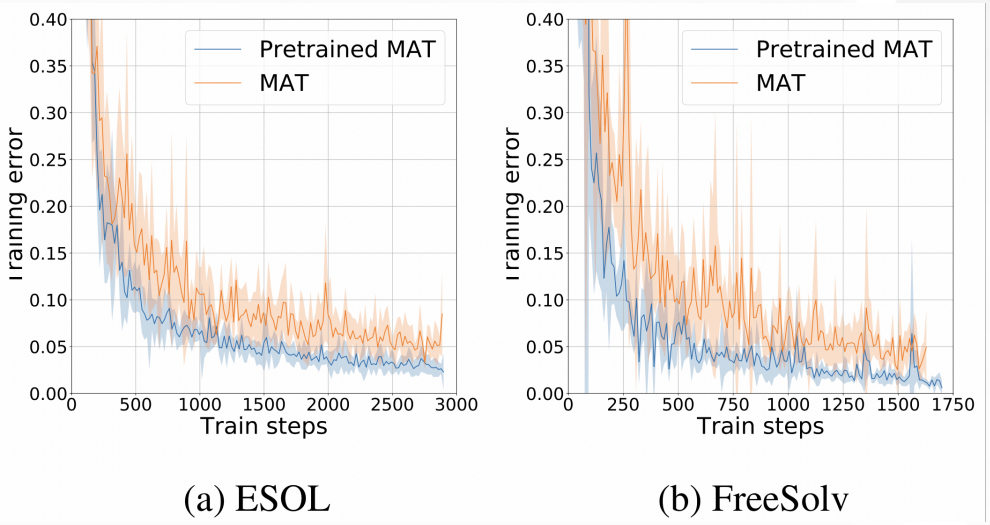
- a와 b가 유사하면 각각의 representation도 유사할 것
  - $a \approx b \Longrightarrow repr(a) \approx repr(b)$
- 다양한 Task, Dataset에 적용 가능할 것
  - Generalization이 잘 되야 유용
- 의미/특징의 Hierarchy가 내재되어 있으면 좋음

Szegedy et al., 2014

Representation Learning

”더이상 전이학습 없이 딥러닝 사용하는 것은 무모하다.”

-- Laurens van der Maaten



(a) ESOL (b) FreeSolv

Szegedy et al., 2014

Representation Learning

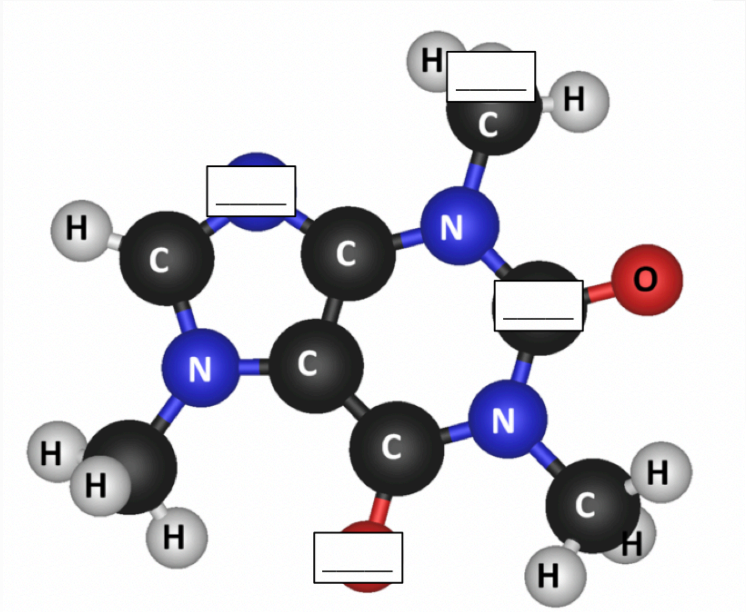
- Supervised
  - X -> explicit Y
  - Ex) “문제지 풀고 답 확인하기”
- Unsupervised
  - No **direct** supervision, but requires **indirect** supervision via task engineering
  - Ex) “읽다가 가려진 단어 뭘까 고민해본다”
  - 답안지 없이도 학습이 가능
    - 대이터에 내재된 패턴을 “답”으로 간주

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

MOLECULE PRETRAINING

Pretraining

- Chemical 데이터에 Unsupervised task 구성은 어떻게?



It was a _____ , cold November day. I
(adjective)
woke up to the _____ smell of _____
(adjective) (type of bird)
roasting in the _____ downstairs. I
(room in a house)
_____ down the stairs to see if I could
(verb (past tense))
help _____ the dinner. My mom said,
(verb)
"See if _____ needs a fresh _____." So I
(relative's name) (noun)
carried a tray of glasses full of _____ into
(a liquid)
the _____ room. When I got there, I
(verb ending in -ing)
couldn't believe my _____! There were
(part of the body (plural))
_____ _____ on the _____!
(plural noun) (verb ending in -ing) (noun)

Pretraining

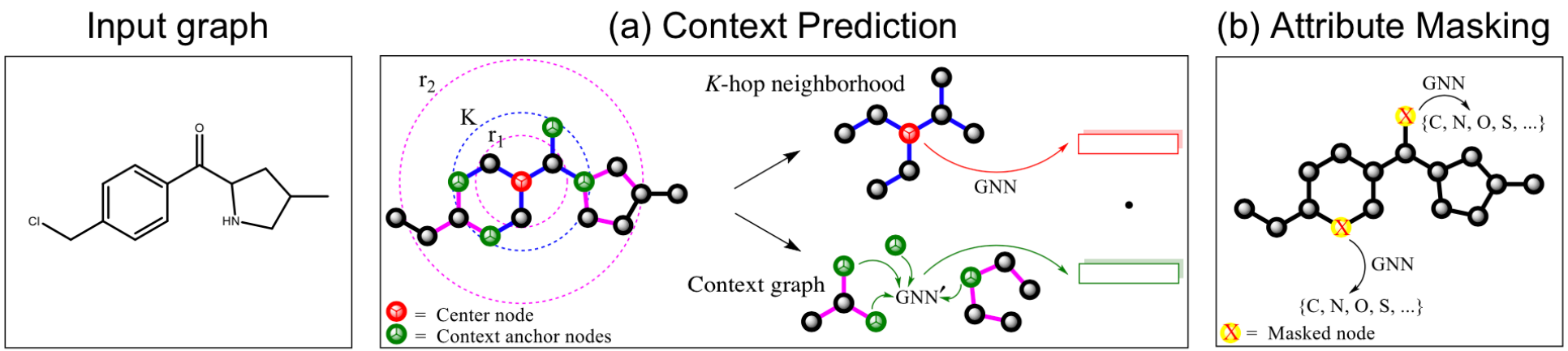


Figure 2: Illustration of our node-level methods, Context Prediction and Attribute Masking for pre-training GNNs. **(a)** In Context Prediction, the subgraph is a K-hop neighborhood around a selected center node, where K is the number of GNN layers and is set to 2 in the figure. The context is defined as the surrounding graph structure that is between r_1- and r_2-hop from the center node, where we use $r_1 = 1$ and $r_2 = 4$ in the figure. **(b)** In Attribute Masking, the input node/edge attributes (*e.g.*, atom type in the molecular graph) are randomly masked, and the GNN is asked to predict them.

Strategies for Pre-Training Graph Neural Networks. Hu et al., 2020

Pretraining

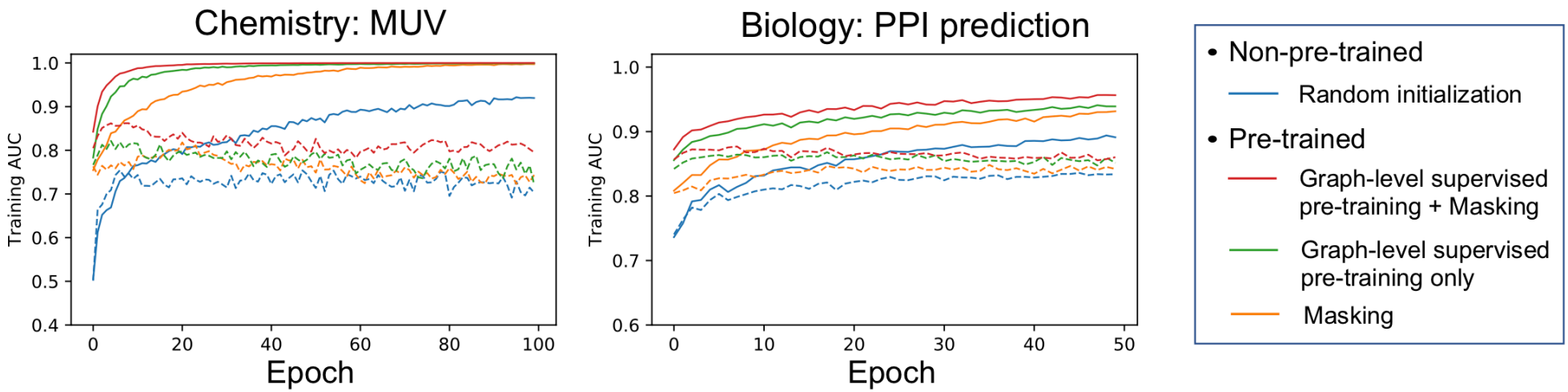


Figure 4: **Training and validation curves of different pre-training strategies on GINs.** Solid and dashed lines indicate training and validation curves, respectively.

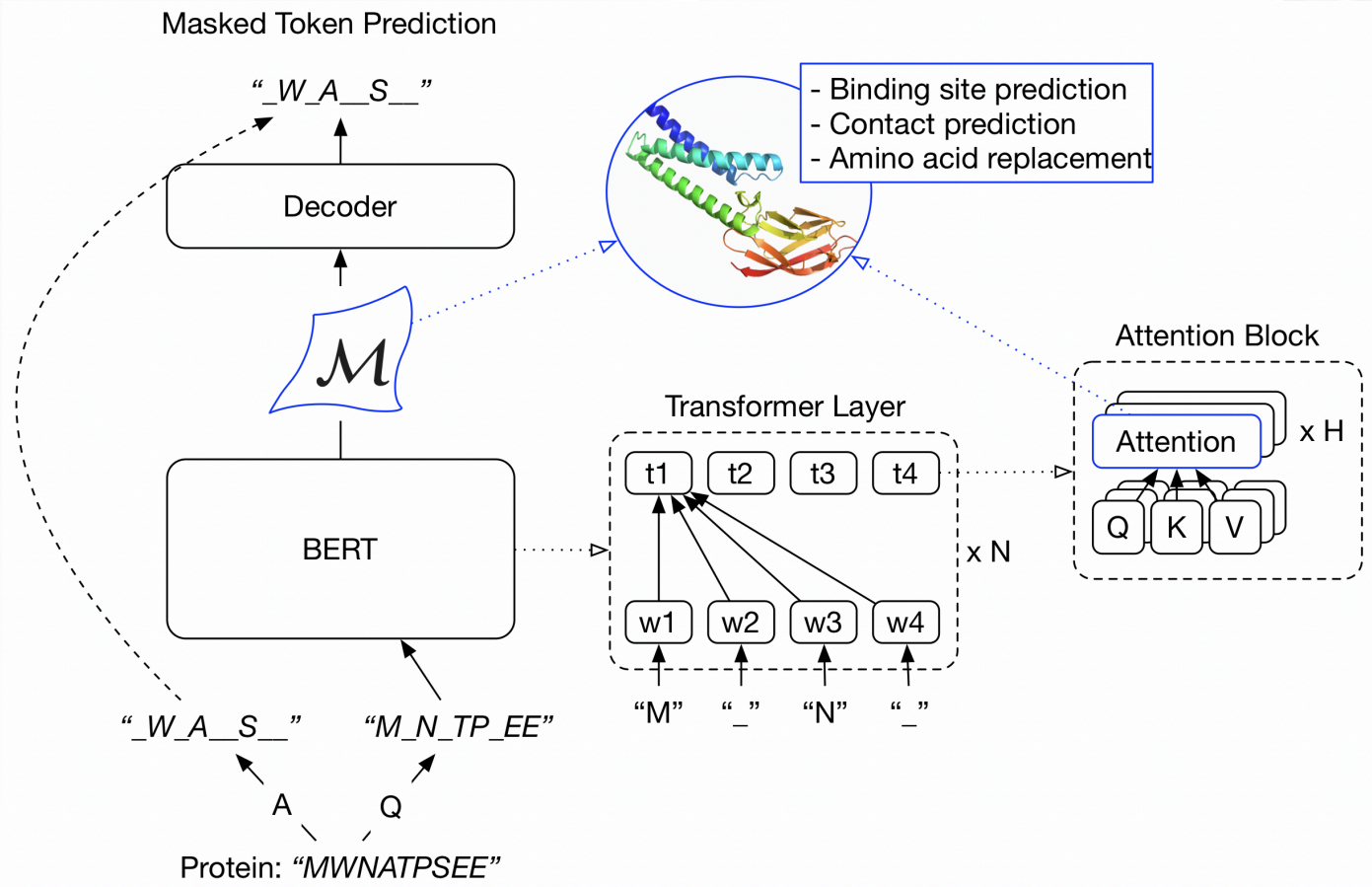
- MUV: Maximum Unbiased Validation set (PCBA subset of 93k mols)
- PPI: 395k unlabeled protein ego-networks

Strategies for Pre-Training Graph Neural Networks. Hu et al., 2020

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

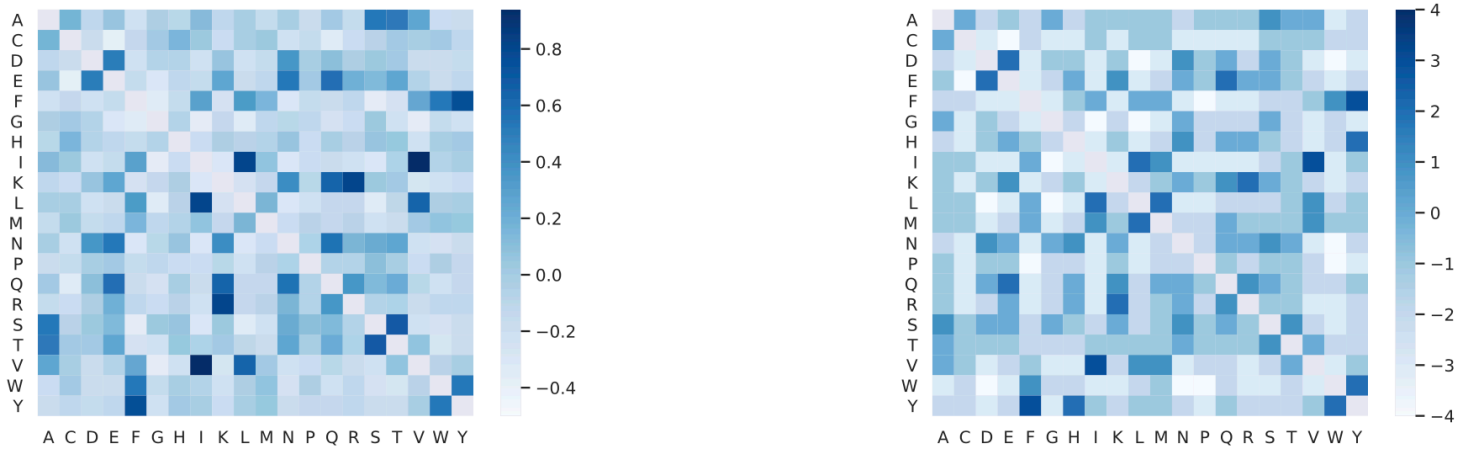
PROTEIN PRETRAINING

Protein BERT



Protein BERT

- Amino acid substitution relationship
  - Pearson r = 0.80 to BLOSUM62 matrix



(a) Attention similarity

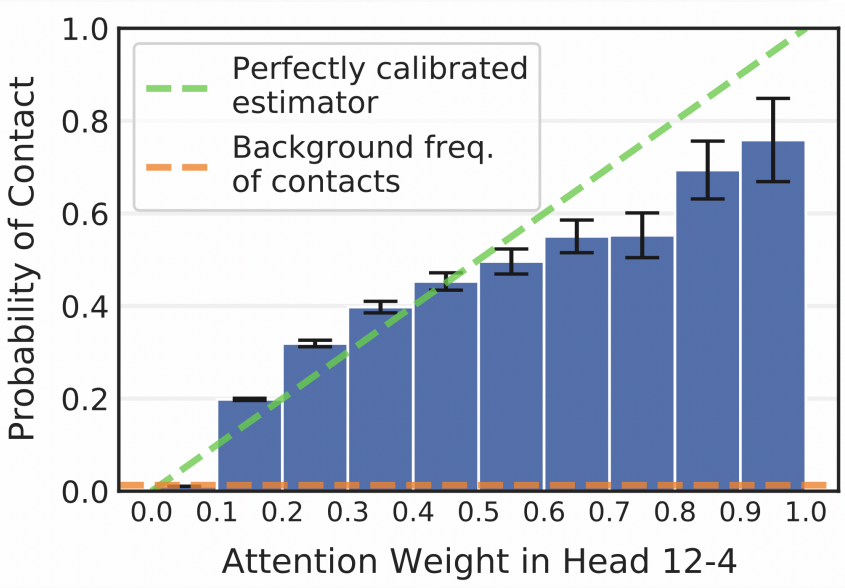
(b) BLOSUM62 substitution scores

Figure 3: Comparison of attention similarity matrix with the substitution matrix. Each matrix entry represents an amino-acid pair (codes in Appendix B.1). The two matrices have a Pearson correlation of 0.80 with one another, suggesting that attention is largely consistent with substitution relationships.

BERTology Meets Biology. Vig et al., 2020

Protein BERT

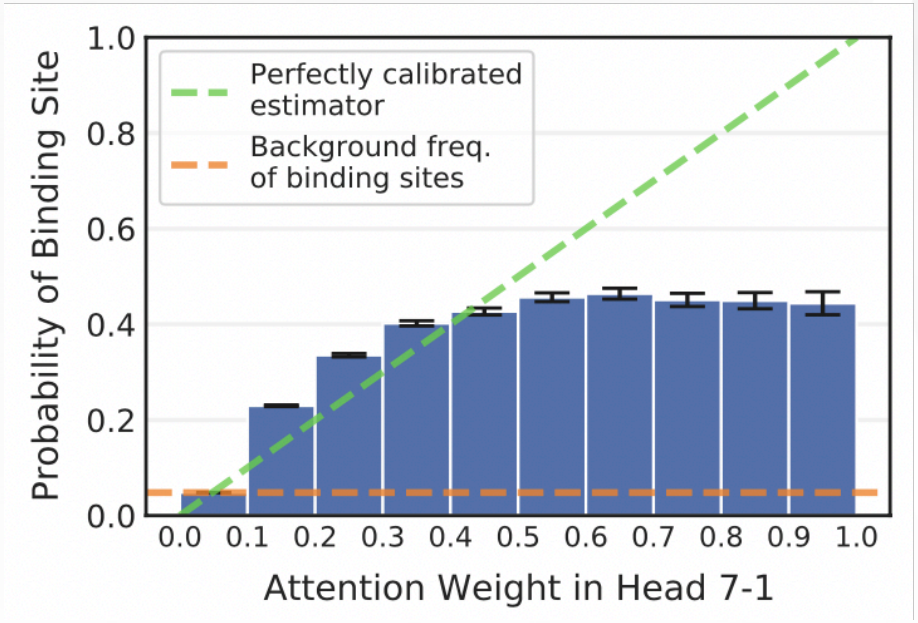
- Contact prediction
  - Ex) de novo TIM-barrel (5BVL) 내 서열상 거리가 머나 구조상 인접한 residue 인식

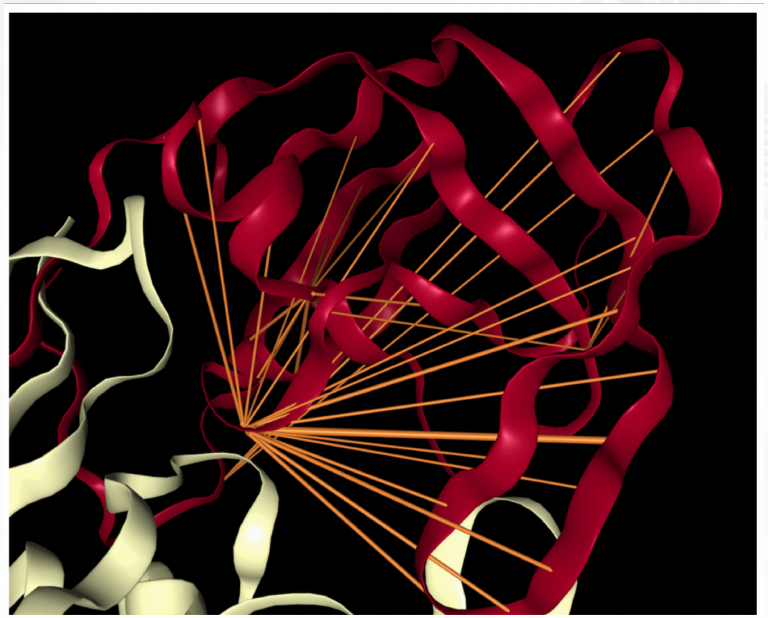


BERTology Meets Biology. Vig et al., 2020

Protein BERT

- Binding site prediction
  - Ex) HIV-1 protease (7HVP) binding site 인식





BERTology Meets Biology. Vig et al., 2020

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

Thank You