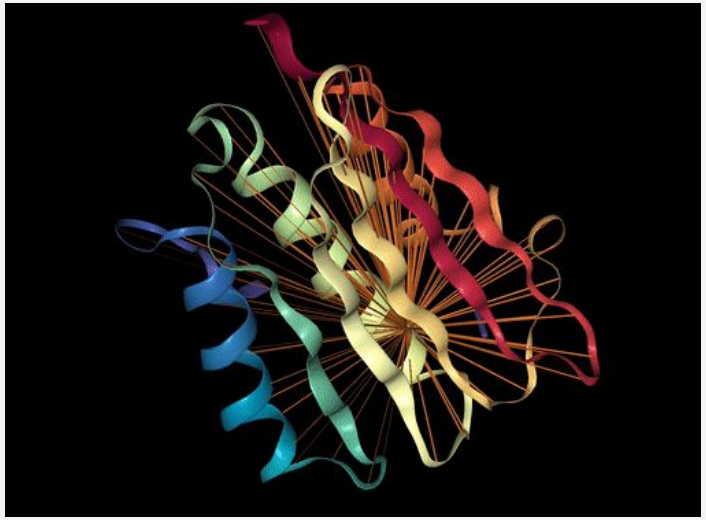
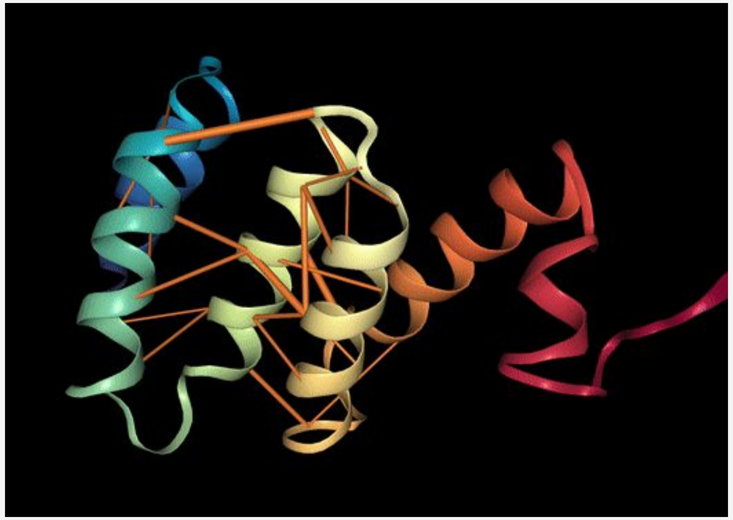
BERTology Meets Biology
Interpreting Attention in Protein Language Models

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong,
Richard Socher, Nazneen Fatema Rajani

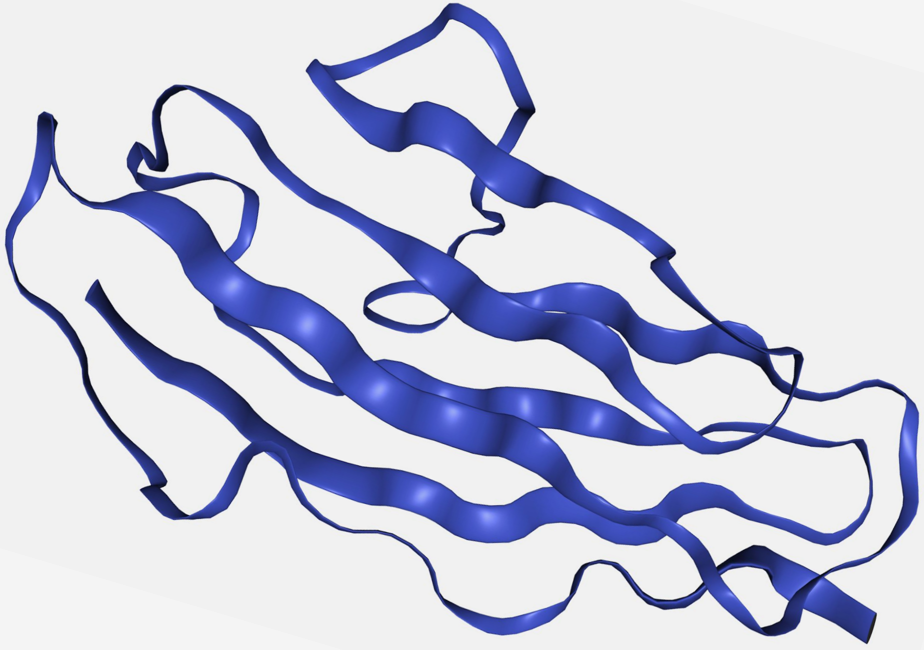
Salesforce Research



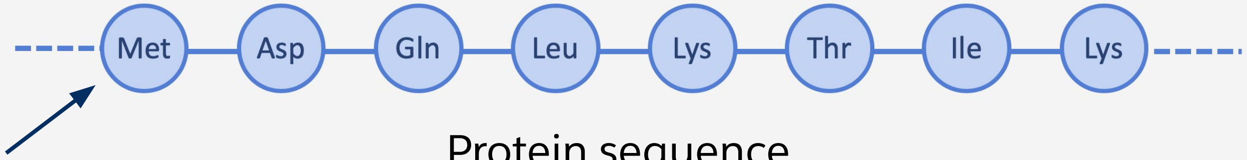


Background: Proteins

- Proteins are complex molecules that play critical role in function and structure of all organisms
- Understanding proteins key to disease therapies
- Other applications such as material science



Protein structure

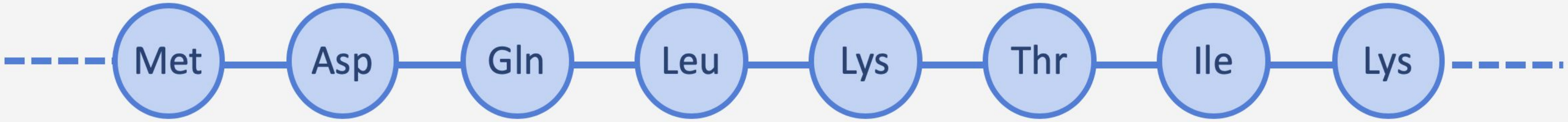


Met Asp Gln Leu Lys Thr Ile Lys

Protein sequence

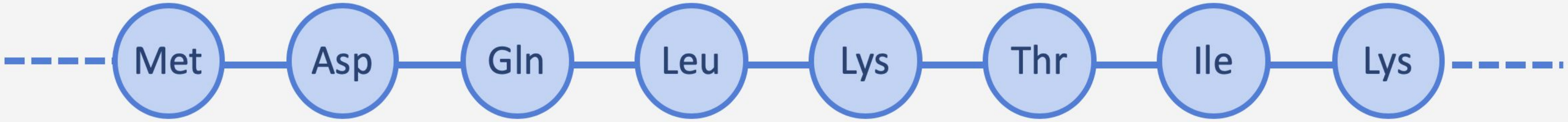
Amino Acid

Background: Proteins as language



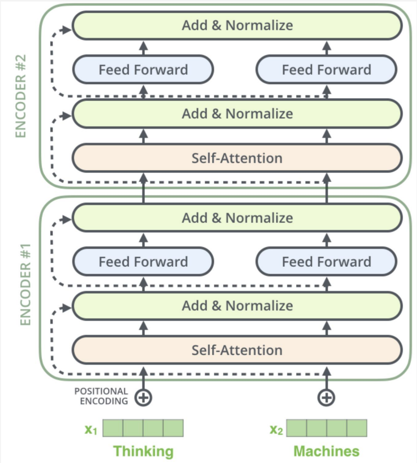
The quick brown fox jumps over the lazy dogs

Background: Proteins as language



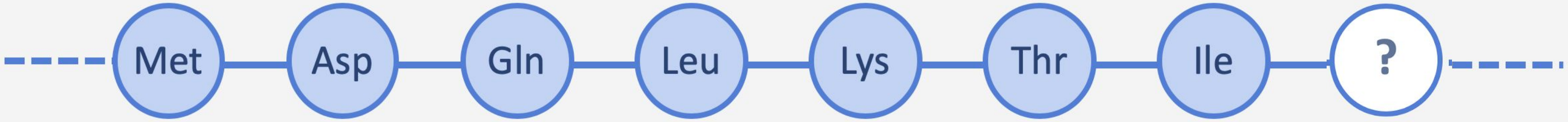
Met Asp Gln Leu Lys Thr Ile Lys

The quick brown fox jumps over the lazy dogs

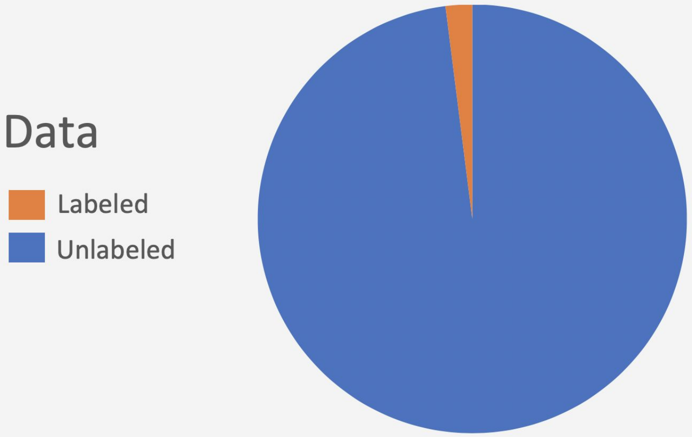


from “The Illustrated Transformer” by Jay Alammar

Background: Proteins as language



The quick brown fox jumps over the lazy ______

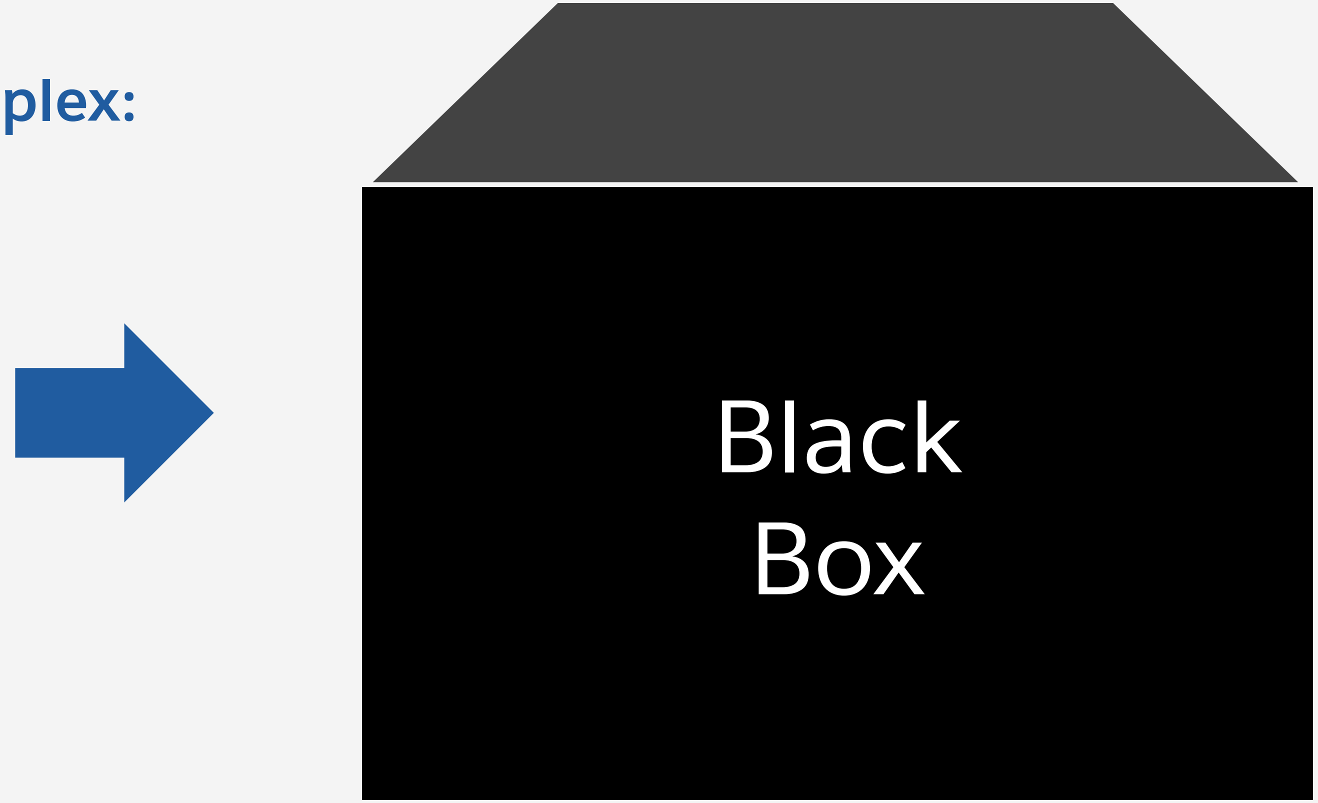


~2,000,000,000 protein sequences available

What does a Transformer language model learn?

Transformers are **large** and **complex:**

- BERT-Base: **110M** Parameters
- BERT-Large: **340M** Parameters
- CTRL: **1.6B** Parameters
- GPT3: **175B** Parameters
- GSHARD: **650B** Parameters



What does a Transformer language model learn?

Transformers are **large** and **complex:**

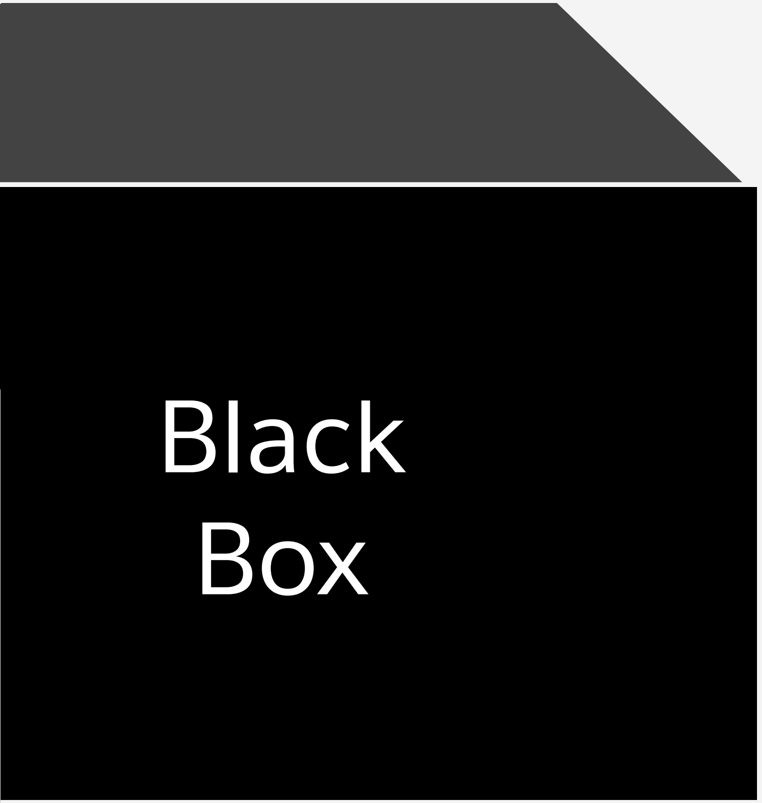
BERT-Base: **110M** Parameters

BERT-Large: **340M** Parameters

CTRL: **1.6B**

GPT3: **175**

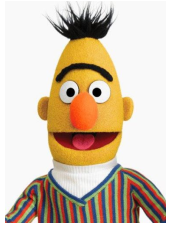
GSHARD: **6**



bert·ol·o·gy

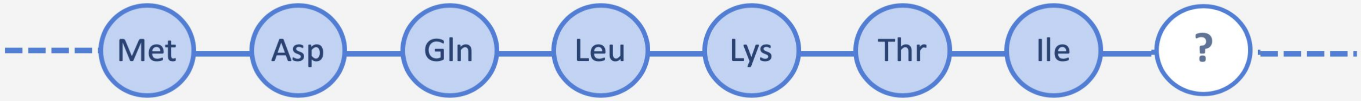
/bərtˈäləjē/

noun

The study of 

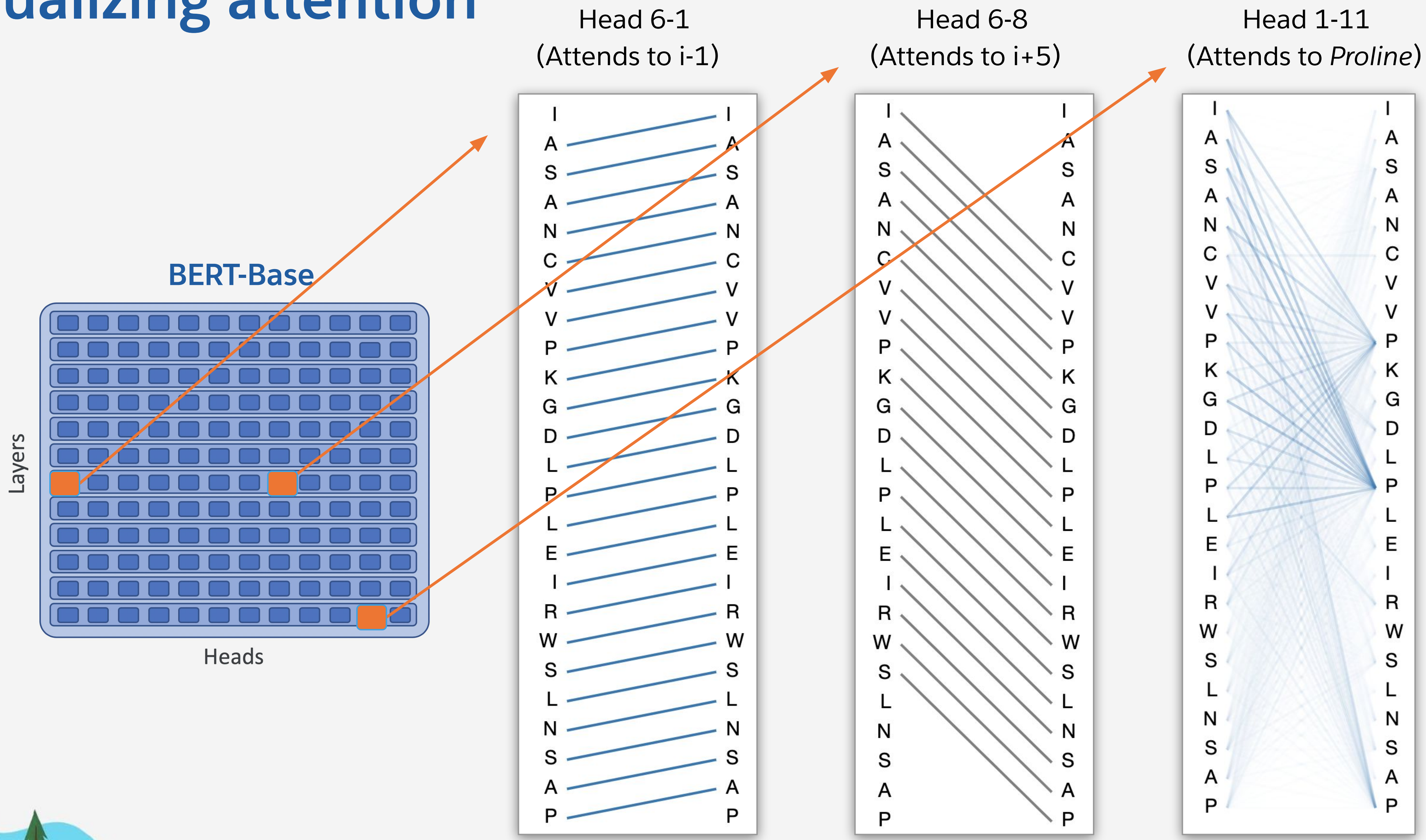
BERTology Meets Biology

- **Models:** BERT (3 variations), XLNet, ALBERT from TAPE, ProtTrans repos
- **Pre-training datasets:** Pfam, Uniref100, BFD
- **Analysis:** Attention for 5000 protein sequences



Language Modeling

Visualizing attention



BERT-Base

Layers

Heads

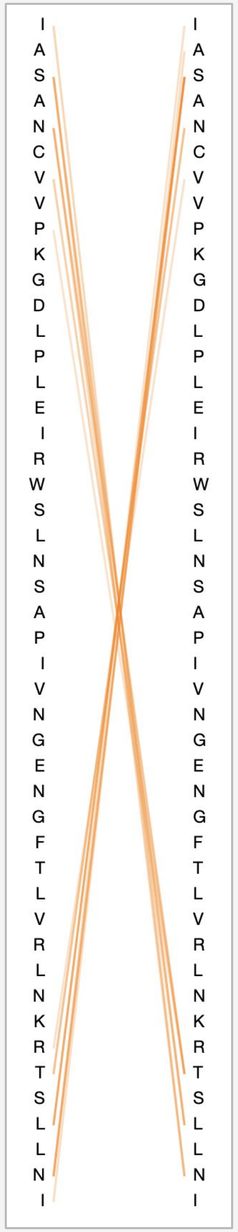
Head 6-1
(Attends to i-1)

Head 6-8
(Attends to i+5)

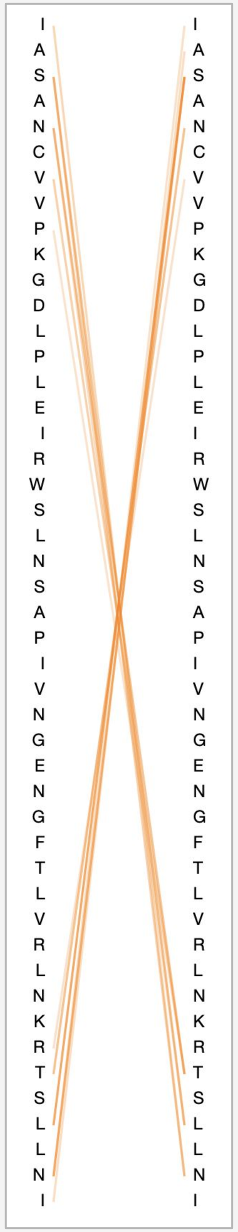
Head 1-11
(Attends to *Proline*)

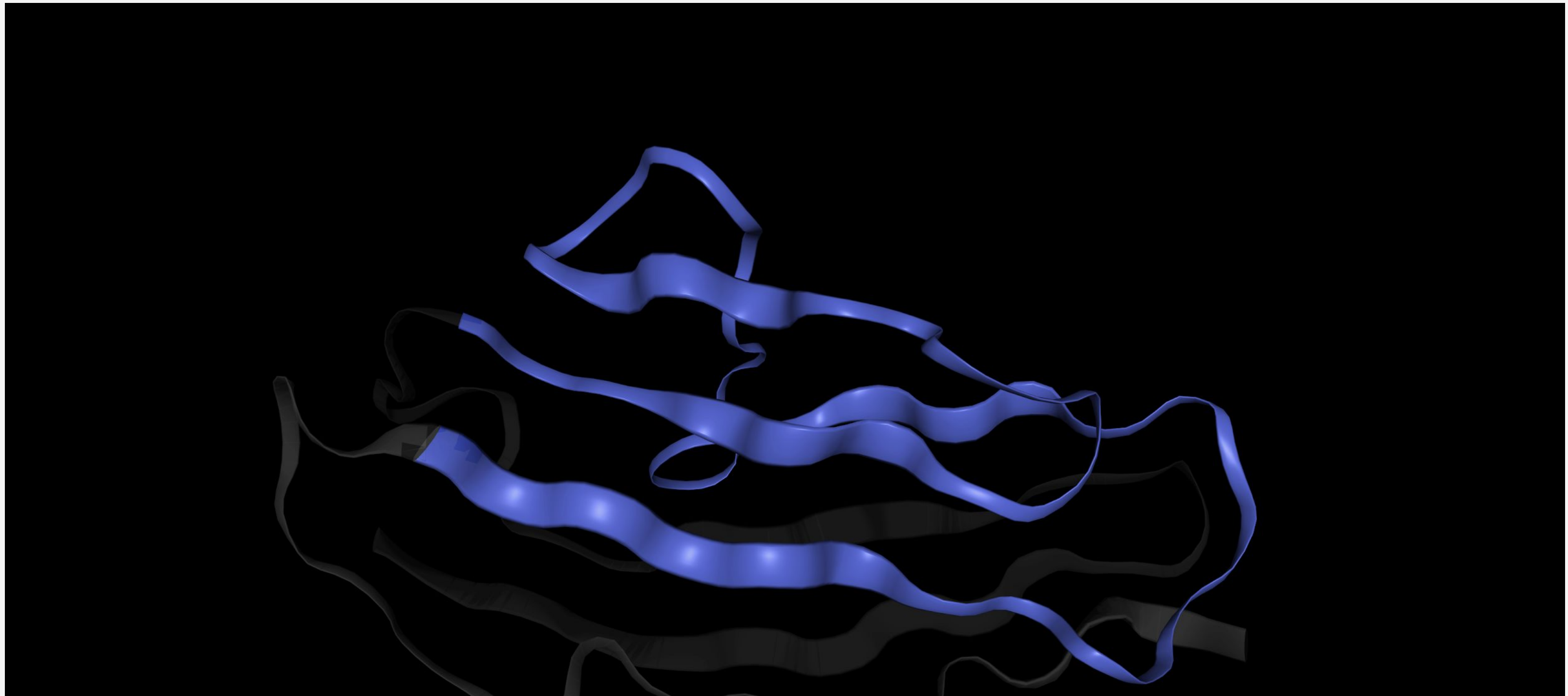
https://github.com/jessevig/bertviz

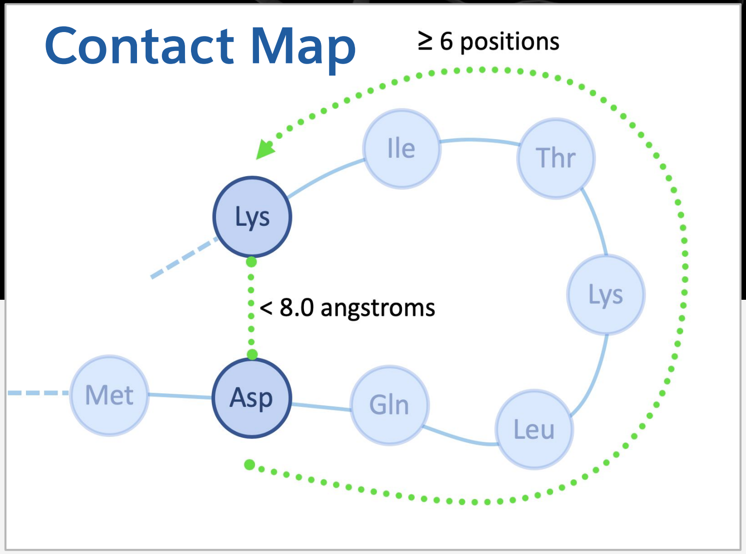
Head 12-4


I
A
S
A
N
C
V
V
P
K
G
D
L
P
L
E
I
R
W
S
L
N
S
A
P
I
V
N
G
E
N
G
F
T
L
V
R
L
N
K
R
T
S
L
L
N
I
I
A
S
A
N
C
V
V
P
K
G
D
L
P
L
E
I
R
W
S
L
N
S
A
P
I
V
N
G
E
N
G
F
T
L
V
R
L
N
K
R
T
S
L
L
N
I

Head 12-4


I A S A N C V V P K G D L P L E I R W S L N S A P I V N G E N G F T L V R L N K R T S L L N I
I A S A N C V V P K G D L P L E I R W S L N S A P I V N G E N G F T L V R L N K R T S L L N I



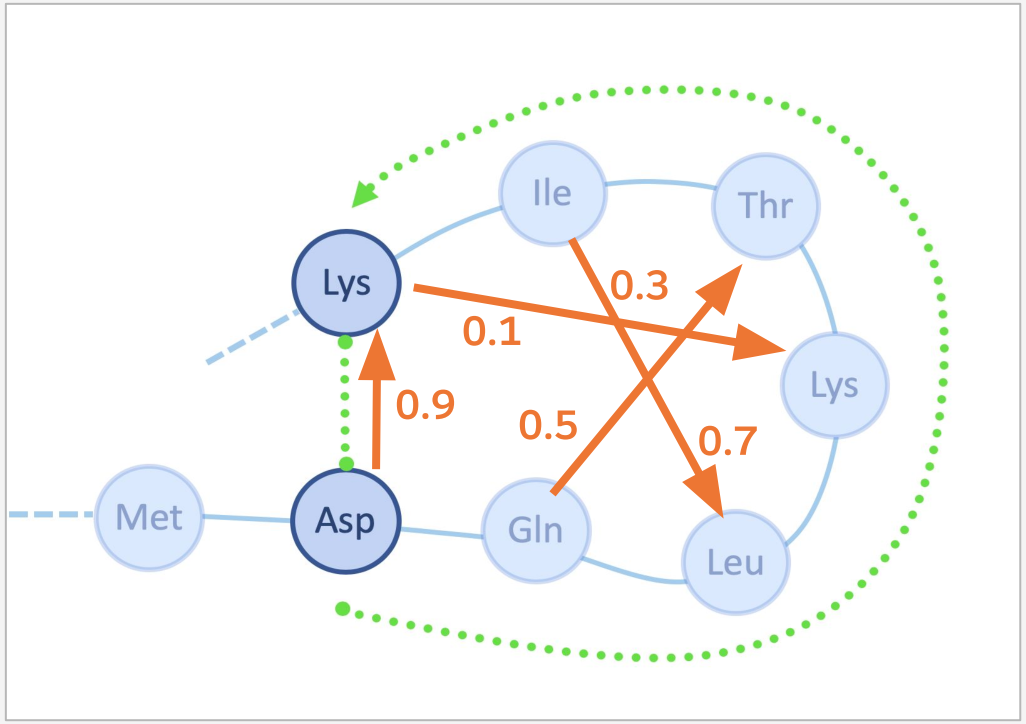


Contact Map

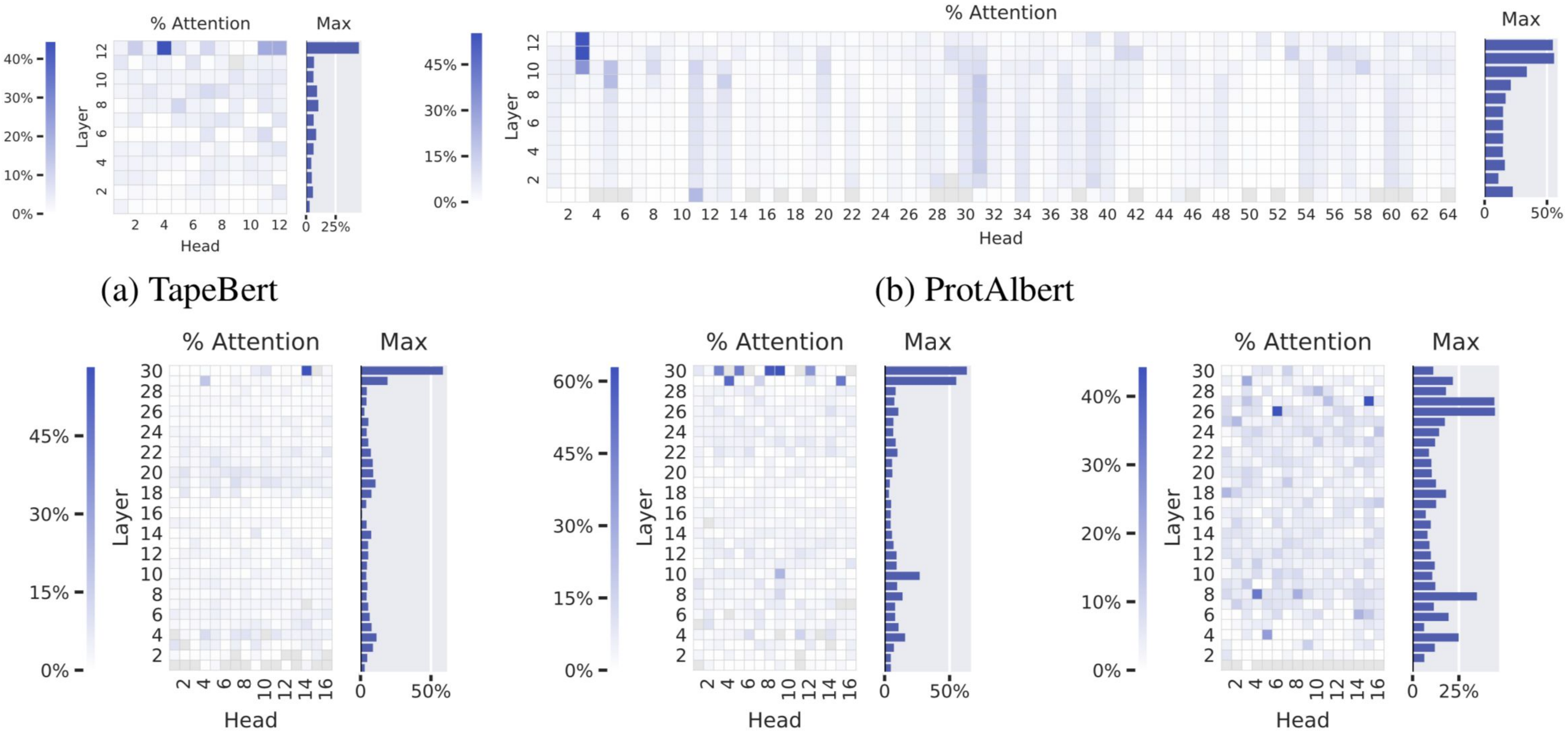
≥ 6 positions

< 8.0 angstroms

Does attention align with contacts?



Dataset: 5000 protein sequences (1M+ amino acids)



(a) TapeBert

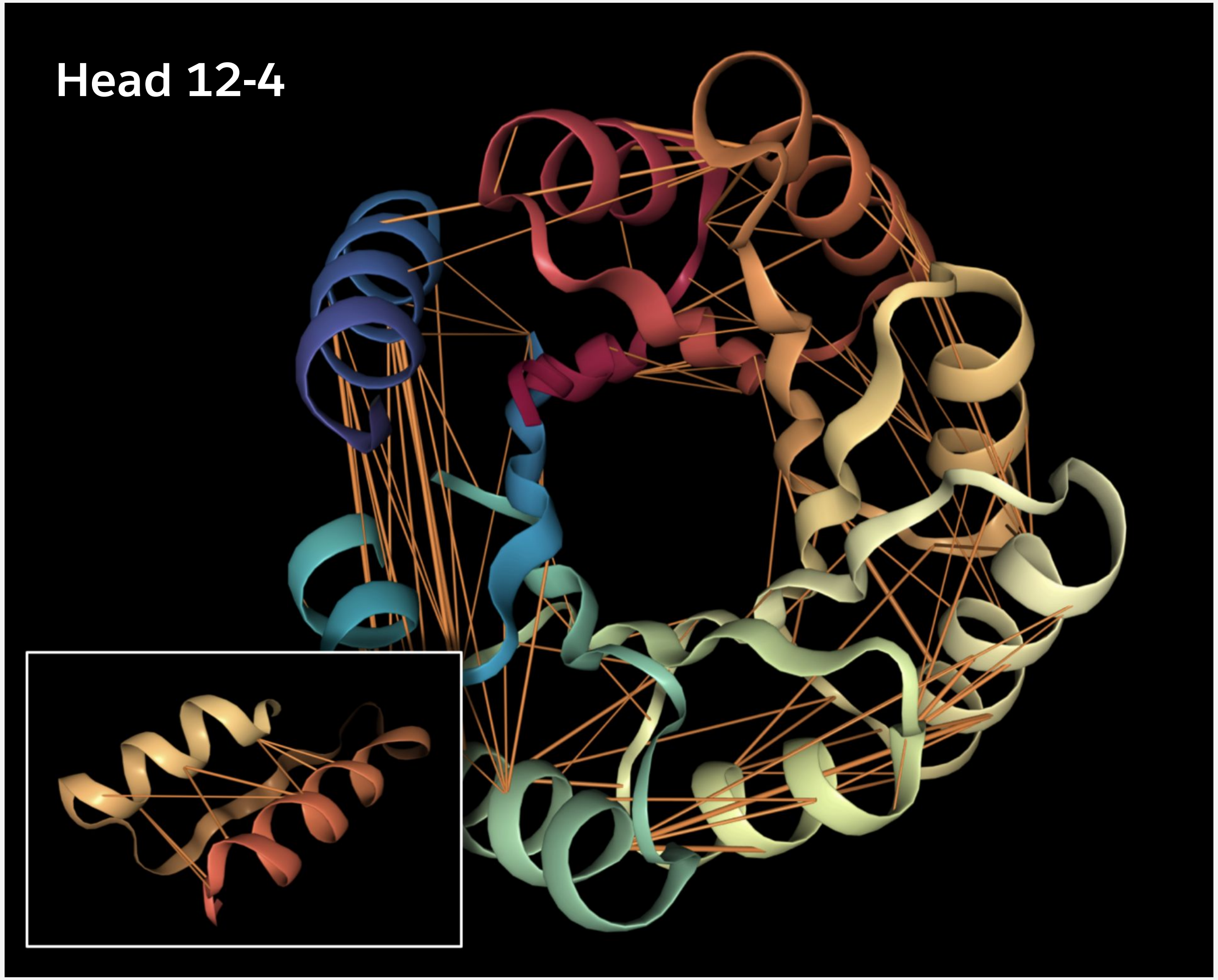
(b) ProtAlbert

(c) ProtBert

(d) ProtBert-BFD

(e) ProtXLNet

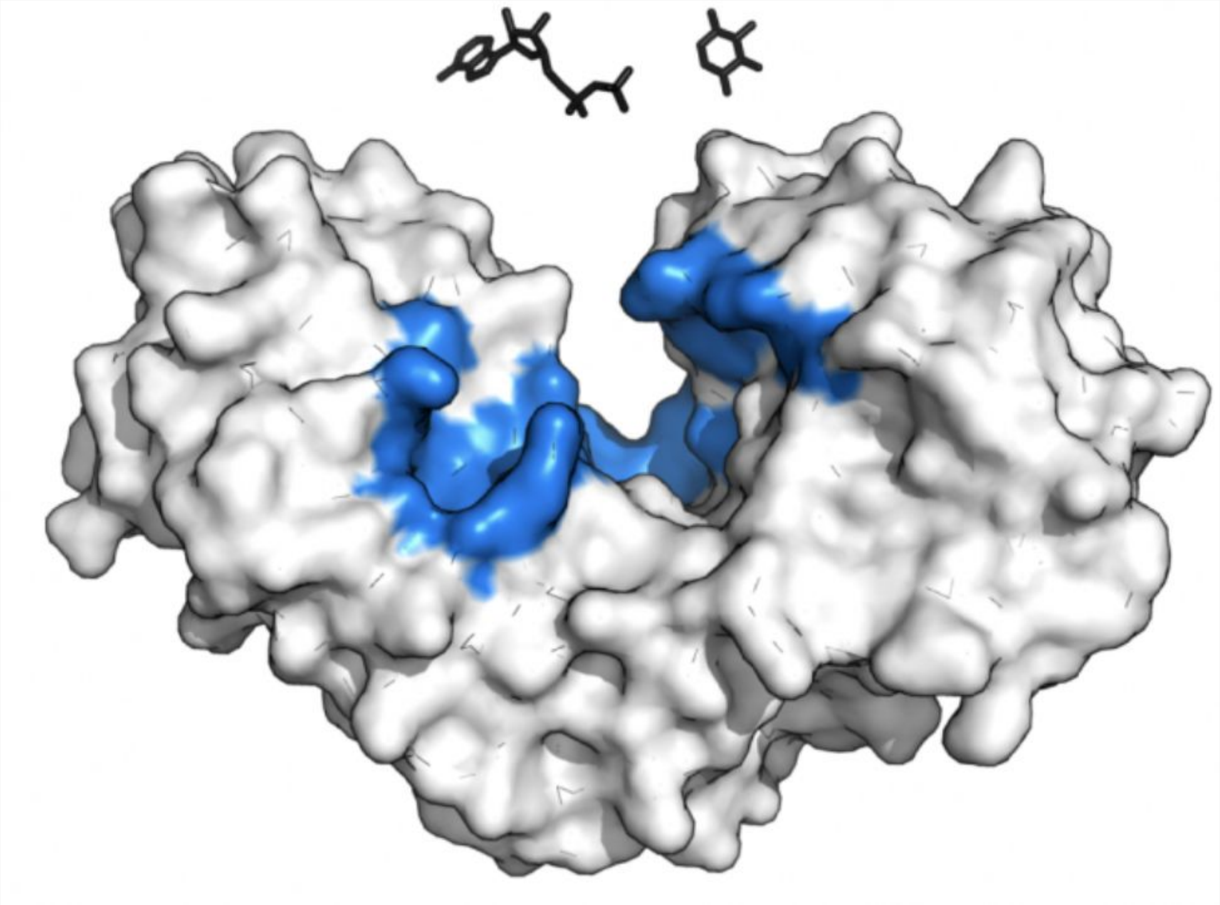
Figure 2: Agreement between attention and contact maps across five pretrained Transformer models from TAPE (a) and ProtTrans (b–e). The heatmaps show the proportion of high-confidence attention weights ($\alpha_{i,j} > \theta$) from each head that connects pairs of amino acids that are in contact with one another. In TapeBert (a), for example, we can see that 45% of attention in head 12-4 (the 12th layer's 4th head) maps to contacts. The bar plots show the maximum value from each layer. Note that the vertical striping in ProtAlbert (b) is likely due to cross-layer parameter sharing (see Appendix A.3).



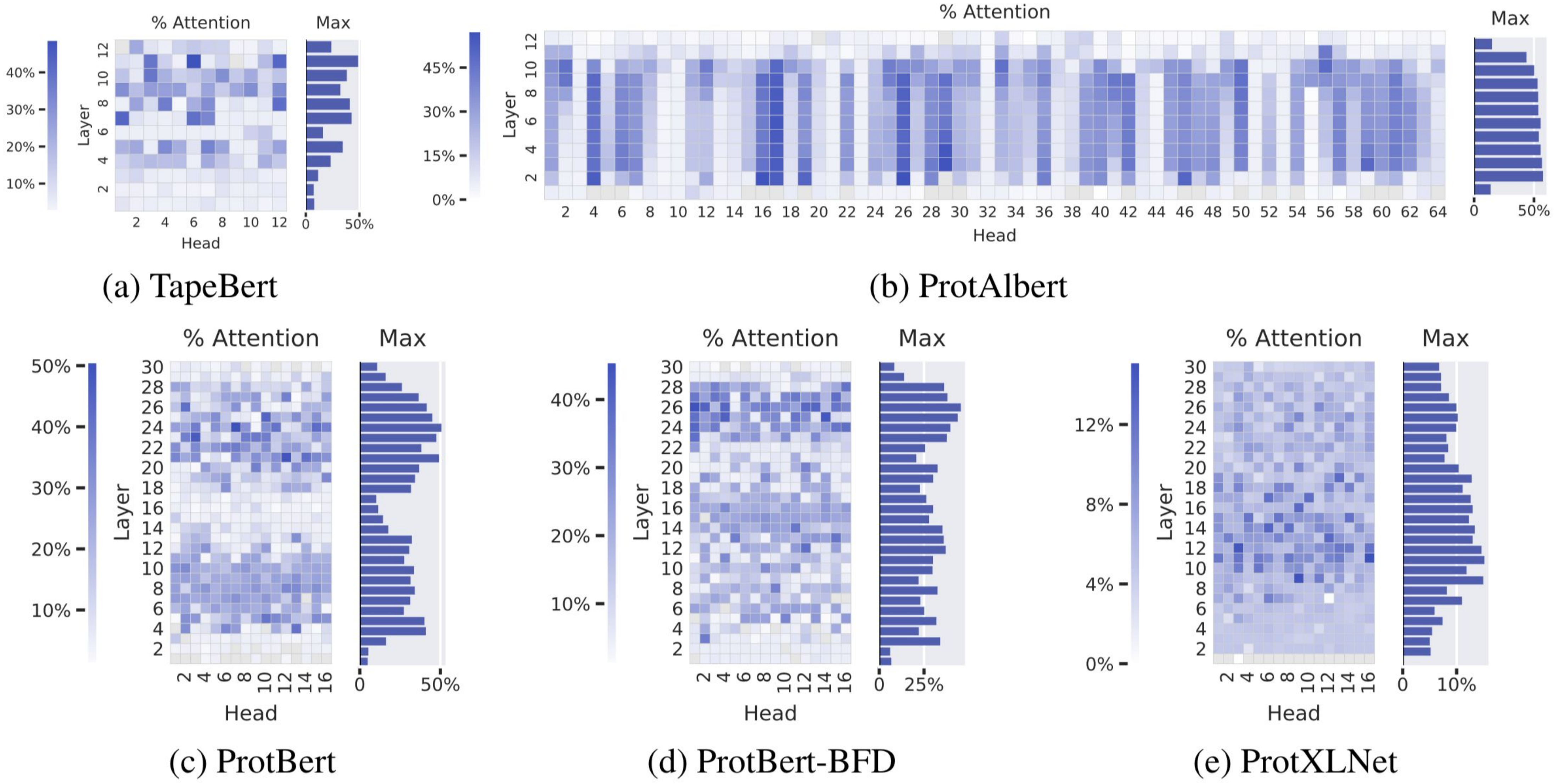
De novo designed TIM-barrel

Inset image shows subsequence

Does attention align with binding sites?



Source: https://teachmephysiology.com/biochemistry/molecules-and-signalling/enzyme-kinetics/



(a) TapeBert

(b) ProtAlbert

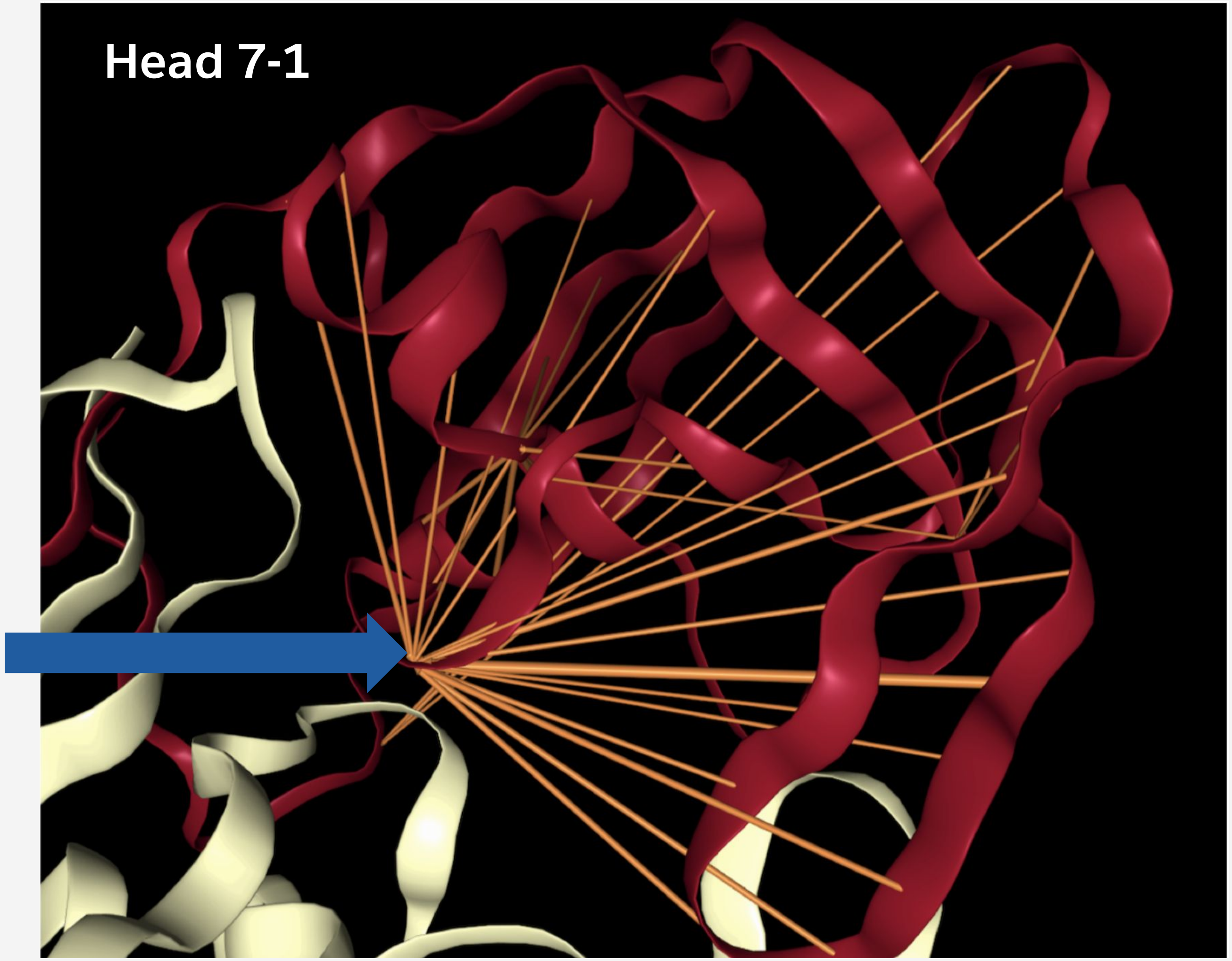
(c) ProtBert

(d) ProtBert-BFD

(e) ProtXLNet

Figure 3: Proportion of attention focused on binding sites across five pretrained models. The heatmaps show the proportion of high-confidence attention ($\alpha_{i,j} > \theta$) from each head that is directed to binding sites. In TapeBert (a), for example, we can see that 49% of attention in head 11-6 (the 11th layer's 6th head) is directed to binding sites. The bar plots show the maximum value from each layer.

Binding Site



HIV-1 protease

Looking forward: Interpretability for scientific discovery

- **NLP** seeks to **automate** capability that humans already have: understanding language
- In contrast, **proteins** are an ongoing subject of **scientific investigation**
- **Interpreting** these models can therefore aid in **scientific discovery**
- Look for **differences** between model's representations and current scientific understanding
- **Contextual visualizations** make representations accessible to domain experts