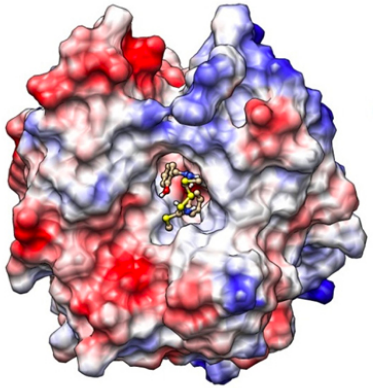
BERTology meets Biology: Interpreting Attention in Protein Language models

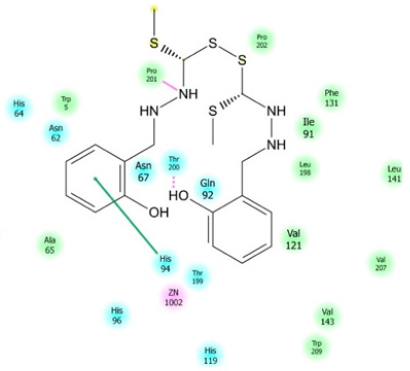
Seongok Ryu
AITRICS, Drug Discovery Team

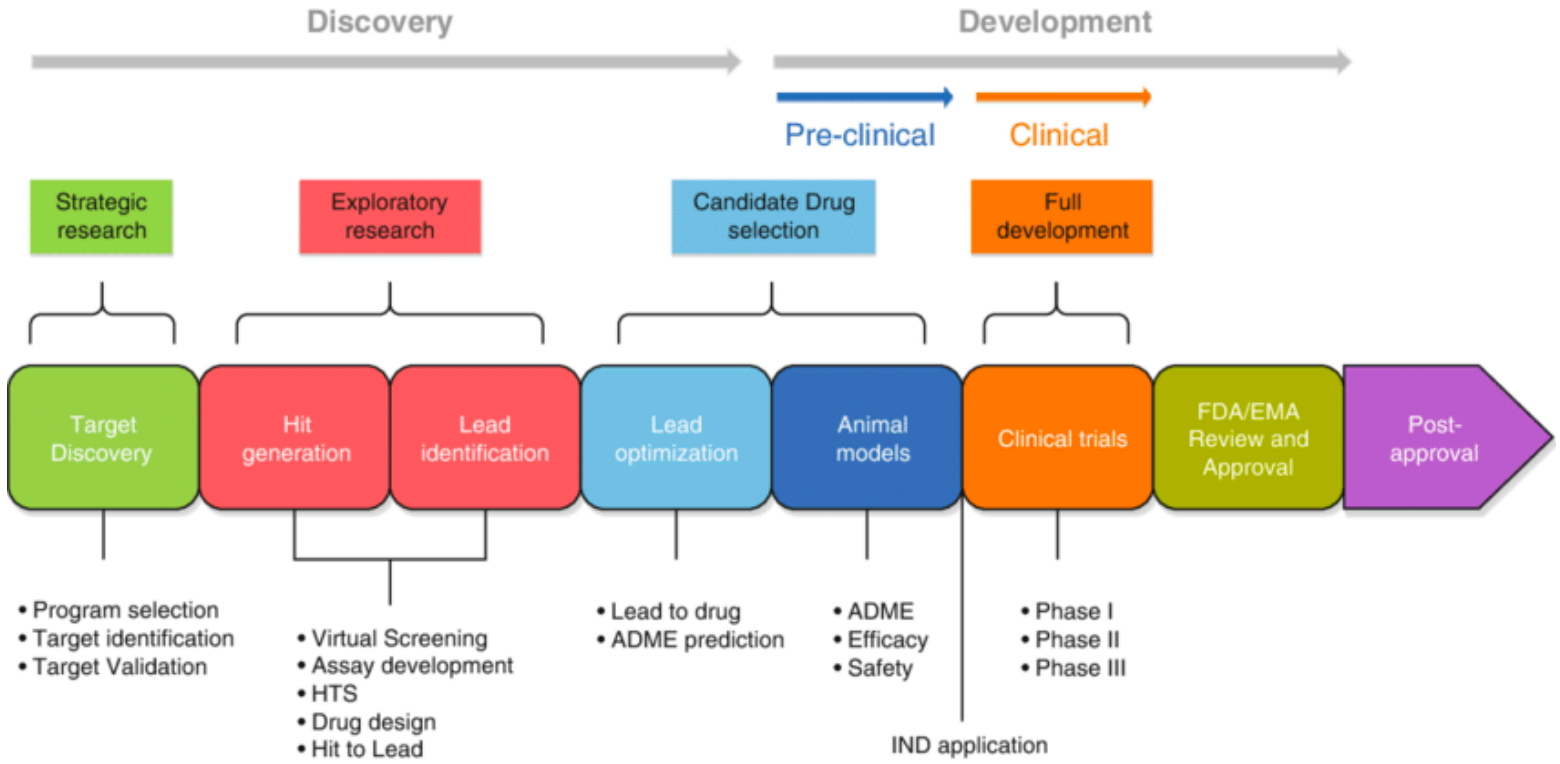
2020. 07. 05.

Preliminaries

Drug discovery process



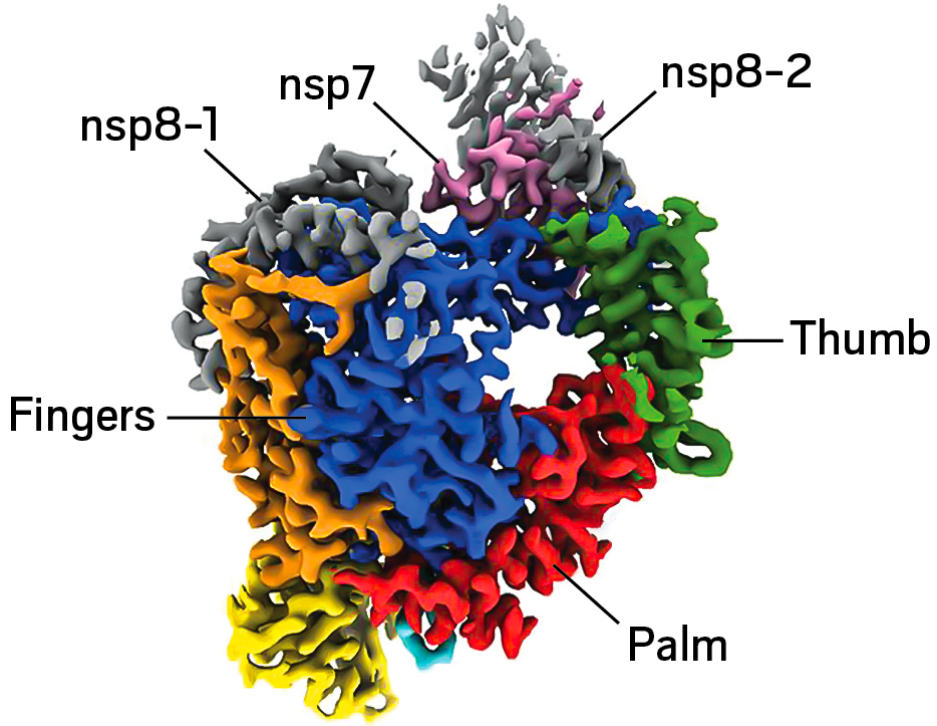




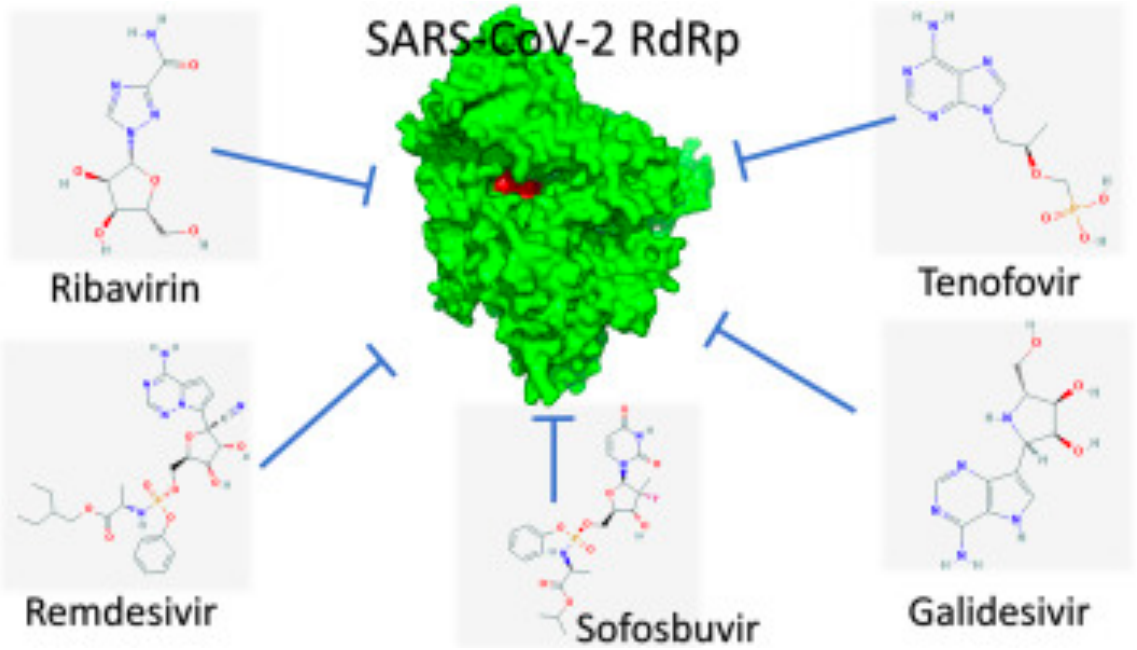
Identifying **what biological targets to regulate** is the first key step in drug discovery.

Preliminaries

Example) COVID-19 and Remdesivir



RNA-dependent RNA polymerase (RDRP)



SARS-COV-2 RDRP inhibitors

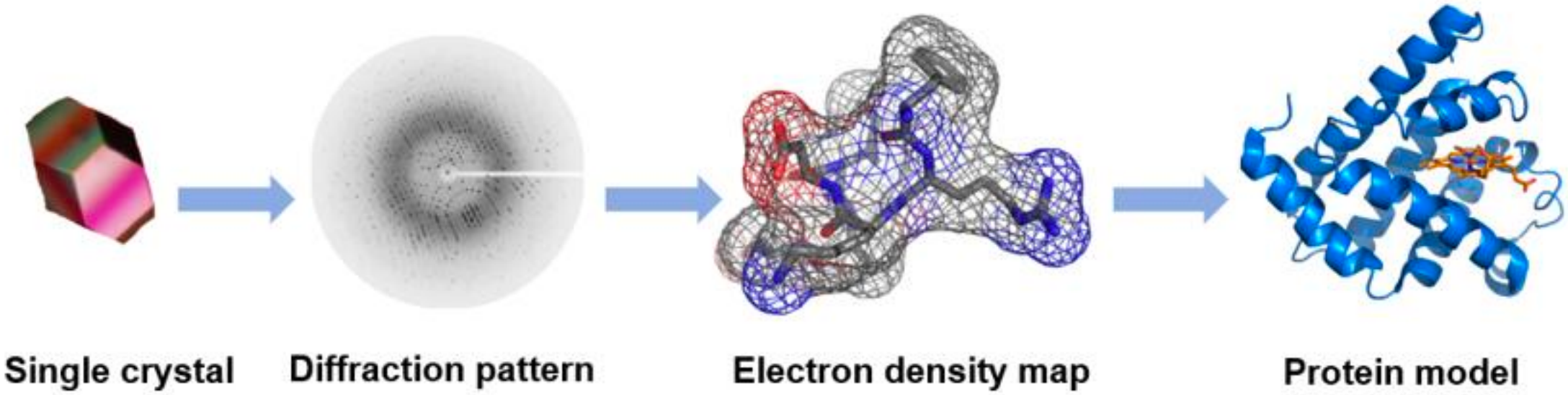
Key question for target identification

- ✓ What biological target (protein) to regulate?
- ✓ Where is the **binding site**? → We have to know **structural information of proteins**.

Preliminaries

However, understanding protein structure is very difficult.

Experiments: X-ray crystallography



- Expensive, Laborious
- Depends on X-ray resolutions
- Usually, crystal structures cannot perfectly reflect the structures *in vivo*.

Preliminaries

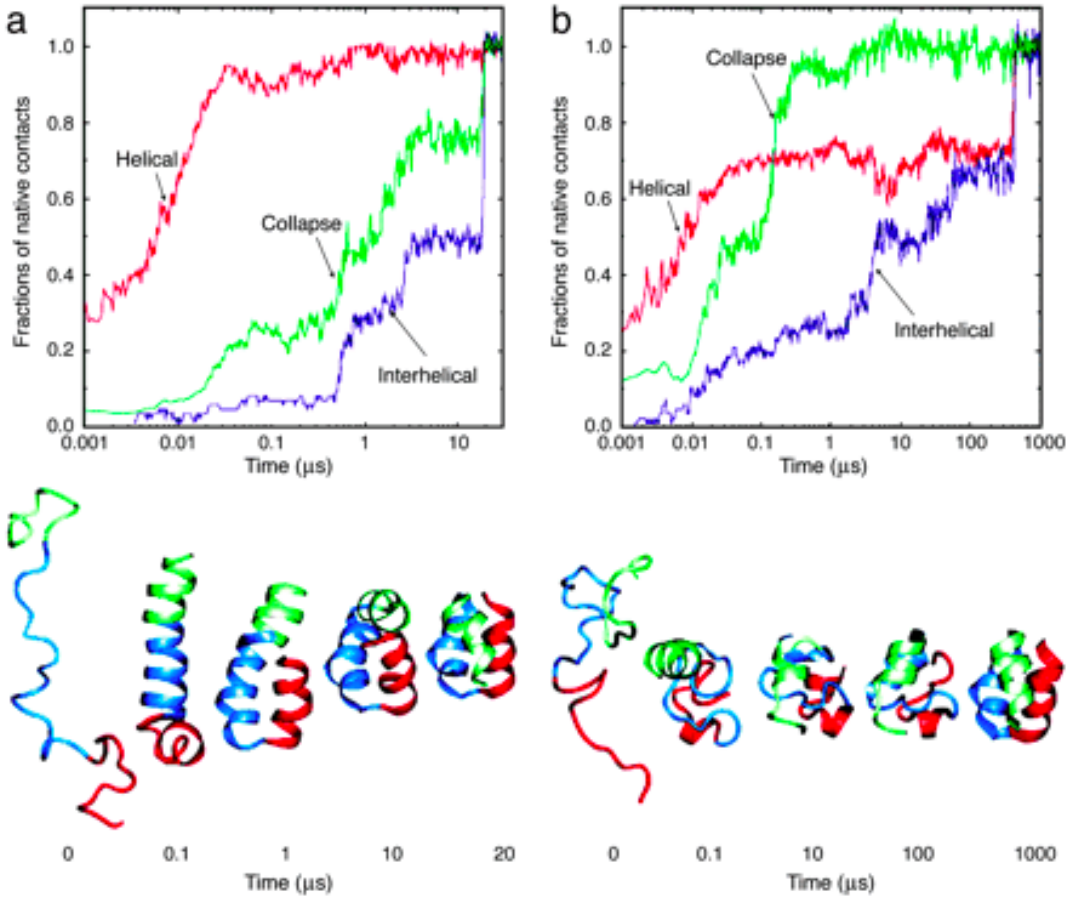
However, understanding protein structure is very difficult.

Computational chemistry (molecular dynamics)

- ✓ Physics-driven solution
- ✓ Cost-expensive:

  ≥ two-weeks for computing single protein structure)
- ✓ Local minima problem:

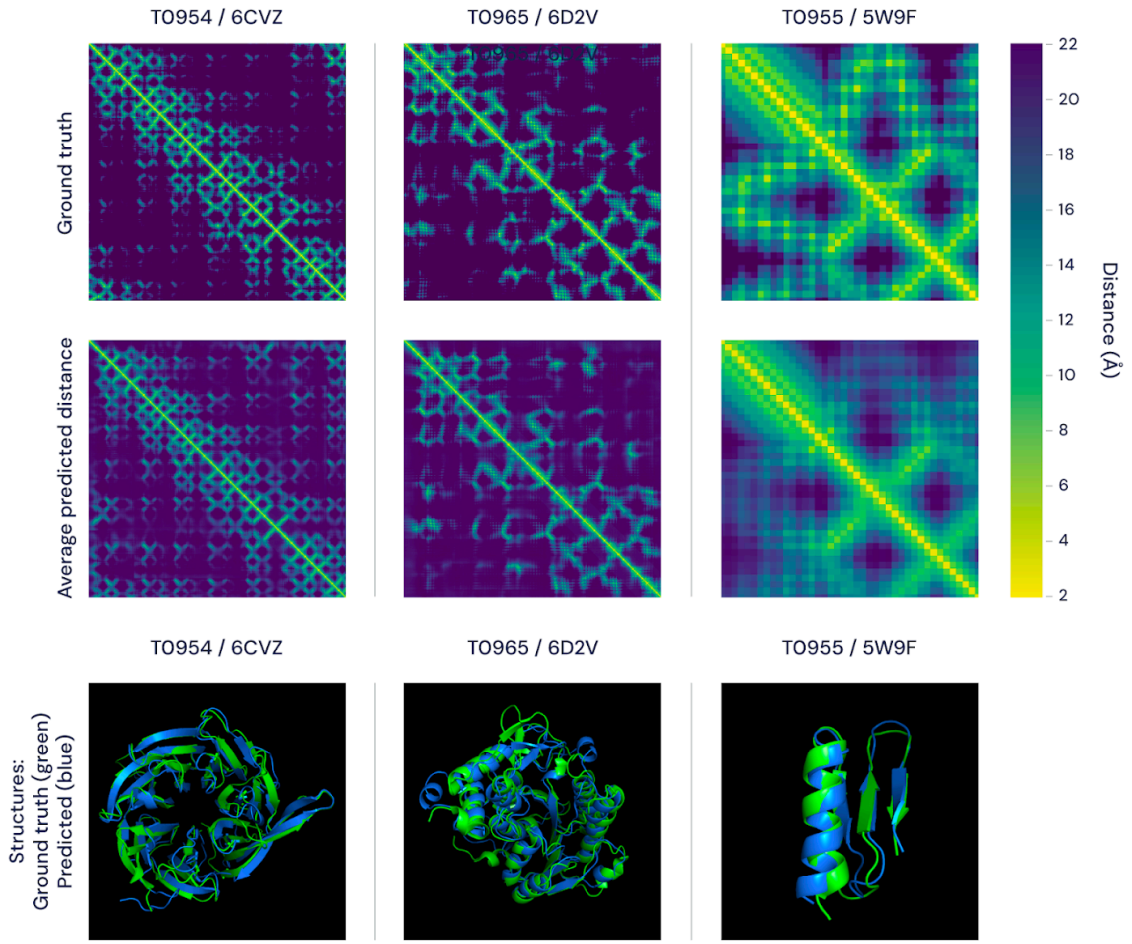
  Since it updates the position and velocity and velocity of atoms by using gradient-descent like algorithm (Verlet algorithm).

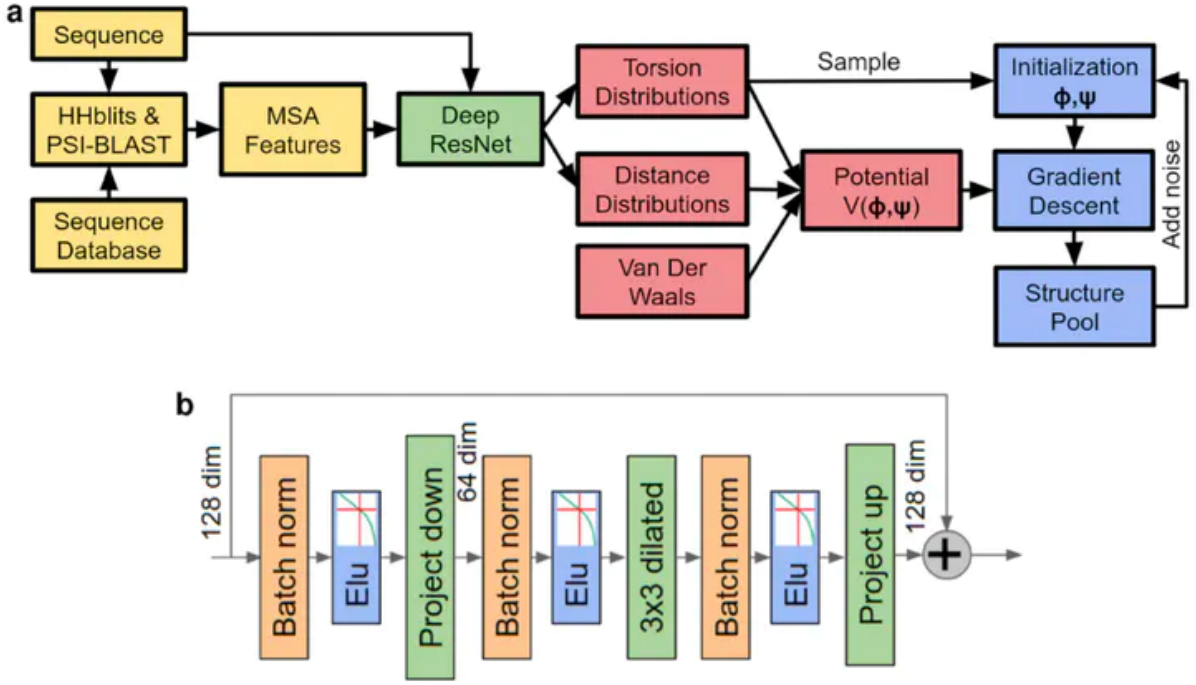


Preliminaries

However, understanding protein structure is very difficult.

AlphaFold: using DNNs for predicting protein structures

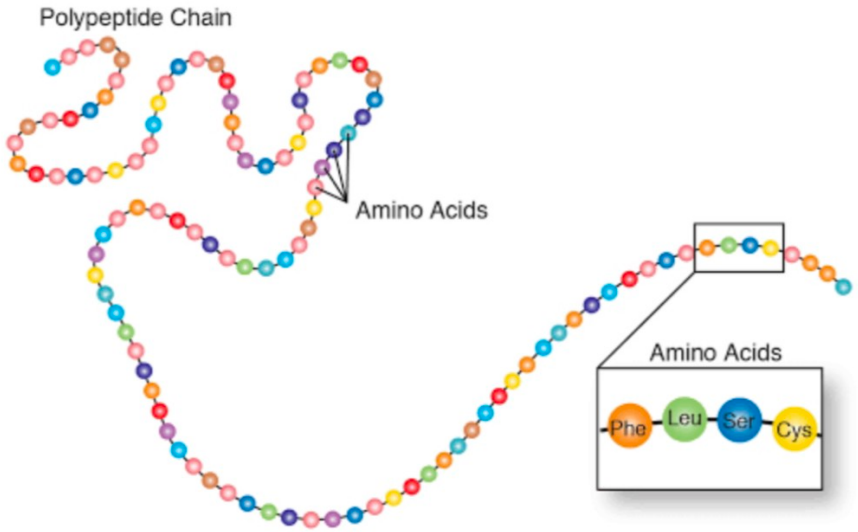




- ✓ Computationally efficient
- ✓ Q) Does it work well for all family of proteins and mutated ones?

BERTology in Biology

What problems have this research addressed?

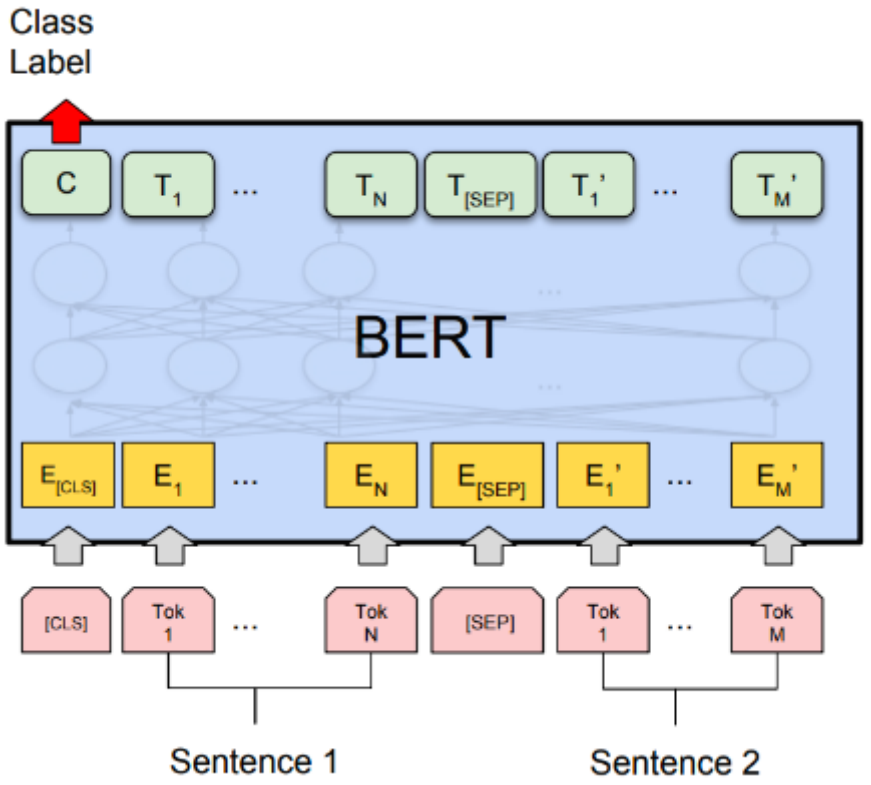


Amino Acids

Ala: Alanine	Gln: Glutamine	Leu: Leucine	Ser: Serine
Arg: Arginine	Glu: Glutamic acid	Lys: Lysine	Thr: Threonine
Asn: Asparagine	Gly: Glycine	Met: Methionine	Trp: Tryptophane
Asp:Aspartic acid	His: Histidine	Phe: Phenylalanine	Tyr: Tyrosisne
Cys:Cysteine	Ile: Isoleucine	Pro: Proline	Val: Valine

Protein can be represented as

the sequence of amino acids.



Using BERT for protein language modeling

BERTology in Biology

Analyzing attention weights in (protein-)BERT

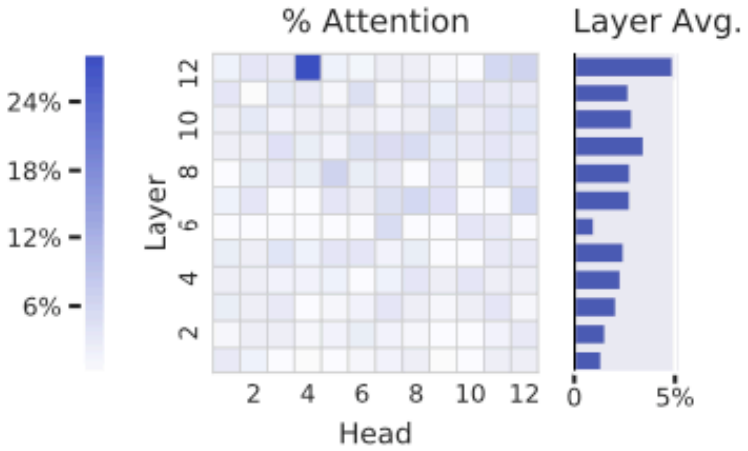


Figure 4: Percentage of each head's attention that is aligned with contact maps, averaged over a dataset, suggesting that Head 12-4 is uniquely specialized for contact prediction.

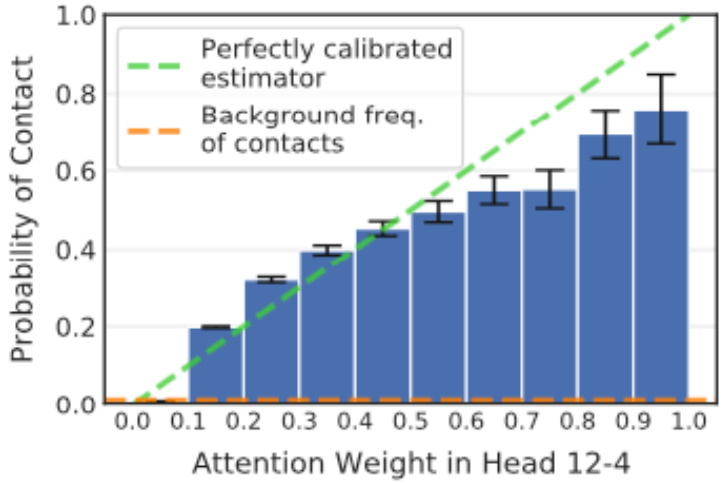


Figure 5: Probability two amino acids are in contact [95% confidence intervals], as a function of attention between the amino acids in Head 12-4, showing attention approximates a perfectly-calibrated estimator (green line).

"Attention captures high-level structural properties of proteins."

→ Will be further discussed in this review.

Method

Modeling

- ✓ **BERT-Base model** from TAPE repository (Evaluating Protein Transfer Learning with TAPE, NIPS2019)
  - 12 layers and 12 attention heads, total 144 distinct attention heads
  - Each attention head generates a distinct set of attention weights α for an input, where $\alpha_{ij} > 0$ represents the attention from token i to token j in the sequence, such that $\sum_j a_{ij} = 1$.
- ✓ Pretrained on **masked language modeling** of amino acids over a dataset of 31 million protein sequence.
- ✓ This model accepts as input a sequence of amino acids $x = (x_1, \ldots, x_L)$ and outputs a sequence of continuous embeddings $z = (z_1, \ldots, z_L)$.

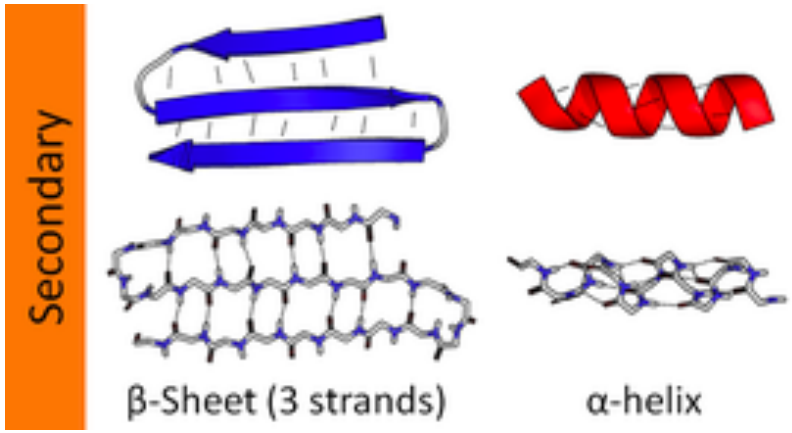
Method

Attention analysis

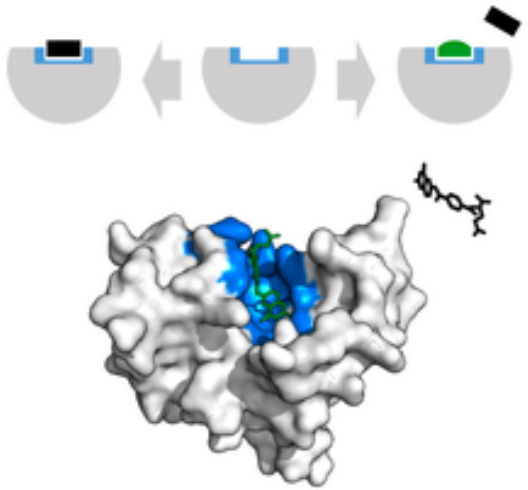
- ✓ How attention aligns with various protein properties, both at
  - the token level: secondary structure, binding sites
  - the token-pair level: contact map
- ✓ Proportion of attention that aligns with property f over a dataset X:

$$p_\alpha(f) = \frac{\sum_{x \in X} \sum_{i=1} \sum_{j=1} f(i,j) \alpha_{ij}(x)}{\sum_{x \in X} \sum_{i=1} \sum_{j=1} \alpha_{ij}(x)}$$

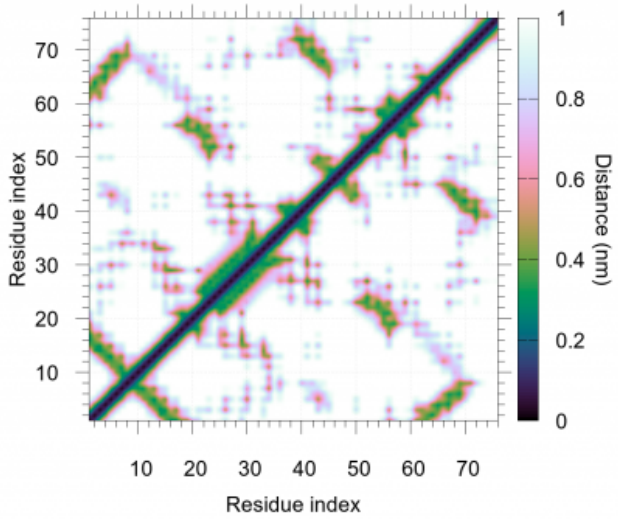
Secondary structure



Binding sites



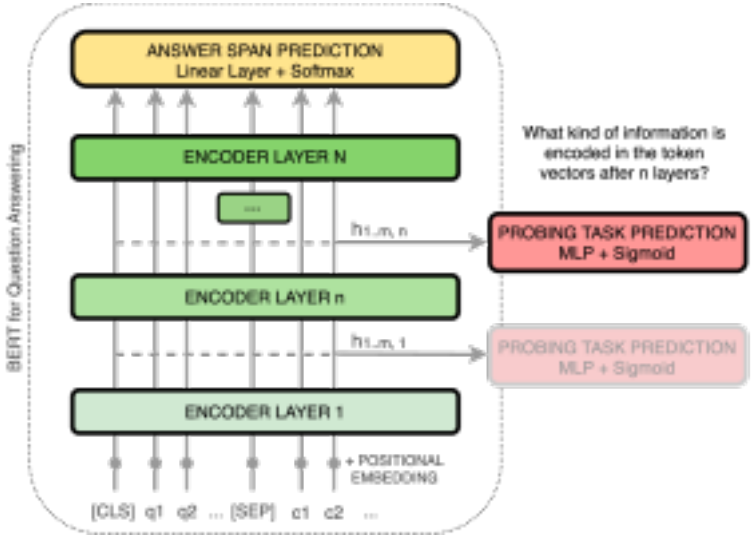
Contact map



Method

Probing tasks

- ✓ A diagnostic classifier to probe the layer outputs to determine what information they contain about properties.
- ✓ A single linear layer followed by softmax, where the weights of the original model were frozen.
- ✓ For token-level probing tasks (binding sites, secondary structures), they fed each token's output vector directly to the classifier.
- ✓ For token-pair probing tasks (contact map), they constructed a pairwise feature vector by concatenating the elementwise differences and products of the two token's output vectors.



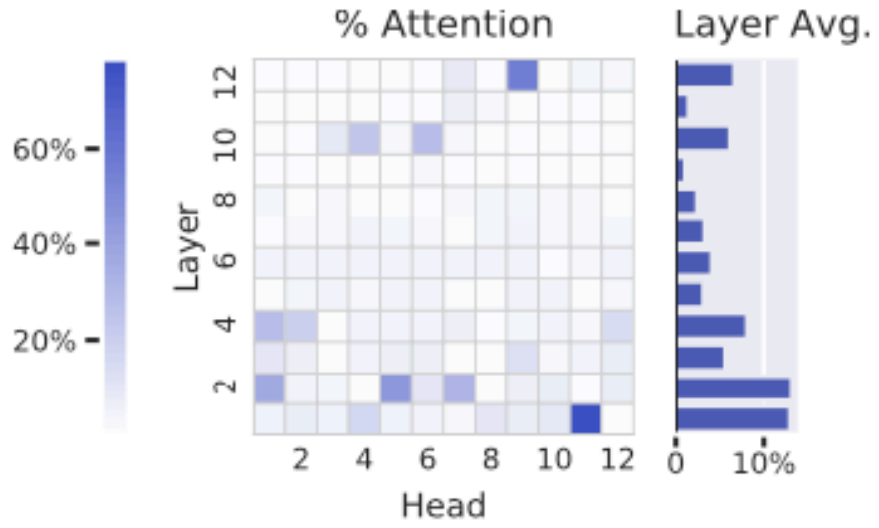
Method

Experimental details

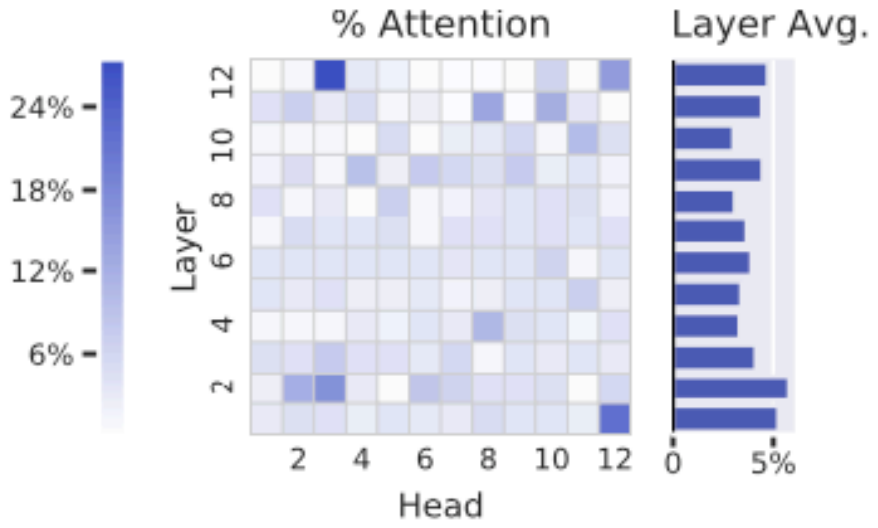
- ✓ Datasets from the TAPE repository:
  - ProteinNet dataset: annotated with spatial coordinates of each AA, used for generating contact maps
  - Secondary structure dataset: annotated with the secondary structure at each sequence position
- ✓ Analysis techniques
  - Excluded attention to [SEP] and [CLS] token (explictly not used in protein LM) in analysis.
  - Filtered attention below a minimum threshold of 0.1 to reduce the effects of very low-confidence attention patterns on the anlysis.
  - Truncated all protein sequences to a maximum length of 512 to reduce the model memory requirements.
- ✓ Single Tesla V-100 GPU with 16GB memory.

Results and Discussion

1. Attention heads specialize in certain types of amino acids



(a) Attention to amino acid *Pro*

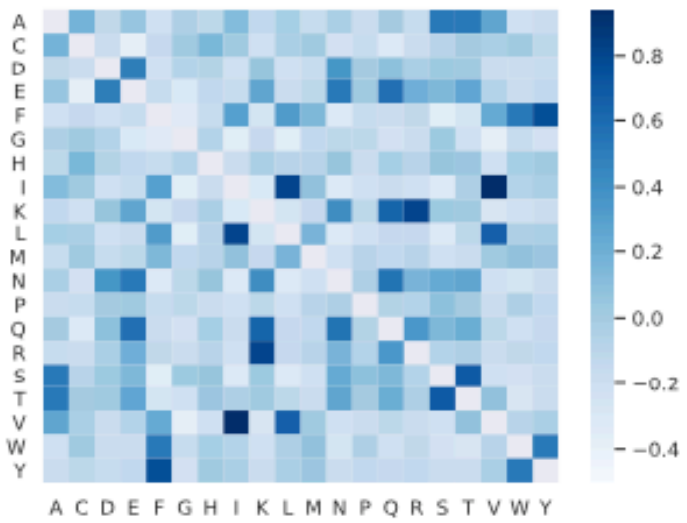


(b) Attention to amino acid *Phe*

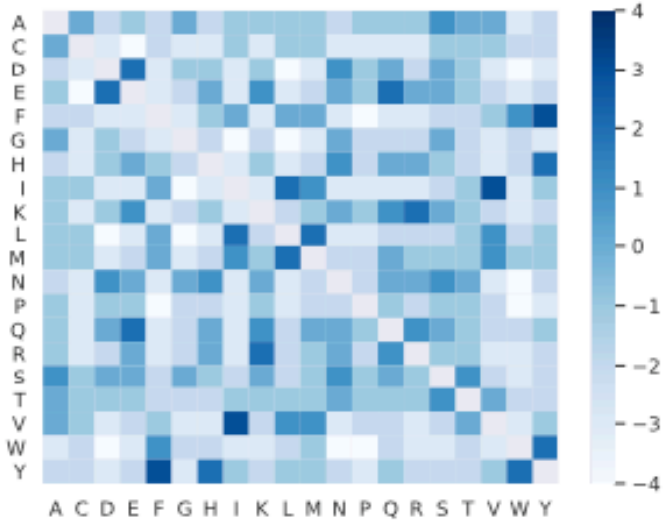
- ✓ For 14 of the 20 types of AAs, there exists an attention head that focuses over 25% of attention on that AA.
  - The head 1-11 focuses 78% of its total attention on “proline (Pro)”.
  - The head 12-3 focuses 27% of its total attention on “phenylalanine (Phe)”.
- ✓ Average results over a dataset of 5,000 sequences with a combined length of 1,067,712 AAs.

Results and Discussion

2. Attention is consistent with substitution relationships



(a) Attention similarity



(b) BLOSUM62 substitution scores

- ✓ Question) Whether each head has “memorized” specific amino acids to target, or whether it has actually learned meaningful properties that correlate with particular amino acids.
- ✓ How the attention received by amino acids relates to an existing measure of structural and functional properties: the substitution matrix.
- ✓ Comparing the attention similarity to the BLOSUM scores → Pearson correlation of 0.80

Results and Discussion

3. Attention aligns strongly with contact maps in one attention head.

- ✓ “Figure 4 shows the percentage of each head’s attention that aligns with contact maps.”
- ✓ A single head “12-4” aligns much more strongly with contact map (28% of attention) than any other heads (maximum 7% of attention).
- ✓ Where the attention weight in head “12-4” is greater than 0.9, the alignment increases to 76%.

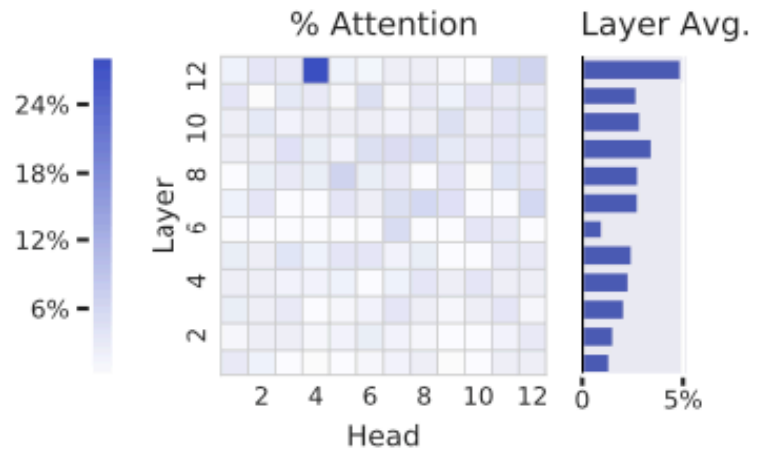
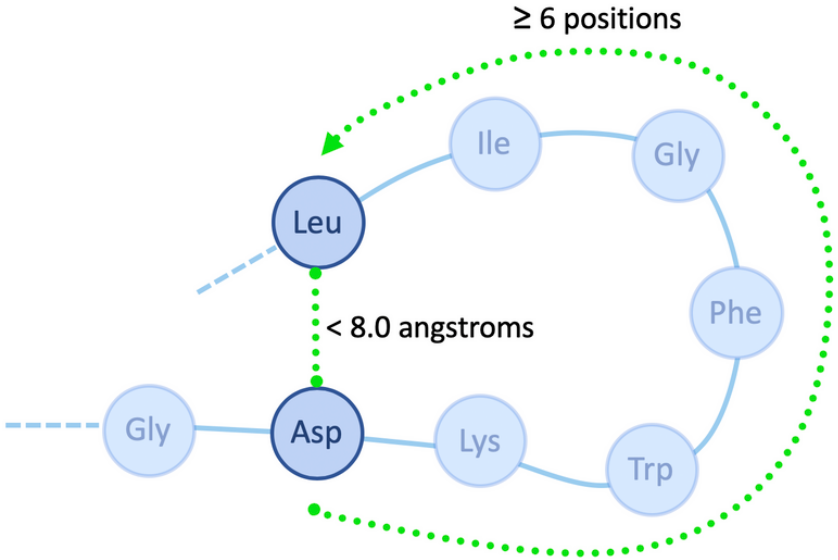


Figure 4: Percentage of each head’s attention that is aligned with contact maps, averaged over a dataset, suggesting that Head 12-4 is uniquely specialized for contact prediction.



AITRICS

Results and Discussion

4. Attention is a well-calibrated predictor of contact maps

- ✓ (In NLP) “Attention weights represent a model’s confidence in detecting certain features.”
- ✓ Comparing attention weight in the head “12-4” with the probability of two amino acids being in contact. (calibration curve in Figure 5.)
- ✓ The Pearson correlation between the estimated probabilities and the attention weights is 0.97.

  → Attention weight is a well-calibrated estimator in this case, providing a principled interpretation of attention as a measure of confidence.

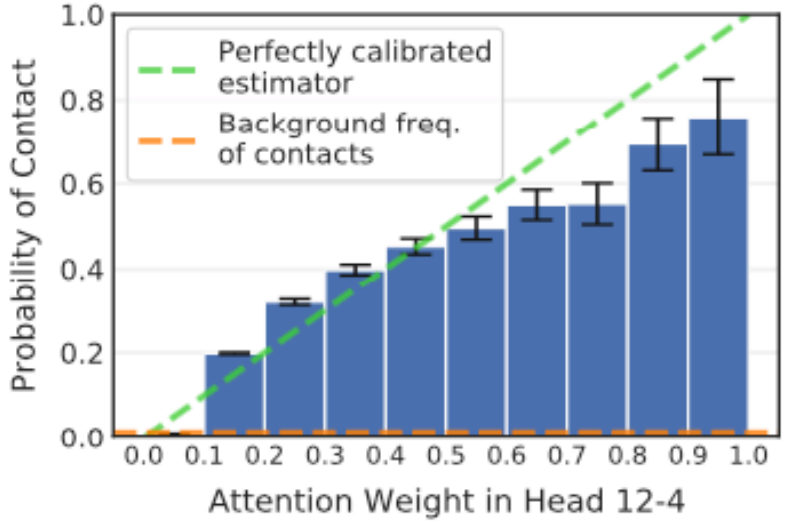
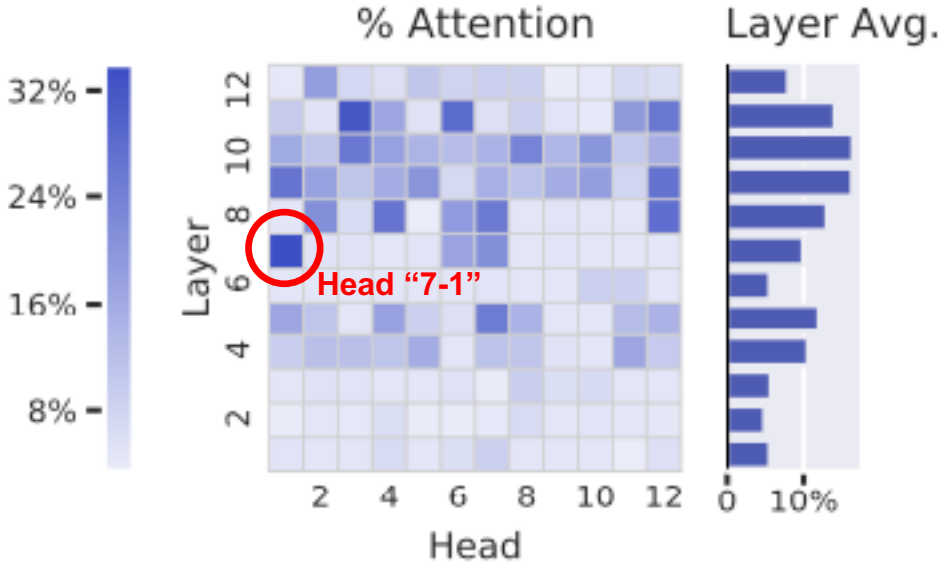


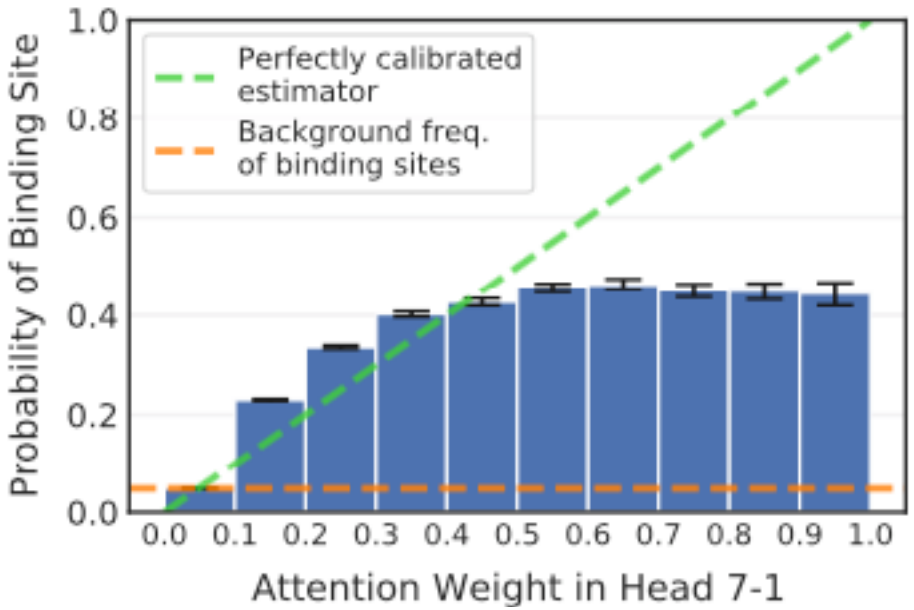
Figure 5: Probability two amino acids are in contact [95% confidence intervals], as a function of attention between the amino acids in Head 12-4, showing attention approximates a perfectly-calibrated estimator (green line).

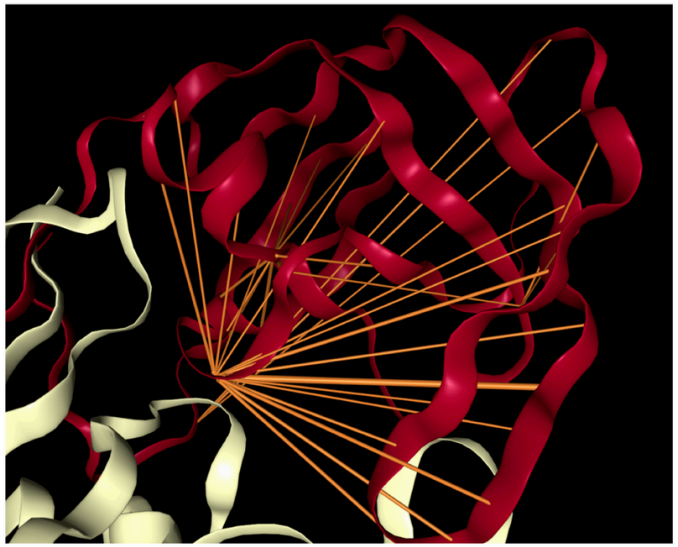
Results and Discussion

5. Attention targets binding sites, especially in the deeper layers

- ✓ Figure 6 shows the proportion of attention focused on binding sites.
- ✓ The effect is strongest in the last 6 layers, which includes 15 heads that each focuses over 20% of their attention on binding sites.
- ✓ The head “7-1” focuses the most attention on binding sites (34%).
- ✓ Figure 7 shows the estimated probability of the head “7-1” targeting a binding site, as a function of the attention weight.







Results and Discussion

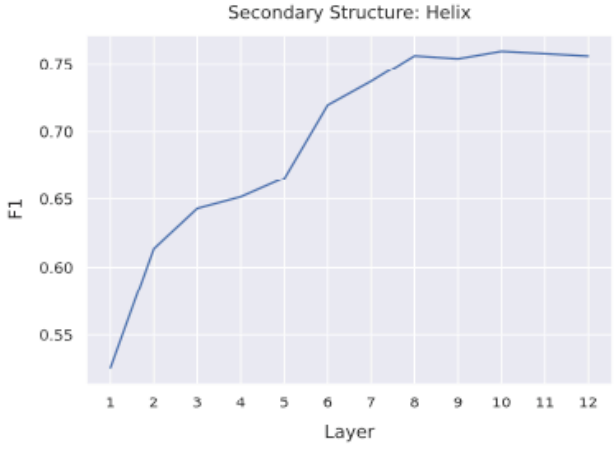
6. Attention targets higher-level properties in deeper layers.

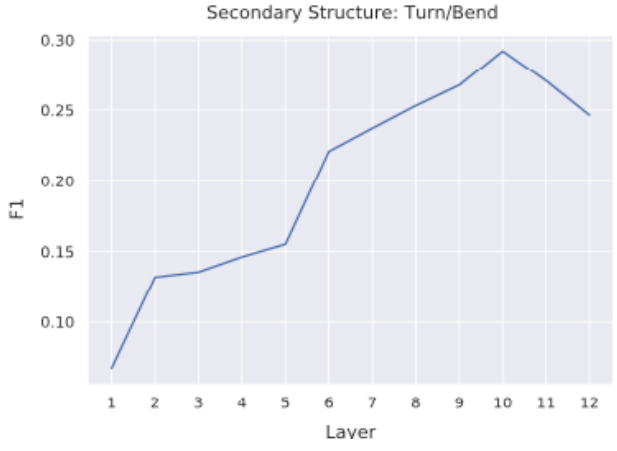
- ✓ **Deeper layers** focus relatively more attention on binding sites and contacts **(high-level contact)**,
- ✓ Whereas secondary structure (**low- to mid-level concept**; Helix, Turn/Bend, Strand) is targeted more **evenly across layers**.
- ✓ Prior works in NLP also suggests that deeper layers in text-based Transformers attend to more complex properties and encode higher-level representations.

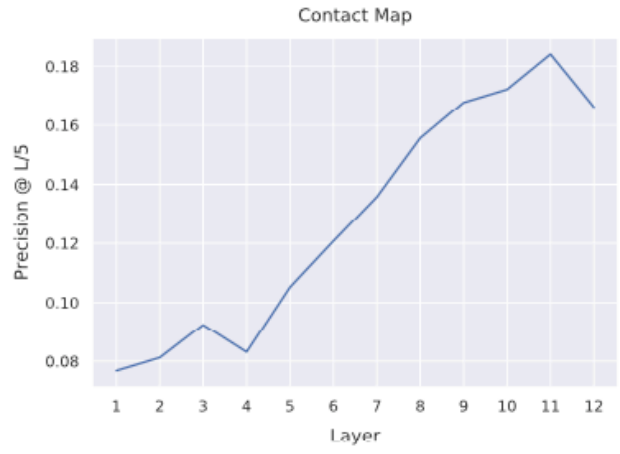
Results and Discussion

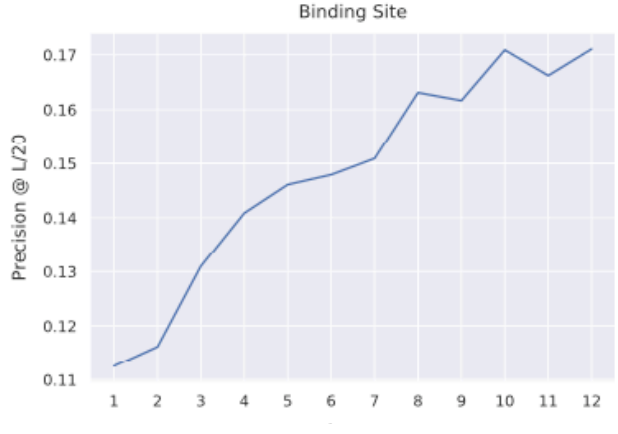
6. Attention targets higher-level properties in deeper layers.

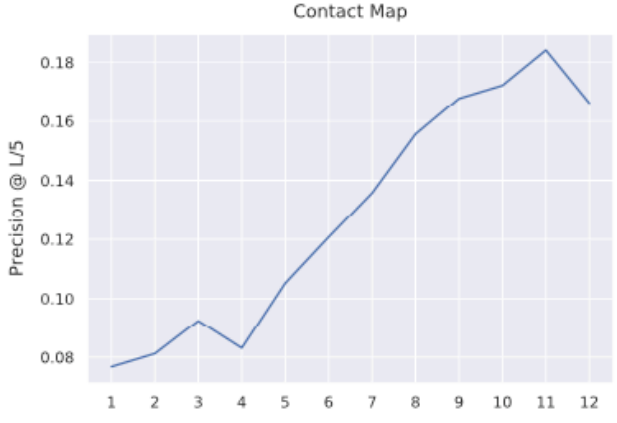
- ✓ The probing analysis similarly shows that the model first forms representations of secondary structure before fully encoding contact maps and binding sites.
- ✓ The model must understand local structure before it can form representations of higher-order structure and function.







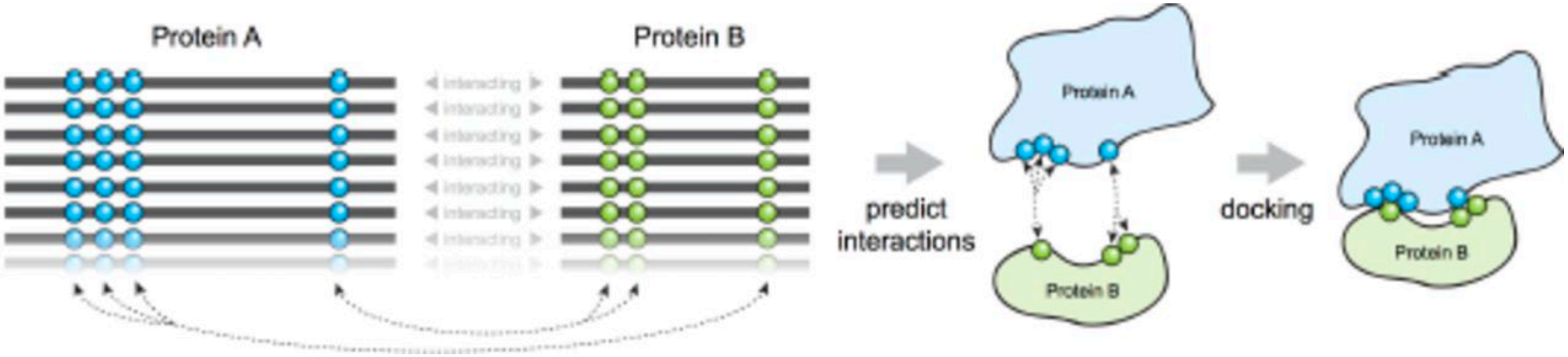




Discussion

Why does attention analysis seem reasonable?

- ✓ Evolutionary pressures have naturally selected proteins among the combinatorial space of possible AA-seqs.
- ✓ Proteins largely function to bind to other molecules – small molecules, proteins, or other macro-molecules.
- ✓ Past work has shown that binding sites can reveal evolutionary relationships among proteins and that particular structural motifs in binding sites are mainly restricted to specific families/superfamilies of proteins.



Co-evolution of two proteins

Thank You!