{'title': 'BERT OLOGY MEETS BIOLOGY : INTERPRETING ATTENTION IN PROTEIN LANGUAGE MODELS exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models',

'authors': [{'name': 'Jesse Vig', 'affiliation': 'Salesforce Research'},

{'name': 'Ali Madani', 'affiliation': 'Salesforce Research'},

{'name': 'Lav R. Varshney',

'affiliation': 'Salesforce Research, University of Illinois at Urbana-Champaign'},

{'name': 'Caiming Xiong', 'affiliation': 'Salesforce Research'},

{'name': 'Richard Socher', 'affiliation': 'Salesforce Research'},

{'name': 'Nazneen Fatema Rajani', 'affiliation': 'Salesforce Research'},

{'name': 'Benjamin Hoover', 'affiliation': ''},

{'name': 'Hendrik Strobelt', 'affiliation': ''},

{'name': 'Sebastian Gehrmann', 'affiliation': ''}],

'abstract': 'Transformer architectures have proven to learn useful representations for protein classification and generation tasks. However, these representations present challenges in interpretability. In this work, we demonstrate a set of methods for analyzing protein Transformer models through the lens of attention. We show that attention: (1) captures the folding structure of proteins, connecting amino acids that are far apart in the underlying sequence, but spatially close in the three-dimensional structure, (2) targets binding sites, a key functional component of proteins, and (3) focuses on progressively more complex biophysical properties with increasing layer depth. We find this behavior to be consistent across three Transformer architectures (BERT, ALBERT, XLNet) and two distinct protein datasets.',

'key\_findings': "Our analysis reveals that attention captures high-level structural properties of proteins, connecting amino acids that are spatially close in three-dimensional structure, but apart in the underlying sequence. Attention also targets binding sites, a key functional component of proteins. Further, attention is consistent with a classic measure of similarity between amino acids—the substitution matrix. Finally, attention captures progressively higher-level representations of structure and function with increasing layer depth. Examples of how specialized attention heads in a Transformer recover protein structure and function, based solely on language model pre-training. Attention in head 12-4 targets amino acid pairs that are close in physical space but lie apart in the sequence, exemplified by a de novo designed TIM-barrel. Attention in head 7-1 targets binding sites, crucial for protein function, with an example of HIV-1 protease where the primary location receiving attention is a known binding site for protease inhibitor drugs. Attention aligns strongly with contact maps in the deepest layers of pretrained Transformer models for protein sequence analysis. The most aligned heads are found in the deepest layers, focusing up to 44.7% (TapeBert), 55.7% (ProtAlbert), 58.5% (ProtBert), 63.2% (ProtBert-BFD), and 44.5% (ProtXLNet) of attention on contacts, compared to a background frequency of 1.3% among all amino acid pairs. The study investigates the attention mechanisms of five pretrained models (TapeBert, ProtAlbert, ProtBert, ProtBert-BFD, and ProtXLNet) towards protein binding sites. It was found that these models, despite being trained on language modeling tasks without explicit spatial information, exhibit structurally-aware attention patterns. ProtAlbert showed the highest attention to binding sites, with 22 heads focusing over 50% of their attention on these sites. The attention to binding sites across models suggests that the models may have learned to recognize biochemical interactions and statistical dependencies between amino acids, which could be valuable for understanding protein functions and interactions. The analysis reveals that deeper layers of the model focus more on high-level concepts such as binding sites and contacts, indicating a hierarchical learning structure where higher-level properties are captured in deeper layers. Additionally, a small number of heads are found to concentrate attention on amino acids associated with post-translational modifications (PTMs), highlighting their importance to protein function despite their low occurrence in the sequence. Attention heads specialize in particular amino acids, with significant attention focused on specific amino acids beyond their background frequencies. Amino acids with similar structural and functional properties are attended to similarly across heads, as evidenced by a Pearson correlation of 0.73 between attention distribution and BLOSUM62 substitution scores. Similar correlations are observed for the ProtTrans models with correlations of 0.68 (ProtBert), 0.75 (ProtBert-BFD), 0.60 (ProtAlbert), and 0.71 (ProtXLNet). The randomized versions of these models yielded significantly lower correlations, indicating the effectiveness of the ProtTrans models in capturing substitution relationships in proteins. Our work extends these methods to protein sequence models by considering particular biophysical properties and relationships. We also present a joint cross-layer probing analysis of attention weights and layer embeddings. This paper demonstrates how a Transformer language model can recover structural and functional properties of proteins and integrate this knowledge directly into its attention mechanism. It also presents a novel tool for visualizing attention in the context of protein structure. The figures illustrate the percentage of each head's attention that is focused on Strand and Turn/Bend secondary structures across different models: TapeBert, ProtAlbert, ProtBert, ProtBert-BFD, and ProtXLNet. For Strand secondary structure, ProtBert-BFD shows a higher percentage of attention, especially in the middle layers, indicating a strong focus on this feature. In contrast, for Turn/Bend secondary structure, attention distribution is more varied across models, with ProtBert-BFD again showing significant attention in certain layers. The differences between the attention proportions and the background frequency of contacts are statistically significant (p< 0.00001). Bonferroni correction applied for both confidence intervals and tests. Differences between attention proportions and the background frequency of binding sites are all statistically significant (p< 0.00001). Bonferroni correction applied for both confidence intervals and tests. The differences between the attention proportions and the background frequency of PTMs are statistically significant (p< 0.00001). Bonferroni correction applied for both confidence intervals and tests.",

'limitation\_of\_sota': 'The strong performance of the Transformer comes at the cost of interpretability, and this lack of transparency can hide underlying problems such as model bias and spurious correlations. Existing protein language models have focused primarily on autoregressive or autoencoding self-supervision objectives for discriminative and generative tasks without a strong emphasis on interpretability. Depending on the task and model architecture, attention may have less or more explanatory power for model predictions.',

'proposed\_solution': 'We analyze Transformer protein models through the lens of attention, and present a set of interpretability methods that capture the unique functional and structural characteristics of proteins. We also compare the knowledge encoded in attention weights to that captured by hidden-state representations. Finally, we present a visualization of attention contextualized within three-dimensional protein structure. The paper discusses the use of Transformer models, pre-trained as language models, to recover protein structure and function. It highlights the ability of specific attention heads within these models to identify amino acid pairs that are physically close or constitute binding sites, crucial for understanding protein functionality. This approach is exemplified through the analysis of a de novo designed TIM-barrel and the HIV-1 protease. The paper demonstrates interpretability methods on five Transformer models that were pretrained through language modeling of amino acid sequences, focusing primarily on the BERT-Base model from TAPE, pretrained on Pfam, and four pre-trained Transformer models from ProtTrans. The methodology includes analyzing how attention aligns with various protein properties, performing probing tasks to test the knowledge contained in model representations, and using curated datasets for analysis. The study explores the relationship between attention mechanisms in Transformer models and the tertiary structure of proteins, as characterized by contact maps. It includes an analysis across five pretrained models, highlighting how attention in the deepest layers aligns with protein contact maps. The study performs a fine-grained analysis of the interaction between attention mechanisms in Transformer-based models and specific amino acids, examining how attention heads specialize and the correlation of attention with amino acid substitution relationships. This work takes an interpretability-first perspective to focus on the internal model representations, specifically attention and intermediate hidden states, across multiple protein language models. It also explores novel biological properties including binding sites and post-translational modifications. Extending visualization and analysis methods to protein sequence models, considering biophysical properties and relationships, and presenting a joint cross-layer probing analysis of attention weights and layer embeddings using a single, unified metric. The paper adapts and extends NLP interpretability methods to protein sequence modeling, focusing on reconciling attention with known properties of proteins and leveraging attention to uncover novel relationships or more nuanced forms of existing measures such as contact maps. The text discusses a method for probing the relationship between attention weights and contact between amino acids in protein structures. By treating the attention weight as a feature of a token-pair (i,j), a probing classifier is trained to predict if amino acids i and j are in contact. In the context of analyzing protein sequences, the study introduces a method to quantify the knowledge embedded within the attention mechanism of multi-head attention models. This is achieved by treating the attention weights across all heads in a given layer as a feature vector and employing a probing classifier to assess the understanding of specific properties (e.g., amino acid contacts, secondary structure) within the attention weights across the entire layer. The performance of the probing classifier is measured using precision at a fraction of the protein sequence length, adhering to standard practices in contact prediction.',

'paper\_limitations': 'The paper does not explicitly mention its limitations within the provided excerpt. The analyses are purely associative and do not attempt to establish a causal link between attention and model behavior, nor to explain model predictions. The analysis reveals limitations in common layerwise probing approaches that only consider embeddings, which may not fully capture how knowledge is operationalized in the model. The analysis was limited to a purely evaluative approach using a random subset of 5000 sequences from the training split of each dataset for attention analysis. Additionally, the study excluded sequences for which annotations were not available, potentially limiting the diversity and representativeness of the datasets used in the analysis.'}