(Re)Discovering Protein Structure and Function Through Language Modeling

Jesse Vig[1] **Ali Madani**[1] **Lav R. Varshney**[1] **Nazneen Fatema Rajani**[1]

Abstract

We show how a Transformer language model, trained simply to predict a masked amino acid in a protein sequence, recovers fundamental structural and functional properties of proteins through its attention mechanism. Specifically, we demonstrate that attention 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. We also show that attention targets binding sites, a key functional component of proteins, and we present a three-dimensional visualization of the interaction between attention and protein structure. Our findings align with biological processes and provide a tool to aid discovery in protein engineering and synthetic biology. The code for visualization and analysis is available at https://github.com/salesforce/provis.

1. Introduction

The study of proteins, the fundamental macromolecules governing biology and life itself, has led to remarkable advances in understanding human health and the development of disease therapies. Protein science, and especially protein engineering, has historically been driven by experimental, wet-lab methodologies along with biophysical, structure-based intuitions and computational techniques (Rosenfeld et al., 2016; Arnold, 1998; Huang et al., 2016). The decreasing cost of sequencing technology has enabled us to collect vast databases of naturally occurring proteins (El-Gebali et al., 2019a), which are rich in information for developing powerful sequence-based AI approaches.

Proteins, as a sequence of amino acids, can be viewed precisely as a language. As such, they may be modeled using neural architectures that have been developed for natural language processing (NLP). In particular, the Transformer (Vaswani et al., 2017), which has led to a revolution in unsupervised learning for text, shows promise for a similar impact on protein sequence modeling. However, the strong performance of the Transformer comes at the cost of interpretability, and much NLP research now focuses on interpreting Transformer models such as BERT (Rogers et al., 2020; Devlin et al., 2019).

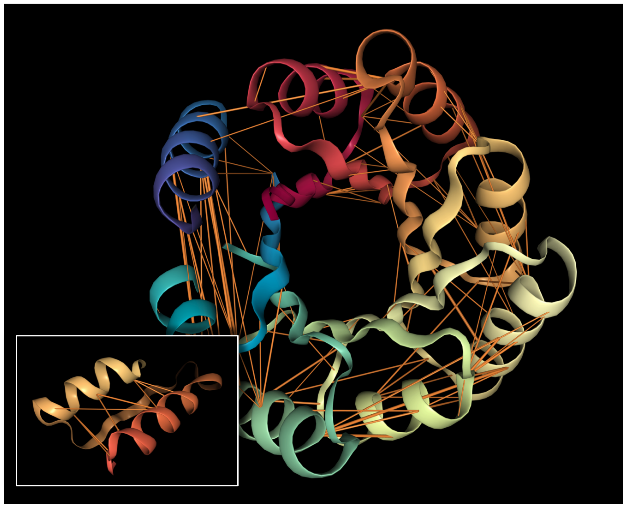

Figure 1. Head 12-4 has learned to focus attention (indicated by orange lines) between amino acids that are spatially close in the folded protein structure but lie apart in the sequence, based solely on language model pre-training. Example is a *de novo* designed TIM-barrel. 76% of high-confidence (> 0.9) attention from this head aligns with ground-truth contact maps on average over a dataset. Visualizations based on the NGL Viewer (Rose et al., 2018; Rose & Hildebrand, 2015).

In this work, we adapt and extend this line of interpretability research to protein sequences. We show how a Transformer model, trained solely to predict a masked amino acid in a sequence, uncovers latent structural and functional properties of proteins through its attention mechanism. In contrast to NLP, which seeks to automate a capability that humans already possess—understanding natural language—protein modeling also seeks to shed light on biological processes that are not yet fully understood. Therefore we also discuss how interpretability can aid scientific discovery. Additional results from this analysis may be found in (Vig et al., 2020b).

[1] Salesforce Research. Correspondence to: Jesse Vig <jvig@salesforce.com>.

ICML Workshop on ML Interpretability for Scientific Discovery, 2020. Copyright 2020 by the author(s).

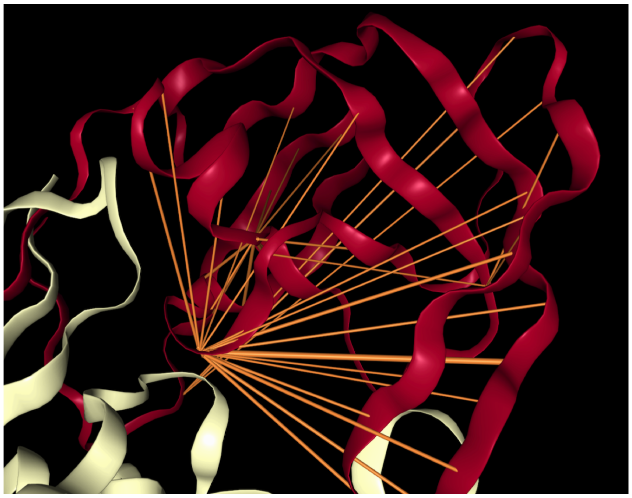

Figure 2. Head 7-1 has learned to focus attention (indicated by orange lines) on binding sites, a key functional component of proteins, based solely on language model pre-training. Example is HIV-1 protease (7HVP). The primary location receiving attention is 27G, a binding site for protease inhibitor small-molecule drugs. 44% of high-confidence (> 0.9) attention from this head focuses on binding sites on average over a dataset.

2. Related Work

Protein language models From sequences alone, Asgari & Mofrad (2015) trained one of the first deep representations of proteins for use as an embedding for downstream tasks. Sequence-only language models have since been trained through autoregressive or autoencoding self-supervision objectives. As examples, Alley et al. (2019); Bepler & Berger (2019); Rao et al. (2019) trained LSTM and transformer-based models to learn representations for protein classification. Rives et al. (2019) showed that the output embeddings from a pretrained Transformer model could be transformed to predict structural and functional properties of proteins, but attention weights were not explored in this analysis. TAPE created a benchmark of tasks to assess protein representation learning models. Riesselman et al. (2019); Madani et al. (2020) trained autoregressive generative models to predict the functional effect of mutations and generate natural-like proteins. Transformer models have also been adapted to incorporate structural information (Ingraham et al., 2019).

Interpreting Transformers. Transformers (Vaswani et al., 2017) are the backbone of state-of-the-art pretrained language models in NLP including BERT (Devlin et al., 2019). BERTology focuses on interpreting what the BERT model learns about natural language by using a suite of probes and interventions (Rogers et al., 2020). So-called *diagnostic classifiers* are used to interpret the outputs from BERT's layers (Veldhoen et al., 2016).

Approaches for interpreting BERT can be categorized into three main categories: interpreting the learned embeddings (Ethayarajh, 2019; Wiedemann et al., 2019; Mickus et al., 2019; Adi et al., 2016; Conneau et al., 2018), BERT's learned knowledge of syntax (Lin et al., 2019; Liu et al., 2019; Tenney et al., 2019; Htut et al., 2019; Hewitt & Manning, 2019; Goldberg, 2019), and BERT's learned knowledge of semantics (Tenney et al., 2019; Ettinger, 2020).

Interpreting attention Interpreting attention in natural language is an active area of research (Wiegreffe & Pinter, 2019; Zhong et al., 2019; Brunner et al., 2020; Clark et al., 2019; Vig & Belinkov, 2019; Htut et al., 2019). Depending on the task, attention may have more or less explanatory power for model predictions (Jain & Wallace, 2019; Serrano & Smith, 2019; Pruthi et al., 2020; Moradi et al., 2019; Vashishth et al., 2019). Visualization techniques have been used to analyze attention in Transformers (Hoover et al., 2019; Kovaleva et al., 2019; Vig, 2019). Recent work has begun to apply attention to guide mapping of sequence models outside of the domain of natural language (Schwaller et al., 2020).

3. Methodology

Model. We study a BERT Transformer model from the TAPE repository that was pretrained on masked language modeling (predicting a masked amino acid) over a dataset of 31 million protein sequences (El-Gebali et al., 2019b). The architecture comprises a series of encoder layers, each of which includes multiple attention heads. Each head generates a distinct set of attention weights α for an input, where $\alpha_{i,j} > 0$ is the attention from token i to token j in the sequence, with $\sum_j \alpha_{i,j} = 1$. Intuitively, $\alpha_{i,j}$ represents the importance that token i assigns to token j when forming its representation for the next layer. The BERT-Base model has 12 layers and 12 heads, yielding a total of 144 distinct attention mechanism. We denote a particular layer-head pair by <*layer*>-<*head*>, e.g. head *3-7* for the 3rd layer's 7th attention head.

Analysis. We explore how attention aligns with known structural and functional properties of proteins. From a structural perspective, we analyze *contact maps*, which describe pairs of non-adjacent amino acids that are in contact in the folded protein structure. Specifically, we measure the proportion of attention that aligns with contact maps, averaged over a large dataset. From a functional perspective, we measure the proportion of attention that targets *binding sites*, regions of a protein that bind with other macromolecules to perform specific functions. Finally, we analyze attention with respect to the *substitution matrix*, a pairwise measure of similarity between amino acids based on how readily they may be substituted for one another. Specifically, we consider the BLOSUM matrix, which is derived from co-occurrence statistics of amino acids in aligned protein

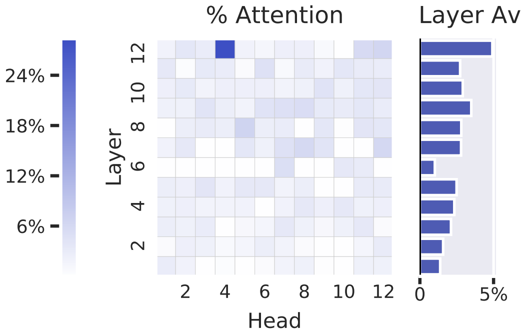

Figure 3. Percentage of each head's attention that is aligned with contact maps, averaged over a dataset. Each heatmap cell shows the value for a single head, indexed by layer (vertical axis) and head index (horizontal axis). For example, the dark blue cell in the upper-middle portion shows that 28% of attention from head 12-4 is aligned with contact maps.

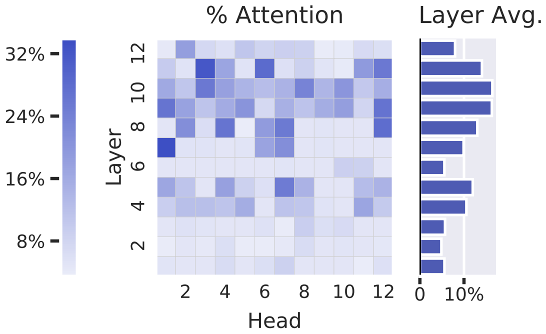

Figure 4. Percentage of each head's attention that focuses on binding sites. Especially in the deeper layers, binding sites are targeted at a much higher frequency than would occur by chance (4.8%). Head 7-1 has the highest percentage (34%).

sequences (Henikoff & Henikoff, 1992). In all analyses, we filter attention below a threshold of 0.1 to reduce the effects of very low-confidence attention patterns on the analysis.

Datasets. We use two protein sequence datasets from the TAPE repository for the analysis: the ProteinNet dataset (AlQuraishi, 2019; Fox et al., 2013; Berman et al., 2000; Moult et al., 2018) and the Secondary Structure dataset (Rao et al., 2019; Berman et al., 2000; Moult et al., 2018; Klausen et al., 2019). Both datasets contain amino acid sequences. The former also contains spatial coordinates of each amino acid, which is used for computing contact maps, and we augment the latter with available token-level binding site annotations from the Protein Data Bank (Berman et al., 2000). For both datasets, we sample a subset of 5000 sequences.

4. What does attention understand about proteins?

4.1. Protein Structure

Attention aligns strongly with contact maps in one attention head. Figure 3 shows the percentage of each head's attention that aligns with contact maps (see Section 3). A single head, 12-4, aligns much more strongly with contact maps (28% of attention) than any of the other heads (maximum 7% of attention). For high-confidence (> 0.9) attention in head 12-4, the alignment increases to 76%. In contrast, the frequency of contact pairs among all token pairs in the dataset is 1.3%. Figure 1 shows an example protein and the induced attention from head 12-4. See Appendix A for a fine-grained analysis and statistical significance test of the relationship between attention and contact maps.

Considering the model was trained with a masked language modeling objective with no spatial information in its inputs or training labels, the presence of a singular head that identifies contacts is surprising. One potential reason for this localizing behavior could be that contacts are more likely to biochemically interact with one another, thereby constraining the amino acids that may occupy these positions. In a language model, therefore, knowing contacts of masked tokens could provide valuable context for token prediction.

While there seems to be a strong correlation between the attention head output and classically-defined contacts, there are also differences. The model may have learned a differing contextualized or nuanced formulation that describes amino acid interactions. These learned interactions could then be used for further discovery and investigation or repurposed for prediction tasks similar to how principles of co-evolution enabled a powerful representation for structure prediction.

4.2. Binding Sites

Attention targets binding sites, especially in the deeper layers. Figure 4 shows the proportion of attention focused on binding sites by each head. In most layers, the mean percentage across heads is significantly higher than the background frequency of binding sites (4.8%). The effect is strongest in the last 6 layers of the model, which include 15 heads that each focus over 20% of their attention on binding sites. Head 7-1, depicted in Figure 2, focuses the most attention on binding sites (34%). See Appendix A for a fine-grained analysis and statistical significance test of the attention patterns in this head. We also find that tokens often target binding sites from far away in the sequence. In head 7-1, for example, the average distance spanned by attention to binding sites is 124 tokens.

Why does attention target binding sites, especially from long distances within the sequence? Evolutionary pressures have naturally selected proteins among the combinatorial space of possible amino acid sequences by the guiding principle that they exhibit critical function to ensure fitness. Proteins largely function to bind to other molecules, whether

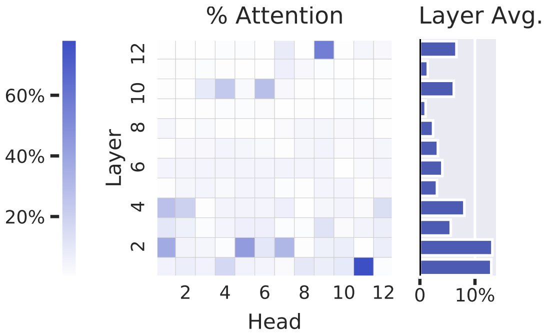

Figure 5. Percentage of each head's attention that focuses on the amino acid *Pro*, averaged over the dataset.

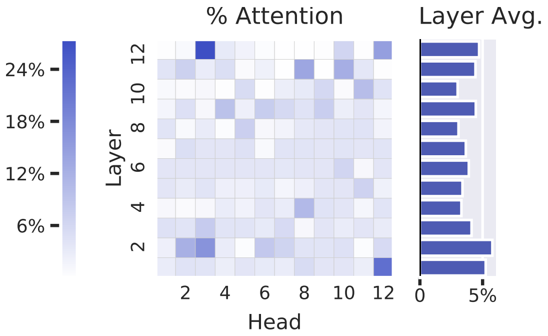

Figure 6. Percentage of each head's attention that focuses on the amino acid *Phe*, averaged over the dataset.

small molecules, proteins, or other macromolecules. Past work has shown that binding sites can reveal evolutionary relationships among proteins (Lee et al., 2017) and that particular structural motifs in binding sites are mainly restricted to specific families or superfamilies of proteins (Kinjo & Nakamura, 2009). Thus binding sites provide a high-level characterization of the protein that may be relevant for the model throughout the sequence.

4.3. Amino Acids and Substitution Relationships

Attention heads specialize in certain types of amino acids. We computed the proportion of attention that each head focuses on particular types of amino acids, averaged over a dataset. We found that for 14 of the 20 types of amino acids, there exists a head that focuses over 25% of attention on that amino acid. For example, head 1-11 (Figure 5) focuses 78% of its total attention on the amino acid *Pro*, and head 12-3 (Figure 6) focuses 27% of attention on *Phe*.

Attention is consistent with substitution relationships. A natural follow-up question is whether each head has "memorized" specific amino acids to target, or whether it has actually learned meaningful properties that correlate with particular amino acids. To test the latter hypothesis, we analyze how the attention received by amino acids relates to an existing measure of structural and functional properties: the *substitution matrix* (see Section 3). We assess whether

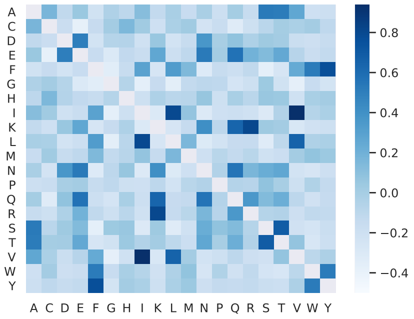

Figure 7. Attention similarity between pairs of amino acids.

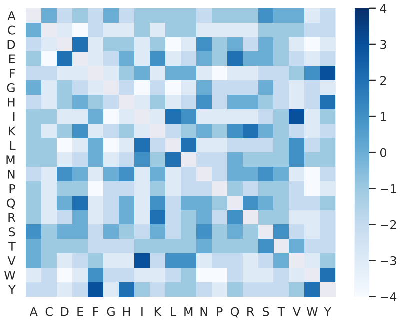

Figure 8. BLOSUM62 substitution matrix.

attention tracks similar properties by computing the similarity of *attention* between a pair of amino acids and then comparing this metric to the pairwise similarity based on the substitution matrix. To measure attention similarity, we compute the Pearson correlation between the proportion of attention that each amino acid receives across heads. For example, to measure the attention similarity between *Pro* and *Phe*, we take the Pearson correlation of the heatmaps in Figures 5 and 6. The values of all such pairwise correlations are shown in Figure 7. We compare these scores to the BLOSUM scores in Figure 8, and find a Pearson correlation of 0.80, suggesting that attention is largely consistent with substitution relationships.

5. Conclusions and Future Work

The present analysis identifies associations between attention and various properties of proteins. It does not attempt to establish a causal link between attention and model behavior (Vig et al., 2020a; Grimsley et al., 2020), nor to *explain* model predictions (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). While this paper focuses on reconciling attention patterns with known properties of protein, one could also use attention to search for novel properties and relationships as a means to aid scientific discovery. But in order for learned representations to be accessible to domain experts, they must be presented in a relevant context, e.g. embedded within protein structure (Figures 1 and 2). We believe there is great potential to develop more such contextual visualizations in biology and other scientific domains.

References

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. arXiv:1608.04207 [cs.CL]., 2016.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering with sequence-only deep representation learning. *bioRxiv*, pp. 589333, 2019.

AlQuraishi, M. ProteinNet: a standardized data set for machine learning of protein structure. *arXiv preprint arXiv:1902.00249*, 2019.

Arnold, F. H. Design by directed evolution. *Accounts of chemical research*, 31(3):125–131, 1998.

Asgari, E. and Mofrad, M. R. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11), 2015.

Bepler, T. and Berger, B. Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., and Wattenhofer, R. On identifiability in transformers. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJg1f6EFDB.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. What does BERT look at? an analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. What you can cram into a single vector: Probing sentence embeddings for linguistic properties, 2018.

Correia, G. M., Niculae, V., and Martins, A. F. T. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, January 2019a. doi: 10.1093/nar/gky995.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1): D427–D432, 2019b. ISSN 0305-1048. doi: 10.1093/nar/gky995. URL https://academic.oup.com/nar/article/47/D1/D427/5144153.

Ethayarajh, K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

Ettinger, A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309, 2013.

Goldberg, Y. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

Grimsley, C., Mayfield, E., and R.S. Bursten, J. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 1780–1790, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.220.

Henikoff, S. and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992. ISSN 0027-8424. doi: 10.1073/pnas.89.22.10915. URL https://www.pnas.org/content/89/22/10915.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4129–4138, 2019.

Hoover, B., Strobelt, H., and Gehrmann, S. exBERT: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276*, 2019.

Htut, P. M., Phang, J., Bordia, S., and Bowman, S. R. Do attention heads in BERT track syntactic dependencies? *arXiv preprint arXiv:1911.12246*, 2019.

Huang, P.-S., Boyken, S. E., and Baker, D. The coming of age of de novo protein design. *Nature*, 537(7620): 320–327, 2016.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, pp. 15794–15805, 2019.

Jain, S. and Wallace, B. C. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, June 2019.

Kinjo, A. and Nakamura, H. Comprehensive structural classification of ligand-binding motifs in proteins. *Structure*, 17(2), 2009.

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Soenderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B., et al. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 2019.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. Revealing the dark secrets of BERT. *arXiv preprint arXiv:1908.08593*, 2019.

Lee, J., Konc, J., Janezic, D., and Brooks, B. Global organization of a binding site network gives insight into evolution and structure-function relationships of proteins. *Sci Rep*, 7(11652), 2017.

Lin, Y., Tan, Y. C., and Frank, R. Open sesame: Getting inside BERT's linguistic knowledge. *arXiv preprint arXiv:1906.01698*, 2019.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M., and Smith, N. A. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019.

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. Progen: Language modeling for protein generation. *bioRxiv*, 2020.

Mickus, T., Paperno, D., Constant, M., and van Deemeter, K. What do you mean, BERT? assessing BERT as a distributional semantics model. *arXiv preprint arXiv:1911.05758*, 2019.

Moradi, P., Kambhatla, N., and Sarkar, A. Interrogating the explanatory power of attention in neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pp. 221–230, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5624. URL https://www.aclweb.org/anthology/D19-5624.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins: Structure, Function, and Bioinformatics*, 86:7–15, 2018. ISSN 08873585. doi: 10.1002/prot.25415. URL http://doi.wiley.com/10.1002/prot.25415.

Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., and Lipton, Z. C. Learning to deceive with attention-based explanations. In *Annual Conference of the Association for Computational Linguistics (ACL)*, July 2020. URL https://arxiv.org/abs/1909.07913.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. S. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, 2019.

Riesselman, A. J., Shin, J.-E., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Accelerating protein design using autoregressive generative models. *bioRxiv*, pp. 757252, 2019.

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, pp. 622803, 2019.

Rogers, A., Kovaleva, O., and Rumshisky, A. A primer in bertology: What we know about how bert works, 2020.

Rose, A. S. and Hildebrand, P. W. NGL Viewer: a web application for molecular visualization. *Nucleic Acids Research*, 43(W1):W576–W579, 04 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv402. URL https://doi.org/10.1093/nar/gkv402.

Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlić, A., and Rose, P. W. NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, 34(21):3755–3758, 05 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty419. URL https://doi.org/10.1093/bioinformatics/bty419.

Rosenfeld, L., Heyne1, M., Shifman, J. M., and Papo, N. Protein engineering by combined computational and *in vitro* evolution approaches. *Trends in Biochemical Sciences*, 41(5):421–433, May 2016. doi: 10.1016/j.tibs.2016.03.002.

Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H., and Laino, T. Unsupervised Attention-Guided Atom-Mapping. 5 2020. doi: 10.26434/chemrxiv.12298559.v1. URL https://chemrxiv.org/articles/Unsupervised_Attention-Guided_Atom-Mapping/12298559.

Serrano, S. and Smith, N. A. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://www.aclweb.org/anthology/P19-1282.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

Vashishth, S., Upadhyay, S., Tomar, G. S., and Faruqui, M. Attention interpretability across NLP tasks. *arXiv preprint arXiv:1909.11218*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Veldhoen, S., Hupkes, D., and Zuidema, W. H. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *CoCo@NIPS*, 2016.

Vig, J. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*, 2019.

Vig, J. and Belinkov, Y. Analyzing the structure of attention in a transformer language model. arXiv:1906.04284 [cs.CL]., 2019.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020a.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. Bertology meets biology: Interpreting attention in protein language models, 2020b. URL https://arxiv.org/abs/2006.15222.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, 2019.

Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*, 2019.

Wiegreffe, S. and Pinter, Y. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, November 2019.

Zhong, R., Shao, S., and McKeown, K. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*, 2019.

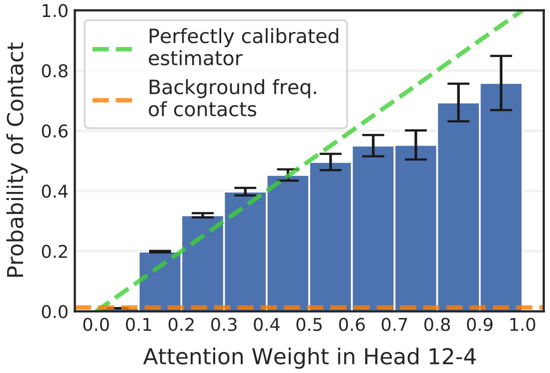

Figure 9. Probability two amino acids are in contact [95% confidence intervals], as a function of attention between the amino acids in Head 12-4, showing attention approximates a perfectly-calibrated estimator (green line).

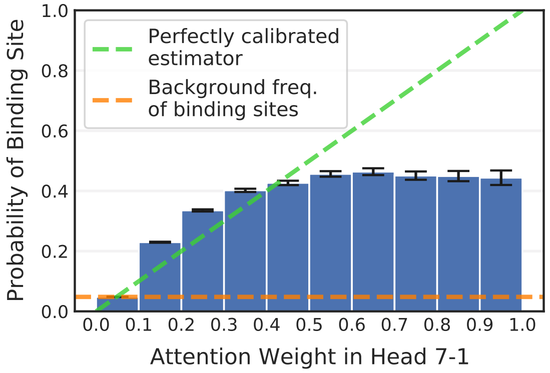

Figure 10. Probability that amino acid is a binding site [95% confidence intervals], as a function of attention received in Head 7-1. The green line represents a perfectly calibrated estimator.

A. Fine-grained analysis of attention heads

It has been suggested that attention weights represent a model's confidence in detecting certain features (Voita et al., 2019; Correia et al., 2019). We test this hypothesis with respect to contact maps by comparing attention weight in head 12-4 with the probability of two amino acids being in contact. We estimate the contact probability by binning all amino acid pairs (i, j) in the dataset by their attention weights $\alpha_{i,j}$, and calculating the proportion of pairs in each bin that are in contact. The results are shown in Figure 9. The Pearson correlation between the estimated probabilities and the attention weights (based on the midpoint of each bin) is 0.97. This suggests that attention weight is a well-calibrated estimator in this case, providing a principled interpretation of attention as a measure of confidence.

We perform a similar comparative analysis between the attention weight in head 7-1 and the probability that the token receiving attention is a binding site, shown in Figure 10. We find that attention weights less than 0.5 approximate the probability, but the predictive power of attention plateaus after this point.