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
{jvig,amadani,lvarshney,cxiong,rsocher,nazneen.rajani}@salesforce.com

Abstract

Transformer architectures have proven to learn useful representations for protein classification and generation tasks. However, these representations present challenges in interpretability. Through the lens of attention, we analyze the inner workings of the Transformer and explore how the model discerns structural and functional properties of proteins. 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 also present a three-dimensional visualization of the interaction between attention and protein structure. Our findings align with known 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`.

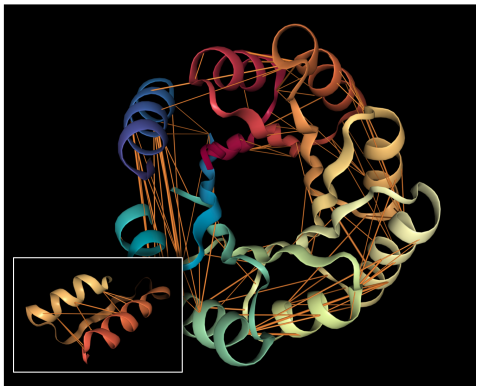
1 Introduction

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 [6, 33, 69]. The decreasing cost of sequencing technology has enabled us to collect vast databases of naturally occurring proteins [21], which are rich in information that is ripe for developing powerful sequence-based approaches. For example, sequence models leveraging principles of co-evolution, whether modeling pairwise or higher-order interactions, have enabled prediction of structure or function [66].

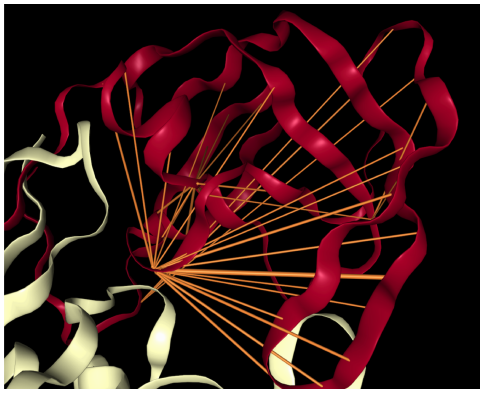
Proteins, as a sequence of amino acids, can be viewed precisely as a language. As such, they may be modeled using neural architectures that have been developed for natural language. In particular, the Transformer [80], which has led to a revolution in unsupervised learning for text, shows promise for a similar impact on protein sequence modeling. However, the strong performance of the Transformer comes at the cost of interpretability, and this lack of transparency can hide underlying problems such as model bias and spurious correlations [41, 55, 75]. In response, a great deal of NLP research now focuses on interpreting the Transformer, e.g., the subspecialty of "BERTology" [65], which studies the BERT [19] Transformer model specifically.

In this work, we adapt and extend this line of interpretability research to protein sequences. We analyze a Transformer protein model through the lens of attention, the core component of the Transformer architecture, and we present a set of analysis methods that capture the unique functional and structural characteristics of protein sequences. In contrast to NLP, which aims to automate a capability that humans already possess—understanding natural language—protein modeling also

Preprint. Under review.



(a) Attention in head 12-4, which targets amino acid pairs that are close in physical space (see inset subsequence 117D-157I) but lie apart in the sequence. Example is a *de novo* designed TIM-barrel (5BVL) with characteristic symmetry.



(b) Attention in head 7-1, which targets binding sites, a key functional component of proteins. Example is HIV-1 protease (7HVP). The primary location receiving attention is 27G, a binding site for protease inhibitor small-molecule drugs.

Figure 1: Two examples of how specialized attention heads in a Transformer recover protein structure and function, based solely on language model pre-training. The orange lines depict attention between amino acids, with line width proportional to attention weight. Heads were selected based on alignment with ground-truth annotations of contact maps and binding sites. Attention values below 0.1 are hidden. Visualizations based on the NGL Viewer [54, 67, 68].

seeks to shed light on biological processes that are not yet fully understood. Therefore we also discuss how interpretability can aid scientific discovery.

Our analysis reveals that attention captures high-level structural properties of proteins, connecting amino acids that are spatially close in three-dimensional structure, but far apart in the underlying sequence (Figure 1a). We also find that attention targets binding sites, a key functional component of proteins (Figure 1b). Further, we show that attention captures substitution properties of amino acids, and constructs progressively higher-level representations of structure and function with increasing layer depth.

2 Background: Proteins

In this section we provide background on the biological concepts discussed in later sections.

Amino acids. Protein sequences are composed from a vocabulary of 20 standard amino acids. Just as different words may convey similar meanings, different amino acids may play similar structural and functional roles in a sequence. While word synonyms are encoded in a thesaurus, similar amino acids are captured in a *substitution matrix*, which scores pairs of amino acids on how readily they may be substituted for one another while maintaining protein viability. One common substitution matrix is BLOSUM [29], which is derived from co-occurrence statistics of amino acids in aligned protein sequences.

Secondary and tertiary structure. Though a protein may be abstracted as a sequence of amino acids, it represents a physical entity with well-defined three-dimensional structure (Figure 1). *Secondary structure* describes the local segments of proteins, and may be divided into three broad categories: *Helix*, *Strand*, and *Turn/Fold*, as well as an *Other* category for local structure that does not fall into one of these categories. *Tertiary structure* encompasses the large-scale formations that determine the overall shape and function of the protein. One way to characterize tertiary structure is by a *contact map*, which describes the pairs of amino acids that are in contact (within 8 angstroms of one another) in the folded protein structure but lie apart (by at least 6 positions) in the underlying sequence [60].

Binding sites. Proteins may also be characterized by their functional properties. *Binding sites* are regions of protein sequences that bind with other molecules to carry out a specific function. For example, the HIV-1 protease is an enzyme responsible for a critical process in replication of HIV [11]. It has a binding site shown in Figure 1b which is a target for drug development to ensure inhibition.

3 Methodology

Model. We study the BERT Transformer model [19], though this analysis could be applied to other Transformers as well. In particular, we use the BERT-Base model from the TAPE repository [60], which was pretrained on masked language modeling of amino acids over a dataset of 31 million protein sequences [22]. This model accepts as input a sequence of amino acids $\boldsymbol{x} = (x_1, ..., x_L)$ and outputs a sequence of continuous embeddings $\boldsymbol{z} = (z_1, ..., z_L)$. The architecture comprises a series of encoder layers, each of which includes multiple attention heads. Each attention head generates a distinct set of attention weights α for an input, where $\alpha_{i,j} > 0$ represents the attention from token i to token j in the sequence, such that $\sum_j \alpha_{i,j} = 1$. Intuitively, attention weights define the influence of every token on the next layer's representation for the current token. 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_index>*, e.g. head *3-7* for the 3rd layer's 7th attention head. A detailed formulation of the attention mechanism may be found in Appendix A.1.

Attention analysis. We analyze how attention aligns with various protein properties, both at the token level (secondary structure, binding sites) and at the token-pair level (contact maps). In the latter case, we define an indicator function $f(i, j)$ for property f that returns 1 if the property is present in token pair (i, j) (i.e., if amino acids i and j are in contact), and zero otherwise. We then compute the proportion of attention that aligns with f over a dataset $\boldsymbol{X}$ as follows:

$$p_\alpha(f) = \sum_{x \in X} \sum_{i=1}^{|x|} \sum_{j=1}^{|x|} f(i, j) \alpha_{i,j}(x) \Big/ \sum_{x \in X} \sum_{i=1}^{|x|} \sum_{j=1}^{|x|} \alpha_{i,j}(x) \tag{1}$$

where $\alpha_{i,j}(x)$ denotes the attention from i to j for input sequence x.

For a *token*-level property f, we define $f(i, j)$ to be an indicator that returns 1 if the property is present in token j (e.g. if j is a binding site) and zero otherwise. In this case, $p_\alpha(f)$ represents the proportion of attention that is directed *to* property f.

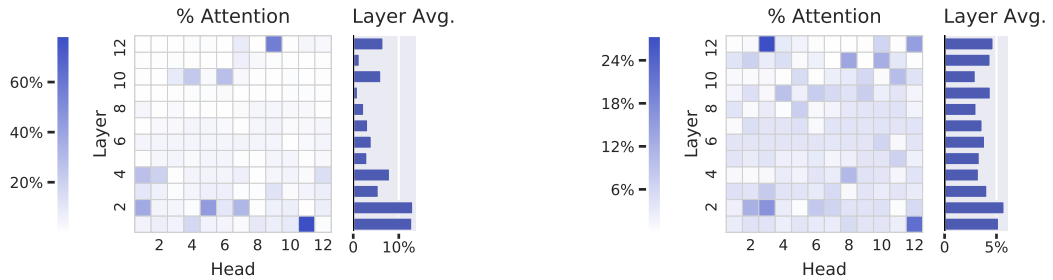
Probing tasks. Besides analyzing attention, we also use a diagnostic classifier [1, 16, 81] to probe the layer outputs to determine what information they contain about the properties described above. For the classifier, we use a single linear layer followed by softmax, and we freeze the weights of the original model. For token-level probing tasks (binding sites, secondary structure), we feed each token's output vector directly to the classifier. For token-pair probing tasks (contact map), we construct a pairwise feature vector by concatenating the elementwise differences and products of the two tokens' output vectors, following the TAPE[1] implementation. We use task-specific evaluation metrics: for secondary structure prediction, we measure F1 score; for contact prediction; we measure precision@$L/5$, where L is the length of the protein sequence, following standard practice [52]; for binding site prediction, we measure precision@$L/20$, since approximately one in twenty amino acids in each sequence is a binding site (4.8% in the dataset described below).

Datasets. We use two protein sequence datasets from the TAPE repository for the analysis: the ProteinNet dataset [3, 10, 26, 52] and the Secondary Structure dataset [10, 38, 52, 60]. Both datasets contain amino acid sequences. ProteinNet is also annotated with spatial coordinates of each amino acid, which is used for generating contact maps, and the Secondary Structure dataset is annotated with the secondary structure at each sequence position. For our analyses of amino acids and contact maps, we use ProteinNet, and for the analysis of secondary structure and binding sites we use the Secondary Structure dataset. For the binding site analysis, we also obtained available token-level binding site annotations from the Protein Data Bank [10]. For analyzing attention, we used a random subset of 5000 sequences from the respective training splits as this analysis was purely evaluative; for training the probing classifier, we used the full training splits for training the model, and the validation splits for evaluation. More details about the datasets may be found in Appendix A.2.

Experimental details We exclude attention to the `[SEP]` delimiter token, as it has been shown to be a "no-op" attention token [15], as well as attention to the `[CLS]` token, which is not explicitly used in language modeling. We also filter attention below a minimum threshold of 0.1 (on a scale from 0 to 1) to reduce the effects of very low-confidence attention patterns on the analysis. We truncate all protein sequences to a maximum length of 512 to reduce the model memory requirements.[2] Experiments are performed using a single Tesla V-100 GPU with 16GB memory.

[1] `https://github.com/songlab-cal/tape`

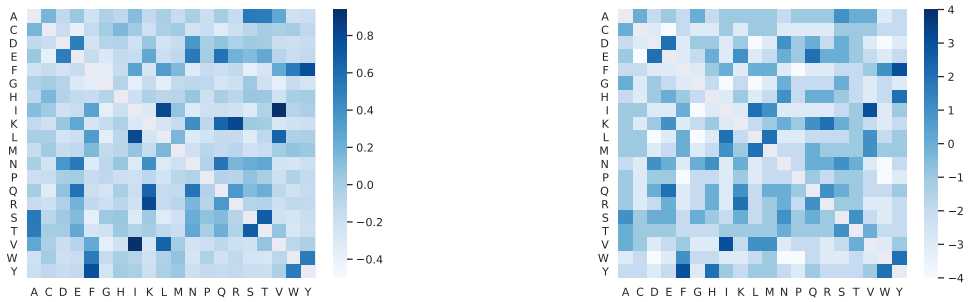
[2] 94% of sequences had fewer than 512 amino acids.



(a) Attention to amino acid *Pro*

(b) Attention to amino acid *Phe*

Figure 2: Percentage of each head's attention that focuses on amino acids *Pro* and *Phe*, respectively, averaged over the dataset. Each heatmap cell shows the value for a single head, indexed by layer (vertical axis) and head index (horizontal axis). For example, in (a), the dark blue cell in the lower right corner shows that head 1-11 focuses 78% of its attention on *Pro*. These and other results suggest that attention heads specialize in certain types of amino acids. See Appendix B.1 for complete results.



(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.

4 What does attention understand about proteins?

4.1 Amino Acids

We begin by examining the interaction between attention and particular types of amino acids.

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 of 5000 sequences with a combined length of 1,067,712 amino acids. We found that for 14 of the 20 types of amino acids, there exists an attention head that focuses over 25% of attention on that amino acid. For example, Figure 2 shows that head 1-11 focuses 78% of its total attention on the amino acid *Pro* and head 12-3 focuses 27% of attention on *Phe*. Note that the maximum frequency of any single type of amino acid in the dataset is 9.4%. Detailed results for all amino acids are included in Appendix B.1.

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 2). We assess whether attention tracks similar properties by computing the similarity of *attention* between each 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 two heatmaps in Figure 2. The values of all such pairwise correlations are shown in Figure 3a. We compare these scores to the BLOSUM scores in Figure 3b, and find a Pearson correlation of 0.80, suggesting that attention is largely consistent with substitution relationships.

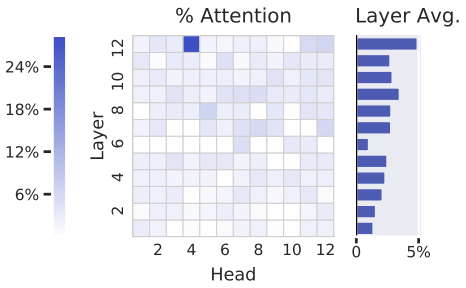


Figure 4: Percentage of each head's attention that is aligned with contact maps, averaged over a dataset, suggesting that Head 12-4 is uniquely specialized for contact prediction.

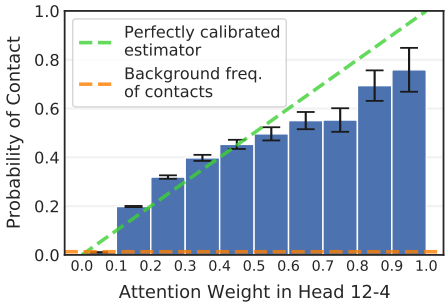


Figure 5: Probability two amino acids are in contact [95% confidence intervals], as a function of attention between the amino acids in Head 12-4, showing attention approximates a perfectly-calibrated estimator (green line).

4.2 Protein Structure

Here we explore the relationship between attention and tertiary structure, as characterized by contact maps (see Section 2). Secondary structure results are included in Appendix B.2.

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 maps (28% of attention) than any of the other heads (maximum 7% of attention). In cases where the attention weight in head 12-4 is greater than 0.9, the alignment increases to 76%. In contrast, the frequency of contact pairs among all token pairs in the dataset is 1.3%. Figure 1a shows an example protein and the induced attention from head 12-4.

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. But *how* is the model able to capture these contact relationships? One possible explanation is that the model learns the underlying geometry of the protein structure. Alternatively, the model may utilize statistical regularities resulting from the contact relationship, e.g., that amino acids in contact often coevolve. The latter explanation seems especially relevant for contacts that are far apart in the underlying sequence, since it would likely be difficult for the model to faithfully capture the geometry over a very long interval.

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 coevolution enabled a powerful representation for structure prediction.

Attention is a well-calibrated predictor of contact maps. It has been suggested that attention weights represent a model's confidence in detecting certain features [17, 85]. We test this hypothesis with respect to contact maps by comparing attention weight in head 12-4, discussed above, 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 5. 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.

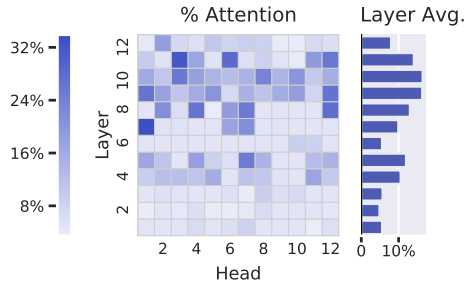


Figure 6: Percentage of each head's attention that focuses on binding sites. Especially in the deeper layers, binding sites are targeted at a much higher frequency than would occur by chance (4.8%). Head 7-1 has the highest percentage (34%).

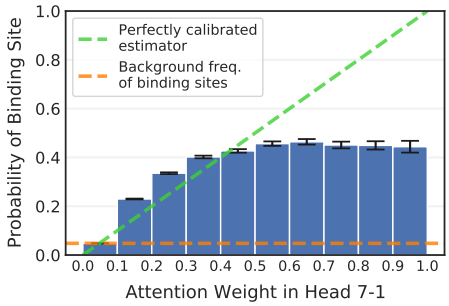


Figure 7: Probability that amino acid is a binding site [95% confidence intervals], as a function of attention received in Head 7-1. The green line represents a perfectly calibrated estimator.

4.3 Binding Sites

We also analyze how attention interacts with binding sites, a key functional component of proteins.

Attention targets binding sites, especially in the deeper layers. Figure 6 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 1b, focuses the most attention on binding sites (34%). Figure 7 shows the estimated probability of this head targeting a binding site, as a function of the attention weight. 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 small molecules, proteins, or other macromolecules. Past work has shown that binding sites can reveal evolutionary relationships among proteins [42] and that particular structural motifs in binding sites are mainly restricted to specific families or superfamilies of proteins [37]. Thus binding sites provide a high-level characterization of the protein that may be relevant for the model throughout the sequence.

4.4 Cross-Layer Analysis

Here we analyze how attention captures properties of varying complexity across different layers of the model, and we compare these results to a probing analysis of layer outputs (see Section 3).

Attention targets higher-level properties in deeper layers. As shown in Figure 8, deeper layers focus relatively more attention on binding sites and contacts (high-level concept), whereas secondary structure (low- to mid-level concept) is targeted more evenly across layers. The probing analysis (Figure 9) similarly shows that the model first forms representations of secondary structure before fully encoding contact maps and binding sites. This result is consistent with the intuition that the model must understand local structure before it can form representations of higher-order structure and function. Prior work in NLP also suggests that deeper layers in text-based Transformers attend to more complex properties [83] and encode higher-level representations [36, 56, 58, 76].

However, one seemingly contrary result is the drop in performance of the contact map classifier in the final layer (Figure 9). This may be explained by the fact that the final layer outputs are not used to compute attention, but rather as an input to the final classifier layer that predicts the masked token for the language model. Whereas attention is a property of token pairs, the predictions are specific to an individual token. Therefore, contact relationships, which exist between tokens, may not be as useful to the model in the final layer. Practitioners may take note to consider the second to last layer when generating embeddings for downstream pairwise tasks.

Figure 8: Each plot shows the percentage of attention focused on the given property, averaged over all heads within each layer. The plots, sorted by center of gravity (red dashed line), show that heads in deeper layers focus relatively more attention on binding sites and contacts, whereas attention toward specific secondary structures (*Strand*, *Turn/Bend*, *Helix*) is more even across layers. More granular analyses of these properties can be found in Figures 4 and 6 and Appendix B.2.

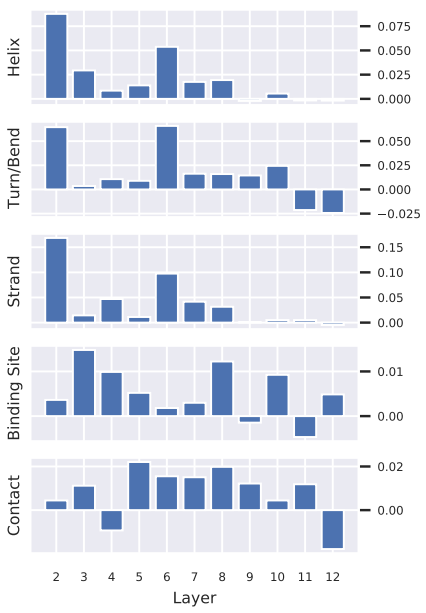


Figure 9: Differential performance of diagnostic classifier by layer, sorted by task order in Figure 8. Each plot shows the change in performance between the given layer and the previous layer. For secondary structure (*Strand*, *Turn/Bend*, *Helix*) the increases occur primarily in the first 5 layers, while increases for binding sites and contact maps are spread more evenly across layers. See Section 3 for a description of evaluation metrics. Absolute performance for each task included in Appendix C.

5 Related Work

Here we contextualize our techniques and findings in the protein modeling and interpretable neural network literatures, especially in natural language processing.

5.1 Protein language models

There is a vast literature dedicated to advancing machine learning (ML) algorithms for proteins. More specifically, deep neural networks for protein language modeling have received broad interest. Early work in this direction applied the Skip-gram model [50] to construct continuous embeddings from protein sequences for downstream tasks [7]. Sequence-only language models have since been trained through autoregressive or autoencoding self-supervision objectives for both discriminative and generative tasks, for example, using LSTMs or Transformer-based architectures [2, 9, 60, 64]. [64] 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 five tasks ranging from remote homology to fluorescence prediction to assess protein representation learning models. In [48, 63], autoregressive generative models were trained to predict the functional effect of mutations and generate natural-like proteins. All aforementioned studies have been the subject of rapid development, with Transformer architectures seemingly providing the most promising avenue for future research. In addition to sequence-only approaches, Transformer architectures have been adapted to incorporate structural information [34].

5.2 Interpreting Models in NLP

The rise of deep neural networks in ML has also led to much work on interpreting these so-called black-box models. Interpretability is a broad topic in ML with many facets [20, 46]; our present work focuses on post hoc interpretation. This section reviews the NLP interpretability literature, which is directly comparable to our work on interpreting protein language models.

General interpretability. Interpretability techniques in NLP mainly focus on explaining instance-level predictions. Influence functions explain a model's prediction on a given instance by identifying the influential training examples for that prediction [14, 39]. Methods that learn to explain by extracting rationales from the input and only using those rationales to make a prediction are *faithful* by design [8, 13, 43, 62]. Model-agnostic methods that are not based on gradients do not require model differentiability. Examples include approaches that fit a linear classifier to a model's prediction on perturbations of the input [4, 61, 72]. Yet another line of research deploys post hoc rationalization for explaining model behavior. Gradient-based techniques are the most commonly used methods in this category [25, 73, 74]. Other approaches include using language models for generating explanations [53, 59].

Interpreting Transformers. The Transformer is a neural architecture that uses attention to accelerate learning [80]. Transformers are the backbone of state-of-the-art pre-trained language models in NLP including BERT [19]. BERTology focuses on interpreting what the BERT model learns about natural language by using a suite of probes and interventions [65]. So-called *diagnostic classifiers* are used to interpret the outputs from BERT's layers [81].

At a high level, mechanisms for interpreting BERT can be categorized into three main categories: interpreting the learned embeddings [1, 16, 23, 49, 86], BERT's learned knowledge of syntax [27, 30, 32, 45, 47, 76], and BERT's learned knowledge of semantics [24, 76].

Interpreting attention specifically. In this paper, we focused on the interpretability of Transformer language models as they apply to protein sequences, with a specific focus on the attention mechanism. Interpreting attention on natural language sequences is a well-established area of research [12, 30, 87, 90]. In some cases, it has been shown that attention correlates with syntactic and semantic relationships in natural language [15, 32, 83]. To our knowledge, no previous work has examined whether attention is a *well-calibrated* estimator of these properties, though previous work as explored the calibration of Transformer outputs from an interpretability perspective [18].

Depending on the task and model architecture, attention may have more or less explanatory power for model predictions [35, 51, 57, 71, 79]. Visualization techniques have been used to convey the structure and properties of attention in Transformers [31, 40, 80, 82]. Recent work has begun to apply attention to guide mapping of sequence models outside of the domain of natural language [70].

Research on using pre-trained language models for protein sequences has been motivated by drawing parallels with natural language. Our work also draws motivation from interpretability techniques in NLP to uncover the properties captured by protein language models. The availability of TAPE suite of protein classification tasks inspired us to use diagnostic classifiers for probing of BERT's layers.

6 Conclusions and Future Work

This paper builds on the synergy between NLP and computational biology by adapting and extending NLP interpretability methods to protein sequence modeling. We show how a Transformer language model recovers structural and functional properties of proteins and integrates this knowledge directly into its attention mechanism. Our analysis reveals that attention not only captures properties of individual amino acids, but also discerns more global properties such as binding sites and tertiary structure. In some cases, attention also provides a well-calibrated measure of confidence in the model's predictions of these properties. We hope that machine learning practitioners can apply these insights in designing the next generation of protein sequence models.

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 [28, 84], nor to *explain* model predictions [35, 87]. While the focus of this paper is reconciling attention patterns

with known properties of proteins, one could also leverage attention to discover novel types of properties and processes. In this way, unsupervised methods like language modeling can serve as tools for scientific discovery. For example, one may be able to discover new insights into evolutionary history by better understanding phylogenetic relationships [89] or into development by understanding ontogenic relationships among cells (recently studied using single-cell RNA sequence data [88] rather than proteins). But in order for learned representations to be useful to domain experts, they must be presented in an appropriate context to facilitate discovery. Visualizing attention in the context of protein structure (Figure 1) is one example of this approach. We believe there is potential to develop other such contextual visualizations of learned representations in biology and other domains.

7 Broader Impact

Interpretability is the hallmark of trustworthy artificial intelligence and is especially important when technologies are directly used by people in consequential activities such as protein engineering or scientific discovery. This work illuminates some of the inner workings of large-scale language models for proteins, which are otherwise inscrutable, and may therefore facilitate their use by scientists and engineers. Many of the safety and governance considerations for pretrained language models for human language also apply for protein language [78], e.g. disparate resources available for training, fine-tuning, and using these models. The governance recommendations of post-market surveillance and co-responsibility among producers and users may be applied to mitigate these potential issues.

Scientists and engineers may pursue beneficent applications of protein language models such as the design of novel disease therapies, antibodies for novel infectious diseases, or food ingredients for low-allergenic infant formulas. They may also, however, pursue hostile applications such as the design of protein toxins [77], which, in the case of developing biological weapons, could have significant geopolitical ramifications. Various strategies could be applied to mitigate such risks, including broad-based governmental regulation, shared responsibility between users and producers of technology, and post-market surveillance.

Since the molecular biology revolution in the 1960s, there has been a large focus on a small number of model organisms, such as *C. elegans*, *Drosophila melanogaster*, and *Mus musculus*, as part of biological research [5, 44]. The protein databases that are used to train protein language models may therefore have unknown biases across the tree of life. These dataset biases may be reflected in the learned models, and further in their interpretation.

8 Acknowledgements

We would like to thank Xi Victoria Lin, Stephan Zheng, and Melvin Gruesbeck for their valuable feedback.

References

[1] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. arXiv:1608.04207 [cs.CL]., 2016.

[2] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-only deep representation learning. *bioRxiv*, page 589333, 2019.

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

[4] David Alvarez-Melis and Tommi Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–421, September 2017.

[5] Rachel A. Ankeny and Sabina Leonelli. What's so special about model organisms? *Studies in History and Philosophy of Science Part A*, 42(2):313–323, June 2011. doi: 10.1016/j.shpsa.2010.11.039.

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

[7] Ehsaneddin Asgari and Mohammad RK Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11), 2015.

[8] Joost Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2963–2977, July 2019.

[9] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019.

[10] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[11] Ashraf Brik and Chi-Huey Wong. HIV-1 protease: mechanism and drug discovery. *Organic & Biomolecular Chemistry*, 1(1):5–14, 2003.

[12] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BJg1f6EFDB`.

[13] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. A game theoretic approach to class-wise selective rationalization. In *Advances in Neural Information Processing Systems*, pages 10055–10065, 2019.

[14] Hongge Chen, Si Si, Yang Li, Ciprian Chelba, Sanjiv Kumar, Duane S. Boning, and Cho-Jui Hsieh. Multi-stage influence function. *Preprint*, 2019.

[15] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[16] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties, 2018.

[17] Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019.

[18] Shrey Desai and Greg Durrett. Calibration of pre-trained transformers, 2020. URL `https://arxiv.org/abs/2003.07892`.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

[20] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[21] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, Lorna J. Richardson, Gustavo A. Salazar, Alfredo Smart, Erik L. L. Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C. E. Tosatto, and Robert D. Finn. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, January 2019. doi: 10.1093/nar/gky995.

[22] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, 2019. ISSN 0305-1048. doi: 10.1093/nar/gky995. URL `https://academic.oup.com/nar/article/47/D1/D427/5144153`.

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

[24] Allyson Ettinger. 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.

[25] Shi Feng, Eric Wallace, Alvin Grissom, Mohit Iyyer, Pedro Rodriguez, and Jordan L. Boyd-Graber. Pathologies of neural models make interpretation difficult. In *EMNLP*, 2018.

[26] Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309, 2013.

[27] Yoav Goldberg. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

[28] Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 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`.

[29] S Henikoff and J G Henikoff. 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`.

[30] John Hewitt and Christopher D Manning. 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)*, pages 4129–4138, 2019.

[31] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exBERT: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276*, 2019.

[32] Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. Do attention heads in BERT track syntactic dependencies? *arXiv preprint arXiv:1911.12246*, 2019.

[33] Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, 2016.

[34] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, pages 15794–15805, 2019.

[35] Sarthak Jain and Byron C. Wallace. 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)*, pages 3543–3556, June 2019.

[36] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019. URL `https://hal.inria.fr/hal-02131630`.

[37] Akira Kinjo and Haruki Nakamura. Comprehensive structural classification of ligand-binding motifs in proteins. *Structure*, 17(2), 2009.

[38] Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjaergaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Soenderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, et al. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 2019.

[39] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017.

[40] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. *arXiv preprint arXiv:1908.08593*, 2019.

[41] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations, 2019.

[42] Juyong Lee, Janez Konc, Dusanka Janezic, and Bernard Brooks. Global organization of a binding site network gives insight into evolution and structure-function relationships of proteins. *Sci Rep*, 7(11652), 2017.

[43] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, 2016.

[44] Sabina Leonelli and Rachel A. Ankeny. Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Science Part C*, 43(1):29–36, March 2012. doi: 10.1016/j.shpsc.2011.10.003.

[45] Yongjie Lin, Yi Chern Tan, and Robert Frank. Open sesame: Getting inside BERT's linguistic knowledge. *arXiv preprint arXiv:1906.01698*, 2019.

[46] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[47] Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019.

[48] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *bioRxiv*, 2020.

[49] Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemeter. What do you mean, BERT? assessing BERT as a distributional semantics model. *arXiv preprint arXiv:1911.05758*, 2019.

[50] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf`.

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

[52] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. 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`.

[53] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*, 2020.

[54] Hai Nguyen, David A Case, and Alexander S Rose. NGLview–interactive molecular graphics for Jupyter notebooks. *Bioinformatics*, 34(7):1241–1242, 12 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx789. URL `https://doi.org/10.1093/bioinformatics/btx789`.

[55] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments, 2019.

[56] Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1179. URL `https://www.aclweb.org/anthology/D18-1179`.

[57] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 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`.

[58] Alessandro Raganato and Jörg Tiedemann. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5431. URL `https://www.aclweb.org/anthology/W18-5431`.

[59] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *Proceedings of the Association for Computational Linguistics (ACL)*, 2019.

[60] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, 2019.

[61] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, 2016.

[62] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[63] Adam J Riesselman, Jung-Eun Shin, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Accelerating protein design using autoregressive generative models. *bioRxiv*, page 757252, 2019.

[64] Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, page 622803, 2019.

[65] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works, 2020.

[66] Nathan J Rollins, Kelly P Brock, Frank J Poelwijk, Michael A Stiffler, Nicholas P Gauthier, Chris Sander, and Debora S Marks. Inferring protein 3d structure from deep mutation scans. *Nature genetics*, 51(7):1170, 2019.

[67] Alexander S. Rose and Peter W. Hildebrand. 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`.

[68] Alexander S Rose, Anthony R Bradley, Yana Valasatava, Jose M Duarte, Andreas Prlić, and Peter W Rose. 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`.

[69] Lior Rosenfeld, Michael Heyne1, Julia M. Shifman, and Niv Papo. 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.

[70] Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Unsupervised Attention-Guided Atom-Mapping. *ChemRxiv*, 5 2020. doi: 10.26434/chemrxiv.12298559.v1. URL `https://chemrxiv.org/articles/Unsupervised_Attention-Guided_Atom-Mapping/12298559`.

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

[72] Jaspreet Singh and Avishek Anand. EXS: Explainable search using local model agnostic interpretability. *arXiv preprint arXiv:1809.03857*, 2018.

[73] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. SmoothGrad: removing noise by adding noise. *ICML workshop on visualization for deep learning*, June 2017.

[74] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328, 2017.

[75] Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems 32*, pages 13230–13241. Curran Associates, Inc., 2019.

[76] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[77] Jonathan B. Tucker and Craig Hooper. Protein engineering: Security implications. *EMBO Reports*, 7:S14–S17, July 2006. doi: 10.1038/sj.embor.7400677.

[78] Lav R. Varshney, Nitish Shirish Keskar, and Richard Socher. Pretrained AI models: Performativity, mobility, and change. arXiv:1909.03290 [cs.CY]., September 2019.

[79] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. Attention interpretability across NLP tasks. *arXiv preprint arXiv:1909.11218*, 2019.

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[81] Sara Veldhoen, Dieuwke Hupkes, and Willem H. Zuidema. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *CoCo@NIPS*, 2016.

[82] Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*, 2019.

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

[84] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020.

[85] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, 2019.

[86] Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*, 2019.

[87] Sarah Wiegreffe and Yuval Pinter. 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)*, pages 11–20, November 2019.

[88] Haizi Yu, Lav R. Varshney, and Genevieve Stein-O'Brien. Towards learning human-interpretable laws of neurogenesis from single-cell RNA-seq data via information lattices. In *Learning Meaningful Representations of Life Workshop at NeurIPS 2019*, December 2019.

[89] Lijia Yu, Deepak Kumar Tanwar, Emanuel Diego S. Penha, Yuri I. Wolf, Eugene V. Koonin, and Malay Kumar Basu. Grammar of protein domain architectures. *Proc. Nat. Acad. Sci. USA*, 116(9):3636–3645, February 2019. doi: 10.1073/pnas.1814684116.

[90] Ruiqi Zhong, Steven Shao, and Kathleen McKeown. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*, 2019.

A Modeling and Experimental Details

A.1 BERT Transformer Architecture

Stacked Encoder: BERT uses a stacked-encoder architecture, which inputs a sequence of tokens $\boldsymbol{x} = (x_1, ..., x_L)$ and applies position and token embeddings followed by a series of encoder layers. Each layer applies multi-head self-attention (see below) in combination with a feedforward network, layer normalization, and residual connections. The output is a sequence of continuous embeddings $\boldsymbol{z} = (z_1, ..., z_L)$.

Self-Attention: Given an input $\boldsymbol{x} = (x_1, ..., x_L)$, the self-attention mechanism assigns to each token pair i, j an attention weight $\alpha_{i,j} > 0$ where $\sum_j \alpha_{i,j} = 1$. Attention in BERT is bidirectional. In the multi-layer, multi-head setting, α is specific to a layer and head. The BERT-Base model has 12 layers and 12 heads.

The attention weights $\alpha_{i,j}$ are computed from the scaled dot-product of the *query vector* of i and the *key vector* of j, followed by a softmax operation. The attention weights are then used to produce a weighted sum of value vectors:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

using query matrix Q, key matrix K, and value matrix V, where d_k is the dimension of K. In a multi-head setting, the queries, keys, and values are linearly projected h times, and the attention operation is performed in parallel for each representation, with the results concatenated.

A.2 Datasets

We used two protein sequence datasets from the TAPE repository for the analysis: the ProteinNet dataset [3, 10, 26, 52] and the Secondary Structure dataset [10, 38, 52, 60]. The former was used for analysis of amino acids and contact maps, and the latter was used for analysis of secondary structure. We additionally created a third dataset for binding site analysis from the Secondary Structure dataset, which was augmented with binding site annotations obtained from the Protein Data Bank's Web API.[3] We excluded any sequences for which binding site annotations were not available. The resulting dataset sizes are shown in Table 10. For the analysis of attention, a random subset of 5000 sequences from the training split of each dataset was used, as the analysis was purely evaluative. For training and evaluating the diagnostic classifier, the full training and validation splits were used.

Dataset	Train size	Validation size
ProteinNet	25299	224
Secondary Structure	8678	2170
Binding Sites	5734	1418

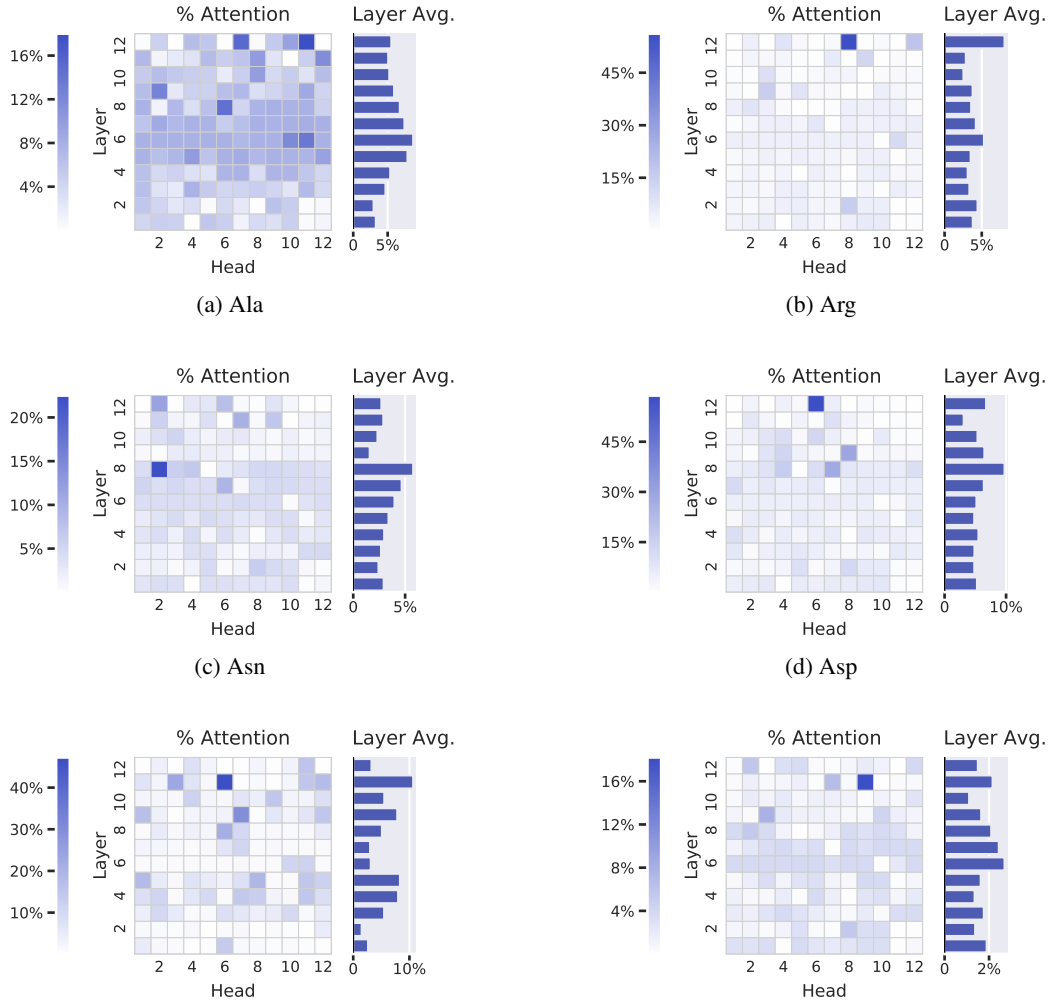
Figure 10: Datasets used in analysis

[3] `http://www.rcsb.org/pdb/software/rest.do`

B Additional Results of Attention Analysis

B.1 Amino Acids

Figures 11-13 show attention analysis results for all amino acids. Amino acid codes are found in Figure 14.

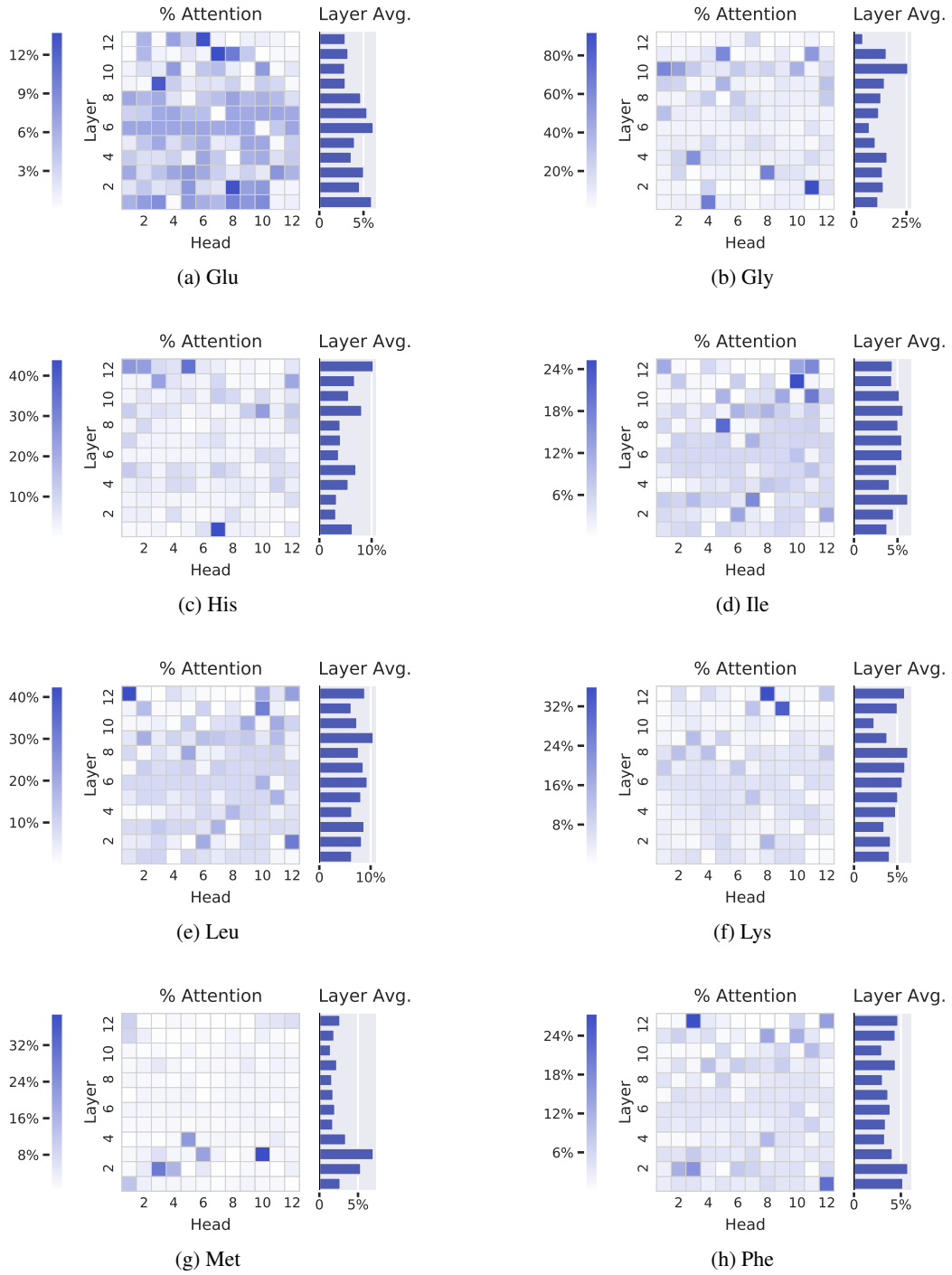


(a) Ala (b) Arg

(c) Asn (d) Asp

(e) Cys (f) Gln

Figure 11: Percentage of each head's attention that is focused on the given amino acid, averaged over a dataset. Each entry in the heatmap shows the value for a single head, while the bar plot on the right of each figure shows the layer average across heads.



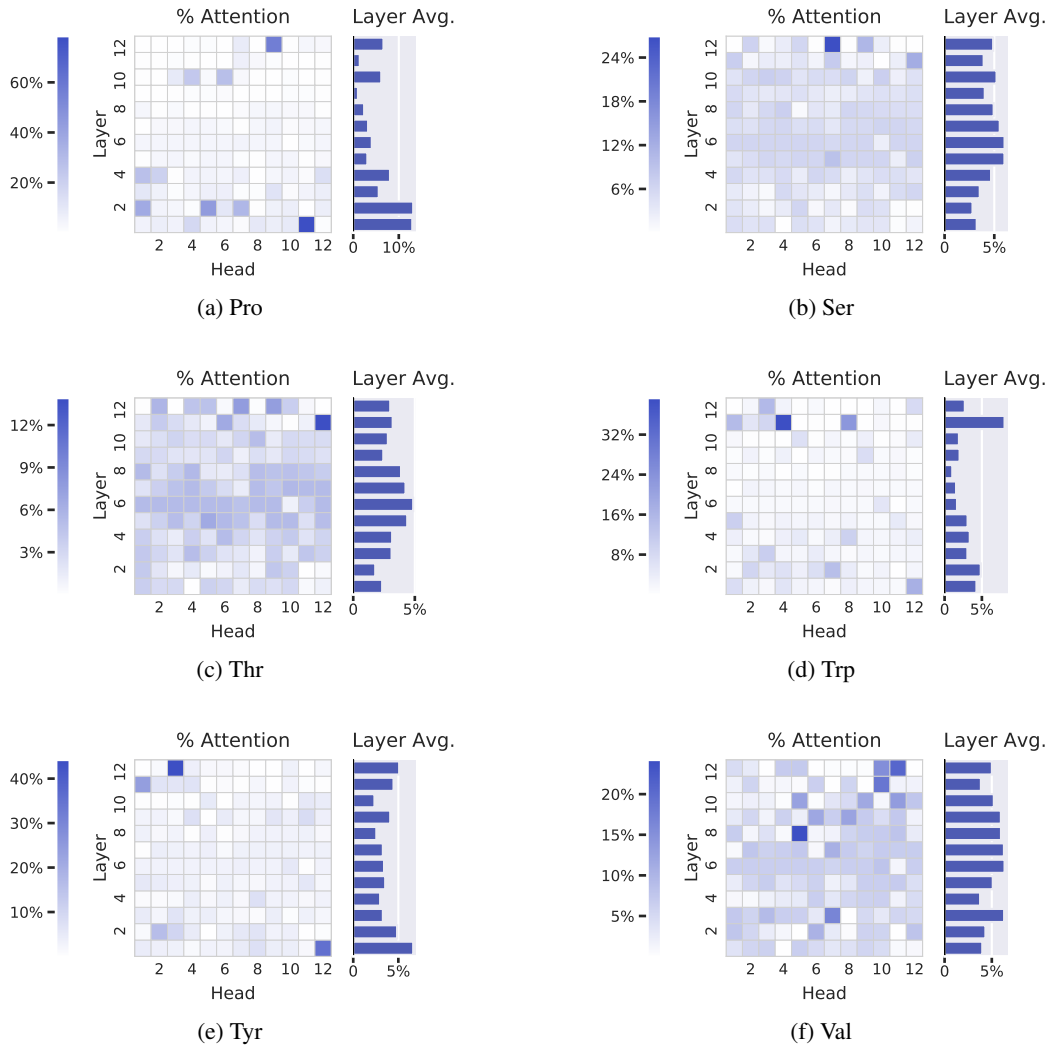
(a) Glu (b) Gly

(c) His (d) Ile

(e) Leu (f) Lys

(g) Met (h) Phe

Figure 12: Percentage of each head's attention that is focused on the given amino acid, averaged over a dataset. Each entry in the heatmap shows the value for a single head, while the bar plot on the right of each figure shows the layer average across heads. (cont.)



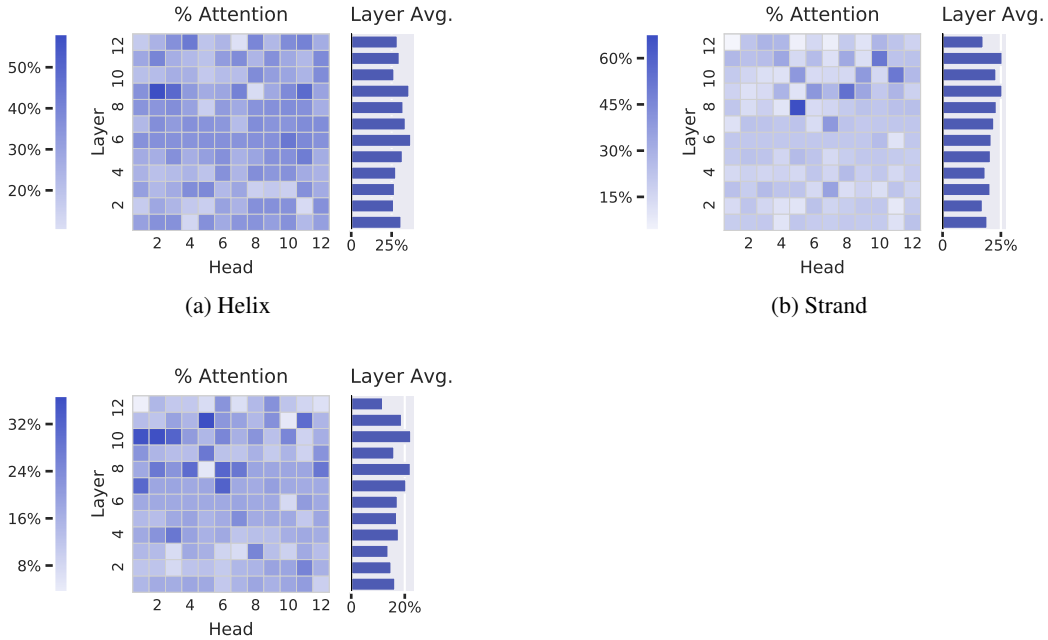
(a) Pro (b) Ser (c) Thr (d) Trp (e) Tyr (f) Val

Figure 13: Percentage of each head's attention that is focused on the given amino acid, averaged over a dataset. Each entry in the heatmap shows the value for a single head, while the bar plot on the right of each figure shows the layer average across heads. (cont.)

Abbrev.	Code	Name
Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine

Figure 14: Amino acids and associated codes.

B.2 Secondary Structure



(a) Helix

(b) Strand

(c) Turn/Bend

Figure 15: Percentage of each head's attention that is focused on the given secondary structure type, averaged over a dataset. Each entry in the heatmap shows the value for a single head, while the bar plot on the right of each figure shows the layer average across heads.

C Detailed Probing Results

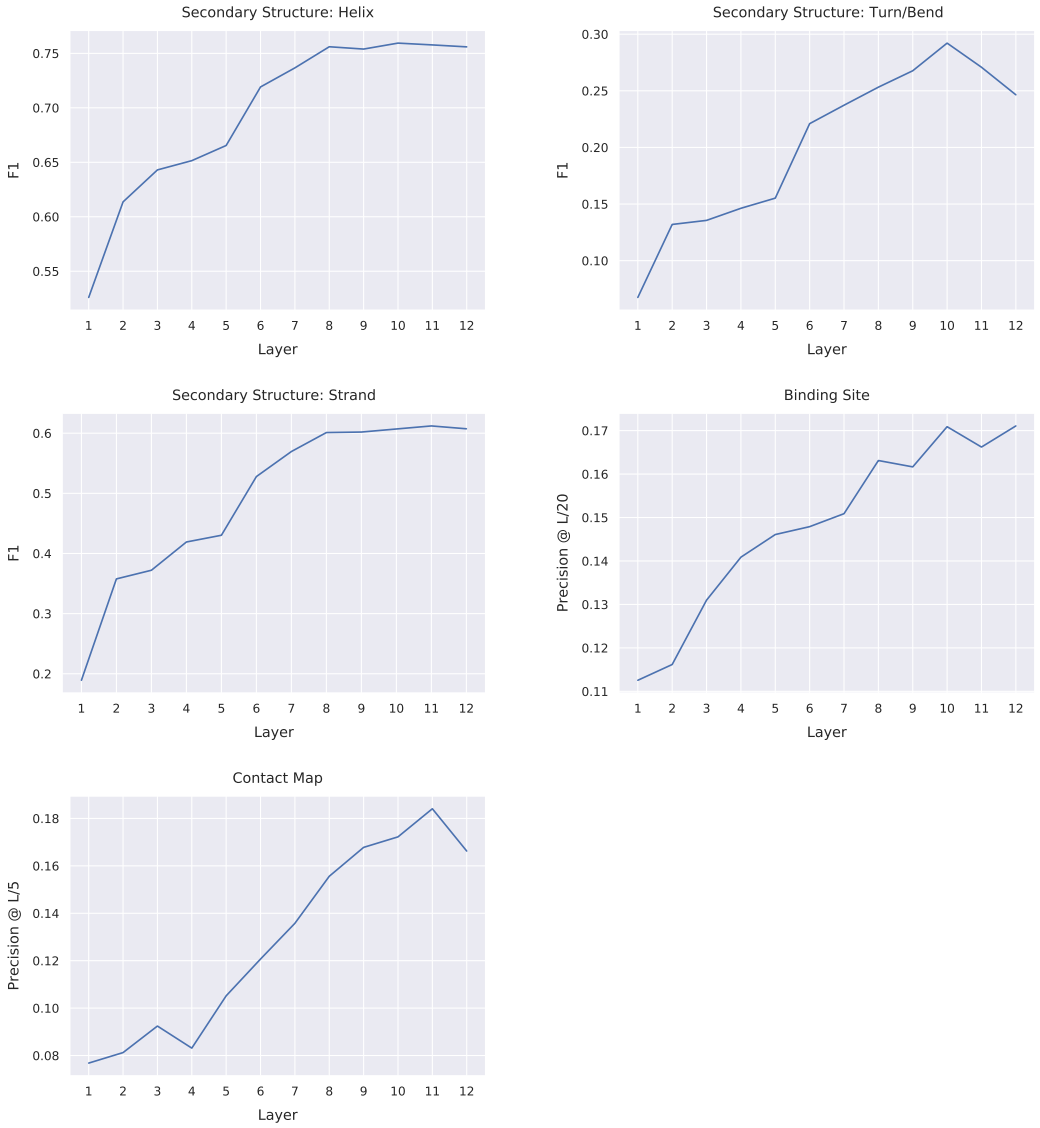


Figure 16: Diagnostic classifier performance for tasks described in Figure 9. For secondary structure prediction, we measure F1 score; for contact map prediction; we measure precision@$L/5$, where L is the length of the protein sequence, following standard practice [52]; for binding site prediction, we measure precision@$L/20$, since approximately one in twenty amino acids in each sequence is a binding site (4.8%). The raw scores of the different metrics cannot be compared directly; the goal of this analysis is to compare the relative shape of the performance curves across tasks.