# DEEP LEARNING PROTEINS USING A TRIPLET-BERT NETWORK

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Modern sequencing technology has produced a vast quantity of proteomic data, which has been key to the development of various deep learning models within the field. However, there are still challenges to overcome with regards to modelling the properties of a protein, especially when labelled resources are scarce. Developing interpretable deep learning models is an essential criterion, as proteomics research requires methods to understand the functional properties of proteins. The ability to derive quality information from both the model and the data will play a vital role in the advancement of proteomics research. In this paper, we seek to leverage a BERT model that has been pre-trained on a vast quantity of proteomic data, to model a collection of regression tasks using only a minimal amount of data. We adopt a triplet network structure to fine-tune the BERT model for each dataset and evaluate its performance on a set of downstream task predictions: plasma membrane localisation, thermostability, peak absorption wavelength, and enantioselectivity. Our results significantly improve upon the original BERT baseline as well as the previous state-of-the-art models for each task, demonstrating the benefits of using a triplet network for refining such a large pre-trained model on a limited dataset. As a form of white-box deep learning, we also visualise how the model attends to specific parts of the protein and how the model detects critical modifications that change its overall function.

## 1 INTRODUCTION

The landscape of bioinformatics research has been transformed with modern sequencing technology, as more data is being collected and refined in laboratories across the world. These vast resources have allowed machine learning and more recently, deep learning to surge in popularity within this field. Deep learning models such as deep neural networks (DNN) are now standard within the area as they can handle large datasets, require minimal feature engineering and are capable of handling complex relationships within the data. This has allowed DNNs to become state-of-the-art architectures for modelling tasks within genomics (Alipanahi et al., 2015; Angermueller et al., 2016; Lanchantin et al., 2017), transcriptomics (Leung et al., 2014; Lee & Yoon, 2015; Zhang et al., 2016), and proteomics (Hou et al., 2017; Klausen et al., 2019; Yang et al., 2018a). Deep learning has proven that it can model a variety of complex processes within biology, as these models provide predictions without any explicit knowledge of the specific physical and biological mechanisms. However, a substantial amount of labelled data is usually required during the development stages. These resources are often not available for certain protein design and engineering tasks, which is inconvenient when modelling critical properties within a protein (Yang et al., 2018b; 2019; Wu et al., 2019). Computational bioinformatics and protein modelling require new approaches to develop robust deep learning models that can combat the lack of labelled data. A majority of the deep learning techniques applied in bioinformatics research originate from applications within image classification (Krizhevsky et al., 2012; He et al., 2016; Ren et al., 2015) and language modelling (Peters et al., 2018; Vaswani et al., 2017; Devlin et al., 2018). State-of-the-art approaches in the fields of natural language processing (NLP) (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018), and computer vision (CV) (Koch et al., 2015; Vinyals et al., 2016; Snell et al., 2017), now commonly employ a technique known as *pre-training*. These methods have revealed that DNNs can still retain their performance as these methods produce robust models with a limited number of training examples. However, such methods have yet to be rigorously tested within the context of protein modelling.

The method of pre-training requires a deep learning model to be trained on a separate task before being fine-tuned to a different dataset (Howard & Ruder, 2018). The utility of pre-training was first demonstrated within the field of the computer vision (Deng et al., 2009; Yosinski et al., 2014), as large convolutional neural networks were initially trained on vast image datasets, before being fine-tuned to specific tasks (Krizhevsky et al., 2012; Szegedy et al., 2017; Simonyan & Zisserman, 2014). In NLP, state-of-the-art language models use vast corpora of text to perform unsupervised or self-supervised pre-training (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018). Recent approaches within protein sequences analysing have employed similar methods during training (Yang et al., 2018b; Rao et al., 2019; Min et al., 2019). However, pre-training is still costly and time-consuming to perform as it requires a considerable amount of computational resources, and so has had a slow adoption rate within computational biology.

Computer vision was again at the forefront of modern machine learning with the application of *metric-learning* for modelling limited datasets. Whereby the original inputs to a DNN are transformed into a feature space that can be used to compare and match examples based on a distance metric (i.e. Euclidean distance, cosine-distance) (Weinberger & Saul, 2009; Hu et al., 2015; Lu et al., 2015). Examples of such deep metric-learning include the use of siamese networks (Koch et al., 2015), triplet-networks (Dong & Shen, 2018), and matching networks (Vinyals et al., 2016). In this work, we aim to determine if both pre-training and metric learning can be implemented simultaneously to develop a deep learning model that is suitable for modern protein sequence analysis.

In this paper, we will consider a pre-trained BERT model (Rao et al., 2019), which is based on a large corpus of unlabeled protein sequences with the goal of re-purposing this model by fine-tuning it using a triplet network for a set of downstream tasks. During training, triplets (i.e. anchor, positive and negative) of the protein sequences will be used along with weight-sharing within the BERT model to cluster the data based on a triplet loss (Hoffer & Ailon, 2015). The BERT model will be used to produce a vector representation for the anchor, positive and negative protein respectfully. During training, a protein is considered to be a positive example to the anchoring protein if its labelled value (i.e. measured property) is closer in absolute value to the anchor's label when compared to the negative instance. Throughout the tuning process, new triplets are formed as the BERT model undergoes semi-supervised training, and begins clustering individual cases within the dataset. As outlined in Figure 1 (e.g. thermostability T50 values), our approach must fine-tune the BERT model to incorporate information about the measured property. Once the model is optimised to the task, the encodings will provide information that reflects expected measured property. Our approach builds on the work of Rao et al. (2019) by tailoring their pre-trained BERT model through the use of metric learning for protein analysis. In doing so, the model should produce an improved feature set for each task without overfitting to the limited labelled examples observed during training.

## 2 Related Work

Determining the critical properties of a protein is one of the most challenging aspects of any downstream task. Traditionally, many of these properties are discovered by examining the physical structure of the protein. However, this is often a very time-consuming and expensive process. Another option is to encode each amino acid with a basic set of physical properties (e.g. its charge or hydrophobicity). Inevitably many physical properties can be missed or poorly represented by such feature engineering, which then leads to overfitting and inadequately modelling of the downstream task. Many have considered encoding the primary structure of the protein (i.e. the sequence of amino acids) (Jaganathan et al., 2019; Sun et al., 2017; Wei et al., 2018), where a vector of real numbers represents each amino acid, and are optimised in a deep learning model. Since deep learning is often used to avoid feature engineering, these model can capture sophisticated features by analysing the original sequence of amino acids. Both supervised and unsupervised deep learning has been applied in proteomic research (Alipanahi et al., 2015; Hochreiter et al., 2007; Almagro Armenteros et al., 2017). As vast resources of proteomic data become available (e.g. the UniProt database (Consortium, 2014)), it provides a third option to perform semi-supervised deep learning. This could be a vital step forward for proteomic research as it is a far less expensive and time-consuming alternative. Rao et al. (2019) displayed the potential semi-supervised deep learning for protein sequence analysis as they introduced the Tasks Assessing Protein Embeddings (TAPE). They benchmarked the current state-of-the-art models to a set of five biologically relevant semi-supervised learning tasks spread across different domains of protein biology. This included a Transformer model (Vaswani

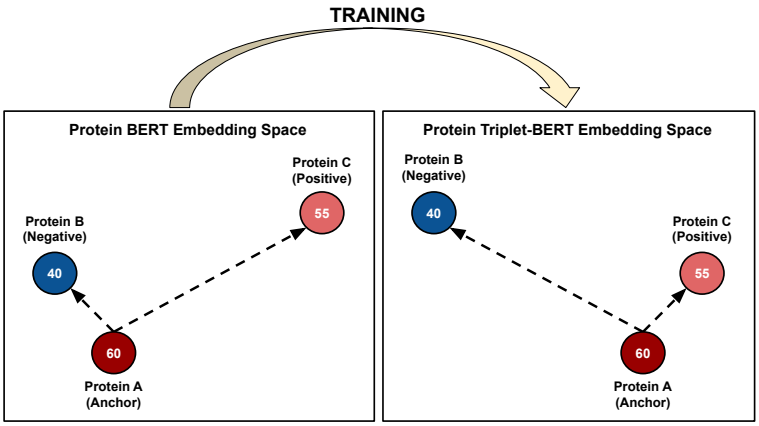



Figure 1: Triplet loss, designed to fine-tune the original pre-trained embeddings produced by the BERT model to a downstream task. During training, a triplet of proteins our organised by calculating the Euclidean distance between each example. Once completed, proteins with similar property values (i.e. the values in the circles) should cluster closer together, producing a more meaningful embedding space.

et al., 2017), a ResNet model (Yu et al., 2017), and a long short-term memory (LSTM) model (Hochreiter & Schmidhuber, 1997) of their own design. In addition to these three models, they also benchmarked two previously proposed architectures; another a bidirectional LSTM model (Bepler & Berger, 2019), and a unidirectional mLSTM (Krause et al., 2016; Alley et al., 2019). Rao et al. (2019) experiments concluded with the transformer model outperforming every other model tested concerning its accuracy, perplexity and exponentiated cross-entropy. These results were not surprising as transformer-style architectures have quickly become the new standard for many NLP tasks (Vaswani et al., 2017; Devlin et al., 2018; Dai et al., 2019).

Given that Rao et al. (2019) has shown that both a transformer is the best candidate for modelling a protein sequence, we aim to build upon their work by testing this pre-trained model on four downstream tasks introduced by Yang et al. (2018b). In this way, we aim to investigate whether such pre-trained embeddings can help predict relevant properties for a set of downstream tasks. We aim to improve this pre-trained model with the use of a triplet style network to fine-tune the model to incorporate additional relevant information about the protein for each specific downstream task. A drawback to Rao et al. (2019) work was that they only tested their transformer network using a character-based encoding. However, state-of-the-art transformer language models now commonly use a subword encoding algorithm before embeddings a sentence (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018). Subword algorithms such as byte-pair encoding algorithm (BPE) (Gage, 1994) or unigram encoding algorithm (Kudo, 2018) can provide more extensive vocabulary during training. These encoding algorithms also have the added benefit of reducing the length of the input sequence. This could also reduce the time and cost required to model protein sequences without using excessive amounts of padding.

In extension to this work, Vig et al. (2020) explored how this BERT model (Rao et al., 2019) was capable of discerning structural and functional properties about the protein. Vig et al. (2020) proved that the model was able to model long-range dependencies within the sequence of amino acids, it was also able to deduce information about the protein based on the folding structure, target binding sites, and additional complex biophysical properties. They concluded that the specific heads within the model attended to individual amino acids, as the attention similarity matrix was positively correlated to the expected substitution scores (i.e. BLOSUM62) for each amino acid. Vig et al. (2020) noted that the deeper layers of the BERT model focused relatively more attention on binding sites and contacts (i.e. high-level concepts). In contrast, information about the secondary structure (i.e. low- to mid-level concepts) within the protein was targeted evenly across each of the layers.

Another example of semi-supervised was by Yang et al. (2018b) as they applied both pre-training and an n-gram encoding strategy to analyse a set of proteins. Their approach consisted of using a

tri-gram encoding to each protein sequence analysing the set of tri-grams with a Doc2Vec model (Le & Mikolov, 2014). The method encodes the protein an in trivial fashion, which makes it susceptible to poorly represented (i.e. infrequent) tri-grams that can later affect pre-training. Lennox et al. (2020b) improved on this work by testing the use of subword algorithms on protein sequences. Both approaches are unfavourable as they implement Doc2Vec models, which return a single vector representation for the entire protein. This makes it difficult to interpret each vector representation when querying specific modifications in the protein.

Metric learning is still uncommon within computational biology deep learning even though it has become standard practice in computer vision (Koch et al., 2015; Dong & Shen, 2018; Vinyals et al., 2016). There have been many improvements to deep metric learning from its introduction with the siamese style networks (Bromley et al., 1994; Chopra et al., 2005; Hadsell et al., 2006). One of the most notable instances was by Hoffer & Ailon (2015) with the triplet network. In this work, we aim to use such a triplet style training procedure to improve the encodings produced by the BERT model (Rao et al., 2019). As we extract embedded representations for four protein property prediction tasks (Yang et al., 2018b) using the pre-trained BERT model (Rao et al., 2019), and then fine-tune our BERT model to each task. The tasked covered in this investigation contain proteins from various families and library designs that were not included in Rao et al. (2019) work. We show that the predictive power of models trained using these embeddings exceeds those trained on the previous state-of-the-art methods. This approach can be an accurate and efficient alternative as it does not require alignment or any additional structural data about the protein. A series of visualisation techniques will be used to present the critical relationships with the data, and how the BERT model attends to specific amino acids in the protein.

## 3 Materials and Methods

Previous applications of pre-training (Yang et al., 2018b; Lennox et al., 2020b) and deep metric learning (Lennox et al., 2020a) have shown a clear benefit to applying either technique to analyse protein data. One key drawback to these approaches is the limited window sizes to which they encode segments of the protein. This can be detrimental to the model's performance as it is unable to capture long-range dependencies within the protein and therefore encodes less information about the proteins final structure. Our approach improves upon these examples by using a BERT style model that is capable of encoding the complete protein in a bidirectional fashion. Past work has justified the importance of either pre-training the model or using metric learning. There is still room for improvement by bringing both approaches together by utilising a state-of-the-art pre-trained network than has been fine-tuned using a triplet style network. Since the BERT has not been set up to handle a subword encoding (e.g. BPE or Unigram), we aim instead to set a stable baseline for the application of both pre-training and deep metric learning that will only use a character-based encoding. However, Lennox et al. (2020b) has shown the potential subword encoding algorithms have in improving and simplifying the pre-training process with a Doc2Vec model. Such a strategy could be beneficial first to encode the proteins before being analysed by the BERT model.

### 3.1 Modelling

The Triplet-BERT network employed in this investigation is outlined in Figure 2. For each task, the proteins are encoded a BERT model (Rao et al., 2019), which has been re-trained on a set of protein sequences used in TAPE investigation. These proteins were collected from the recently curated Pfam database (El-Gebali et al., 2018), which holds approximately thirty-one million protein domains, and forms the corpus used to train large sequence models as featured in TAPE (Rao et al., 2019). The architects of the Pfam database have organised the proteins into clusters that share evolutionary-related groups, also known as families. In summary, the BERT model consists of 12-layers with a hidden size of 512 units and eight attention heads, leading to a $\sim 38M$ - parameter model, and was trained using masked-token prediction (Devlin et al., 2018). Every layer of the BERT model is frozen except for the last layer, which will allow the model to be easily tuned to each task. The features produced by the final layer of the model are then pooled to form the vector representation for each protein. Initially the model will encode a triplet of proteins, $(x_a, x_p, x_n)$, whereby $x_a$, $x_p$, and $x_n$ denote the anchor, positive and negative proteins respectively. The BERT model will then outputs the following:

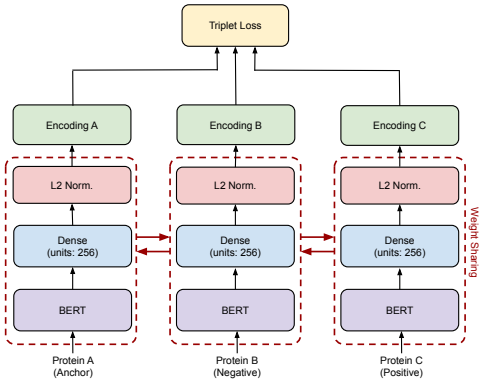



Figure 2: Overview of the Triplet-BERT Approach.

$$\begin{aligned} a &= f(x_a) \\ p &= f(x_p) \\ n &= f(x_n) \end{aligned} \tag{1}$$

$$\begin{aligned} D(a,p) &= \|a - p\| \\ D(a,n) &= \|a - n\| \end{aligned} \tag{2}$$

$$L(a,p,n) = \frac{1}{2}\{\max(0, m + D(a,p) - D(a,n)\}. \tag{3}$$

In this example, we are applying inter-domain learning as the weights of the BERT model are shared. The model is represented by the encoding function $f$, which is applied to each branch of the triplet network. Once the BERT model has encoded the triplet, it is then passed through one final dense and L2-normalisation layer respectfully (Schroff et al., 2015). The triplet of encodings can then be used to train the BERT to rank the triplet based on the anchoring protein via the triplet-loss (Equation. 3), as visualised in Figure 2. Once the BERT model has been fine-tuned, we will use its final generated encodings to build a simple regression model to predict the given properties of each task. As in past examples, Gaussian process (GP) regression models (Matérn kernels with $\nu = 5/2$) (Rasmussen, 2003) will be used to model the properties as to remain unbiased in this investigation.

### 3.2 Data

To thoroughly evaluate the performance of using our approach, we included four downstream tasks in this investigation that cover a range of potential properties in which deep learning could be applied. As we are mainly interested in the modelling capabilities of our approach and not the data itself, we will only briefly review each resource. The peak absorption wavelength dataset comprises of the Gloeobacter violaceus rhodopsin (GR) parent protein and an additional 80 protein sequences (1–5 mutations) (Engqvist et al., 2015). The enantioselectivity dataset contains the epoxide hydrolase (EH) parent protein and a further 151 protein sequences (1–8 mutations) (Zaugg et al., 2017). The plasma membrane localisation dataset includes three-parent proteins and an additional 248 protein sequences (1–108 mutations) (Bedbrook et al., 2017). Finally, the thermostability (T50) dataset contains three-parent proteins and an additional 261 protein sequences (1–109 mutations) (Romero et al., 2013). Please see citations for a more in-depth explanation into the nature of each property and the methods used to collect the data.

## 4 Results and Discussion

In this study, we began by evaluating the original BERT model as an example of a pre-training strategy. This model was used as a baseline to our investigation on a set of downstream tasks. We

Table 1: Results for the four protein downstream tasks.

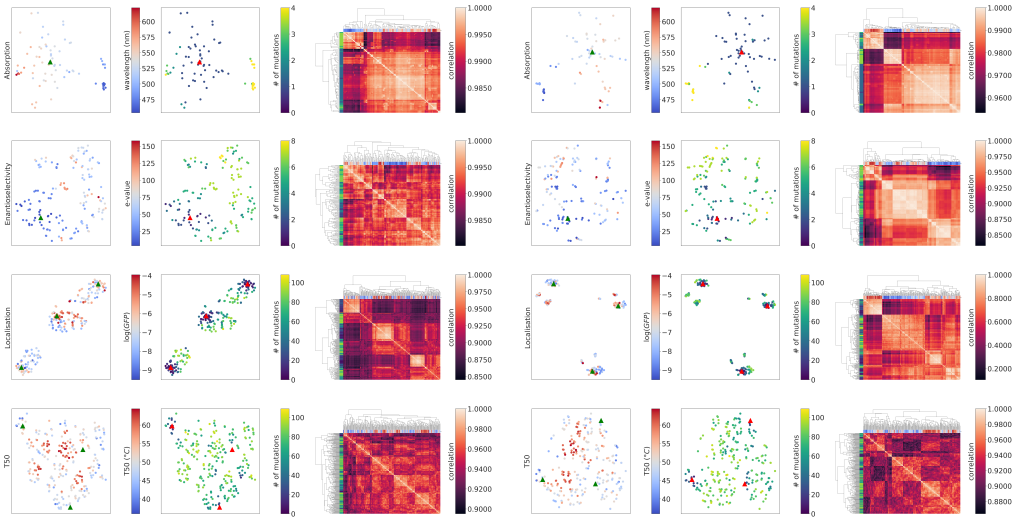
| Model | Encoding | Vocabulary Size | Absorption | Enantioselectivity | Localisation | T50 |
|---|---|---|---|---|---|---|
| BERT (Triplet) (Ours) | Character | 20 | **14.06** | **3.85** | **0.50** | **2.36** |
| BERT (Non-Triplet) (Rao et al., 2019) | Character | 20 | 16.57 | 7.57 | 0.70 | 2.47 |
| CNN (Triplet) (Lennox et al., 2020a) | Character | 20 | 17.14 | 5.93 | 0.63 | 2.58 |
| CNN (Non-Triplet) (Lennox et al., 2020a) | Character | 20 | 25.28 | 8.01 | 0.67 | 3.32 |
| Doc2Vec (Lennox et al., 2020b) | Unigram | 2000 | 26.41 | 6.77 | 0.65 | 2.98 |
| | | 4000 | 18.09 | 6.90 | 0.76 | 2.80 |
| | | 8000 | 20.92 | 8.58 | 0.86 | 2.59 |
| | | 16000 | 24.05 | 7.07 | 0.77 | 3.33 |
| | | 32000 | 21.98 | 9.53 | 0.76 | 2.96 |
| Doc2Vec (Lennox et al., 2020b) | BPE | 2000 | 23.83 | 10.38 | 0.66 | 2.70 |
| | | 4000 | 20.80 | 9.76 | 0.67 | 3.01 |
| | | 8000 | 18.46 | 6.72 | 0.75 | 2.75 |
| | | 16000 | 20.64 | 6.08 | 0.73 | 2.76 |
| | | 32000 | 24.27 | 7.03 | 0.67 | 2.80 |
| Doc2Vec (Yang et al., 2018b) | Tri-gram | 8000 | 23.30 | 9.14 | 0.73 | 2.91 |
| Doc2Vec (Lennox et al., 2020b) | Character | 20 | 46.08 | 12.55 | 0.81 | 4.32 |

then built on this approach by testing the advantages of combining both pre-training and deep metric learning to the same tasks, as shown in Table 1. Just as past studies, we adopted an eighty-twenty split of the generated triplets from the training data to train and validate the performance of the model for each dataset. Doing so provided a stable training setup during the fine-tuning stage of development. Unsurprisingly, the triplet tuned version of the BERT model easily outperformed the original pre-trained baselines along with the other examples that included both CNN and Doc2Vec based models with an improved mean absolute error (MAE) score in each task. Our results indicate that the fine-tuning stage does alter the latent space produced by the original model, and tailors it to each specific downstream task improving the final representation of each protein. The real value in applying pre-training is observed when the model can successfully encode a protein without any prior knowledge of biochemistry. Only during the pre-training stage does the model begin to learn these complex relationships between the amino acids within the protein sequence. Through proper fine-tuning can these pre-trained embeddings be improved by using deep metric learning to model subtle mutations within a set of amino acids.

The encodings produced by both strategies are visualised using a set of t-distributed stochastic neighbour embedding (t-SNE) (Maaten & Hinton, 2008) plots along with cluster maps, as shown in Figures 3a - 3b (with all t-SNE projections using a perplexity of 30) were produced for each downstream task. Figure 3a depicts the encodings of the original BERT model for each downstream task. While Figure 3b is the final encodings produced once the BERT model had been tuned using our triplet network approach. By considering the combination of both plots, it is easier to envision how the BERT model perceives each protein sequence when using either strategy and observe the contribution of each mutation in the final feature vector representation.

For the absorption task, we see how there is less of an order to the original encoding when compared to the triplet tuned counterparts, as shown in Figures 3a-3b. In Figure 3b, it is far easier to determine which modifications will have a more significant effect on the proteins absorption value as the triplet tuned encodings become more tailored to the task. Figures 3a-3b presents the model's ability to capture even the most minor modifications to the original parent protein, regardless of the length of the sequence. The high correlations observed in the cluster map in Figure 3a reflect the fact that all the modified proteins were based on one protein and indicated two main clusters within the dataset. However, in Figure 3b the cluster map based on the triplet tuned encodings provide a more detailed depiction of how the proteins are correlated to one another as we observe smaller sub-clusters within the dataset.

In the enantioselectivity task, we again see the best performance from the triplet-BERT model, as shown in Table 1. Still, both versions of the BERT model were capable of detecting any modifications present within the parent protein. However, when considering the cluster maps in Figures 3a-3b we can see the triplet tuned encodings provide more distinct clusters when compared to the originals. In Figures 3a-3b, we observe that the triplet encodings incorporate information with regards to the measured property and the modifications present within the protein. In the absorption task, we notice that an increase in the number of modifications could lead to either an increase or decrease in absorption values. However, for enantioselectivity, the more modifications that are present in the protein, the higher its expected e-value will be for this particular dataset.



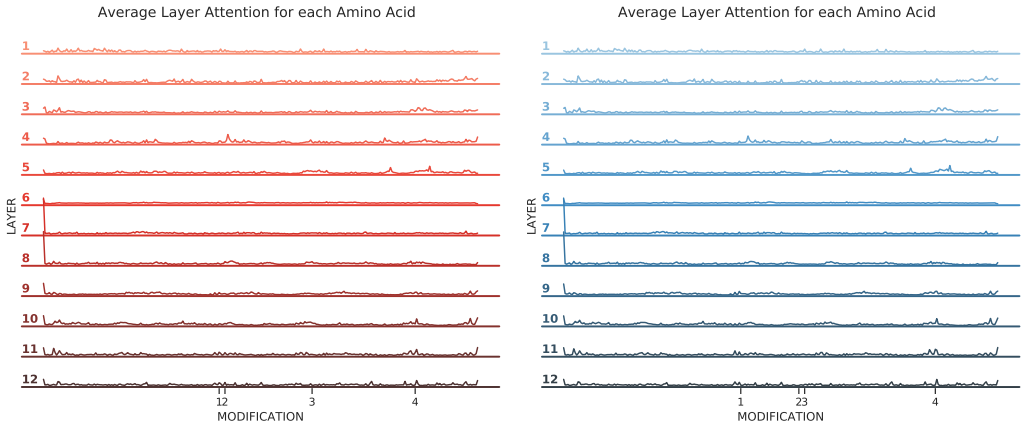
(a) Original BERT. (b) Triplet-BERT.

Figure 3: A set of t-SNE and cluster plots for both versions of the BERT model, thereby visualising the correlations within each learned embedding space (e.g. the number of modifications present and the functional property of each protein) (see text for details).

When examining the plasma membrane localisation task, we again observed the best results from the triplet-BERT encodings, as shown in Table 1. When visualising both sets of encodings in Figures 3a-3b, we can see how the BERT model easily clusters the three families tested in this specific task. Interestingly in Figure 3b, the triplet encoding clusters these families further away from one another. With smaller sub-clusters appearing for the proteins that have a higher localisation value. The number of mutations appears to have an opposite effect to that of absorption and enantioselectivity. From Figures 3a-3b, we see that the fewer mutations present within this protein leads to a higher overall localisation value. Just as in the case of the absorption and enantioselectivity, we again notice a far sharper cutoff between suspected groups within the data when using the triplet encodings for Figure 3b. In considering the cluster maps in Figures 3a-3b, it becomes easier to recognise which parent protein is more or less receptive to the task as the model becomes better at detecting the relevant modifications.

In the final task, we again observed the triplet-BERT model producing the best encodings for modelling thermostability (i.e. T50) values, as outlined in Table 1. When we visualise the encodings produced by each strategy in Figures 3a-3b, the proteins that possess the highest T50 values are clustered into the centre of each plot. Unlike in the localisation task, the encodings produced by the BERT model for the thermostability task are not initially separated into three distinct clusters based on the parent proteins. Instead, we see a series of smaller groups with most of the proteins with high thermostability values congregating in the centre. Similarly to absorption and enantioselectivity, the number of mutations present in the protein is positively correlated to the thermostability value. Similarly to the localisation task, when we consider the cluster maps in Figures 3a-3b, we can see that triplet-BERT model is better at clustering the proteins with higher thermostability values together.

To reinforce the utility and interpretability of this approach, we have also included a set of Figures 5-8 that focus on a few examples from the peak absorption task. In Figure 5, we have mapped the attention weights of the final layer onto the parent protein and two mutated (i.e. the most and least absorbent) versions of this protein. In Figure 5, we can see which what parts of the protein and specific mutations (i.e. red lettering) contribute the most to the final vector representation. Moreover, we consider the average attention in each head of the final layer for these modified proteins in Figures 6a-6b. These figures illustrate the complexity of this model as no two heads attend to the same parts of the proteins. In Figures 4a-4b, by taking an average over each head of the BERT model, can we ascertain the critical parts of the protein within each layer. Likewise, In Figures 7a-7b, by taking an average over each layer of the BERT model, can we observe similar patterns within each head.




(a) Individual Layers

(b) Individual Layers

Figure 4: Average Attentions within the BERT model. The least and most absorbent mutated versions of the parent protein are coloured blue and red respectfully.

Finally, we compare the parent protein to both of its modified counterparts in Figure 8, by taking the absolute difference in attention for the final layer. In Figure 8a, we can discern that mutations one and four play a significant role in modifying the function of the original protein. While in Figure 8b, we can see that it is mutations two and four respectfully. In both cases, these particular modifications cause the model to attend to the surrounding amino acids within the protein, thereby altering the final vector representation.

## 5 Conclusion

In this work, we have illustrated how pre-training can be utilised for robust modelling of a protein's functional properties, and with some additional fine-tuning through the use of a triplet-network, these models can be further improved. From the results, the triplet-BERT network produced more detailed encodings in each downstream task when compared to the original pre-trained BERT encodings and previous baselines. When using both strategies of pre-training and metric learning, we observed state-of-the-art results for all downstream tasks when compared to using just one of these approaches. This investigation has shown that deep learning can still be applied to produce state-of-art regardless of the limited number of examples within the dataset. More specifically, we have highlighted the potential for pre-training and metric learning within the field of proteomics. By visualising the intermediate features generated by the BERT model, we also provided insight into the function of a protein as we measured the impact of specific modifications featured in all four downstream tasks.

As modern sequencing technology continues to improve proteomics to provide more data on the properties of a protein, it will become paramount to link these extensive resources to specific tasks through the use of techniques such as pre-training and metric learning. In future work, we postulate that subword encodings could improve the encodings generated during pre-training by the BERT model. This will allow the network to model the learn a far more substantial vocabulary for each protein and will reduce the overall sequence length of the protein, which in turn will reduce the time and cost required to perform pre-training.

## References

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33 (8):831–838, 2015.

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

pp. 589333, 2019.

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

Christof Angermueller, Heather J Lee, Wolf Reik, and Oliver Stegle. Accurate prediction of single-cell dna methylation states using deep learning. *BioRxiv*, pp. 055715, 2016.

Claire N Bedbrook, Austin J Rice, Kevin K Yang, Xiaozhe Ding, Siyuan Chen, Emily M LeProust, Viviana Gradinaru, and Frances H Arnold. Structure-guided schema recombination generates diverse chimeric channelrhodopsins. *Proceedings of the National Academy of Sciences*, 114(13): E2624–E2633, 2017.

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

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pp. 737–744, 1994.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546. IEEE, 2005.

UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1): D204–D212, 2014.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 459–474, 2018.

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, et al. The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2018.

Martin KM Engqvist, R Scott McIsaac, Peter Dollinger, Nicholas C Flytzanis, Michael Abrams, Stanford Schor, and Frances H Arnold. Directed evolution of gloeobacter violaceus rhodopsin spectral properties. *Journal of molecular biology*, 427(1):205–220, 2015.

Philip Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Sepp Hochreiter, Martin Heusel, and Klaus Obermayer. Fast model-based protein homology detection without alignment. *Bioinformatics*, 23(14):1728–1736, 2007.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pp. 84–92. Springer, 2015.

Jie Hou, Badri Adhikari, and Jianlin Cheng. Deepsf: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 34(8):1295–1303, 2017.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Deep metric learning for visual tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(11):2056–2068, 2015.

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3): 535–548, 2019.

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*, 87(6): 520–527, 2019.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

Ben Krause, Liang Lu, Iain Murray, and Steve Renals. Multiplicative lstm for sequence modelling. *arXiv preprint arXiv:1609.07959*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*, 2018.

Jack Lanchantin, Ritambhara Singh, Beilun Wang, and Yanjun Qi. Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks. In *Pacific Symposium on Biocomputing 2017*, pp. 254–265. World Scientific, 2017.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196, 2014.

Taehoon Lee and Sungroh Yoon. Boosted categorical restricted boltzmann machine for computational prediction of splice junctions. In *International Conference on Machine Learning*, pp. 2483–2492, 2015.

Mark Lennox, Neil Robertson, and Barry Devereux. Deep metric learning for proteomics. *Submitted to: The International Conference on Machine Learning and Applications*, 2020a.

Mark Lennox, Neil Robertson, and Barry Devereux. Expanding the vocabulary of a protein: Application of subword algorithms to protein sequence modelling. *The Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2020b.

Michael KK Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–i129, 2014.

Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou. Multi-manifold deep metric learning for image set classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1137–1145, 2015.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Seonwoo Min, Seunghyun Park, Siwon Kim, Hyun-Soo Choi, and Sungroh Yoon. Pre-training of deep bidirectional protein sequence representations with structural information. *arXiv preprint arXiv:1912.05625*, 2019.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. *arXiv preprint arXiv:1906.08230*, 2019.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pp. 63–71. Springer, 2003.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Philip A Romero, Andreas Krause, and Frances H Arnold. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3):E193–E201, 2013.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Tanlin Sun, Bo Zhou, Luhua Lai, and Jianfeng Pei. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC bioinformatics*, 18(1):277, 2017.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

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*, pp. 5998–6008, 2017.

Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Leyi Wei, Yijie Ding, Ran Su, Jijun Tang, and Quan Zou. Prediction of human protein subcellular localization using deep learning. *Journal of Parallel and Distributed Computing*, 117:212–217, 2018.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

Zachary Wu, SB Jennifer Kan, Russell D Lewis, Bruce J Wittmann, and Frances H Arnold. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences*, 116(18):8852–8858, 2019.

Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine learning in protein engineering. *arXiv preprint arXiv:1811.10775*, 2018a.

Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, 2018b.

Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328, 2014.

Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 472–480, 2017.

Julian Zaugg, Yosephine Gumulya, Alpeshkumar K Malde, and Mikael Bodén. Learning epistatic interactions from sequence-activity data to predict enantioselectivity. *Journal of computer-aided molecular design*, 31(12):1085–1096, 2017.

Sai Zhang, Jingtian Zhou, Hailin Hu, Haipeng Gong, Ligong Chen, Chao Cheng, and Jianyang Zeng. A deep learning framework for modeling structural features of rna-binding protein targets. *Nucleic acids research*, 44(4):e32–e32, 2016.

# A APPENDIX

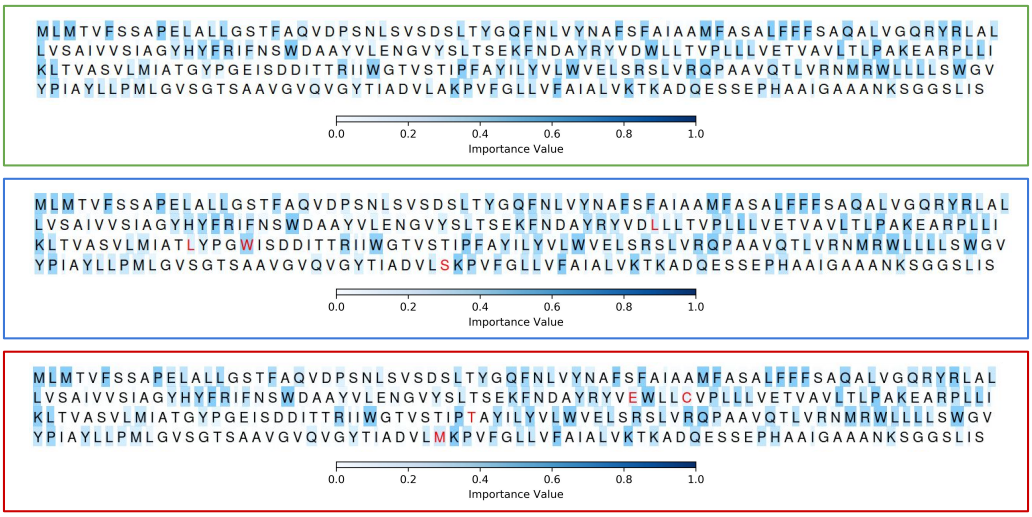
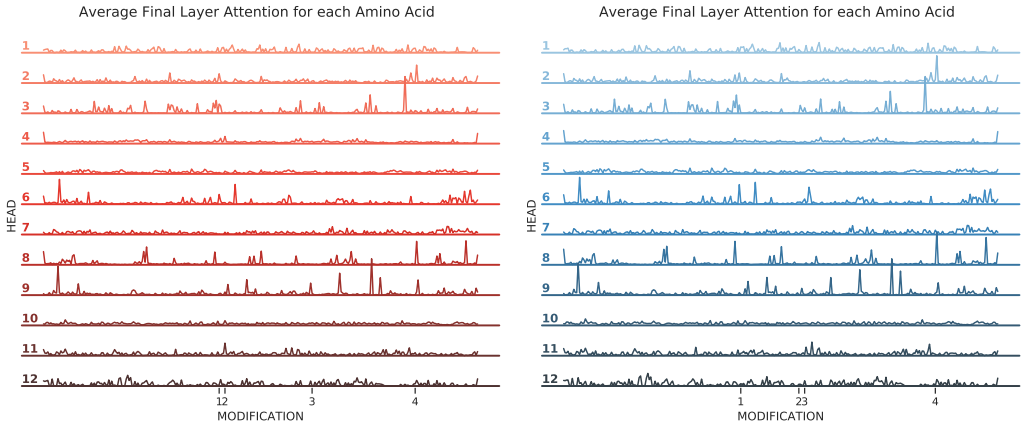


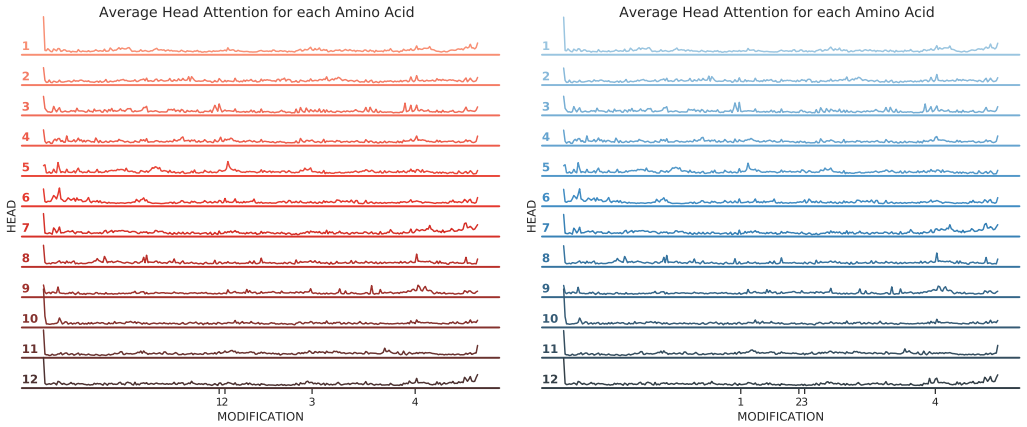

Figure 5: A set of attention maps highlighting the importance of each amino-acid with three proteins from the peak absorption task. The parent protein is outlined in green, and its least and most absorbent versions are outlined in blue and red respectfully. Any modifications are represented by red lettering.




(a) Final Layer

(b) Final Layer

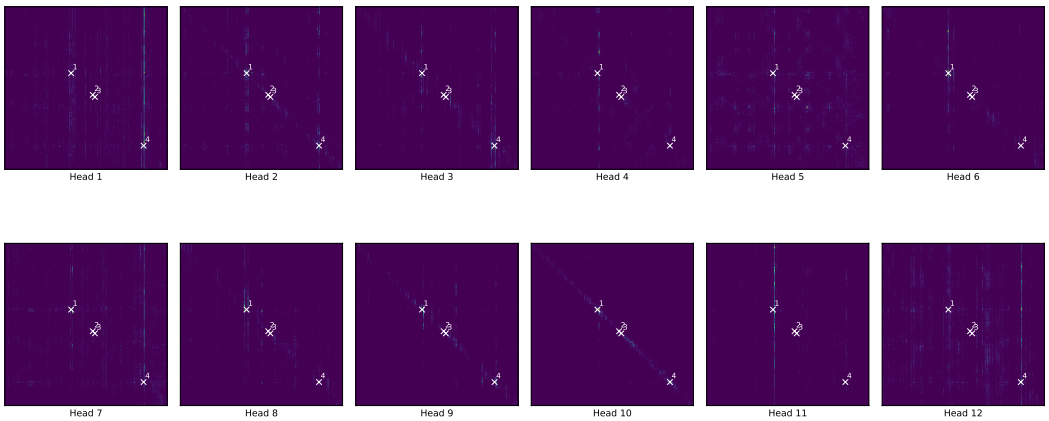
Figure 6: Average Attentions within the BERT model. The least and most absorbent mutated versions of the parent protein are coloured blue and red respectfully.



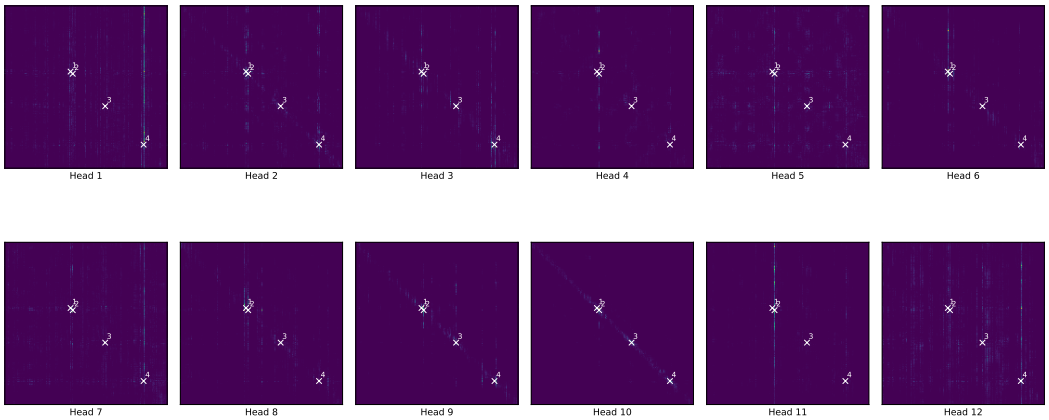

(a) Individual Heads

(b) Individual Heads

Figure 7: Average Attentions within the BERT model. The least and most absorbent mutated versions of the parent protein are coloured blue and red respectfully.




(a) Least absorbent mutated version.




(b) Most absorbent mutated version.

Figure 8: The absolute difference in attention between a protein and its parent for each head of the final layer. The modification are highlighted by the white crosses.