Predicting Protein B-factors from Amino Acid Sequence with Deep Neural Networks

Marco Carbullido

Tulane University

Abstract

Despite significant advancements in protein structure prediction, understanding structural dynamics remains challenging. This exploration of protein B-factor prediction from amino acid sequences using deep learning models, focuses on the application of transformer architectures and learned protein embeddings. These experiments reach a Pearson Correlation Coefficient of 0.815, being competitive with other state-of-the-art approaches.

Keywords: B-factor

Predicting Protein B-factors from Amino-Acid Sequence with Deep Neural Networks

The conformational dynamics of proteins are responsible for their biological functions. In recent years, while deep learning has enabled researchers to predict protein structures with high accuracy, the prediction of structural dynamics still remains a challenge. This means that while models like AlphaFold2 have enabled scientists to accurately predict the three-dimensional shapes of millions of proteins just from their amino acid sequence, these models are not trained to provide any insight into the structural changes that they might undergo, essential to understanding the mechanisms function. Experimentally, protein structure determination yields *temperature factors*, also known as *b-factors*, for each atom, which are proportional to their mean displacement from their average position (see Figure 1). B-factors give insight into which parts of the enzyme are more flexible and might destabilize at high temperatures. For example, scientists can target these regions for mutation and optimize their stability to enhance enzyme performance (See Citation 1). It has been shown previously that neighboring atoms have an effect on b-factors (Pandey et al.). However, it is important to remember that atoms close to one another in 3D-space do not necessarily come from adjacent regions in the protein sequence.

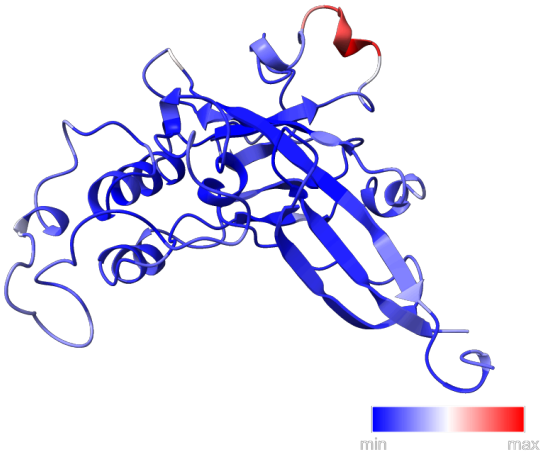

Figure 1: Experimentally determined 3D structure of 1AOL, colored by b-factor

This semester, our aim was to test the performance of various approaches to predicting protein b-factors. Given the long-range dependencies inherent to this prediction task, we hypothesized that implementing self-attention would outperform other methods such as LSTM. In addition to evaluating

various architectures, we found that training our models on protein embeddings learned through an autoencoding task was more effective and led to faster model convergence when compared to the raw amino acid sequences. Our experiments seem to confirm our hypothesis by showing that transformer-based approaches achieved the lowest losses, with our best model achieving state-of-the-art performance with a Pearson Correlation Coefficient (PCC) of 0.815. These results are in agreement with the general notion that self-attention mechanisms improve the modeling of long-range dependencies in sequences, extending to those at play in protein folding.

Background

A common metric used to evaluate the quality of b-factor predictions is the PCC, with the highest reported average test PCC reaching 0.8 utilizing an LSTM architecture, outperforming vanilla RNNs due to the vanishing gradients problem observed in longer sequences (see Figure 2, Pandey et. al). Because

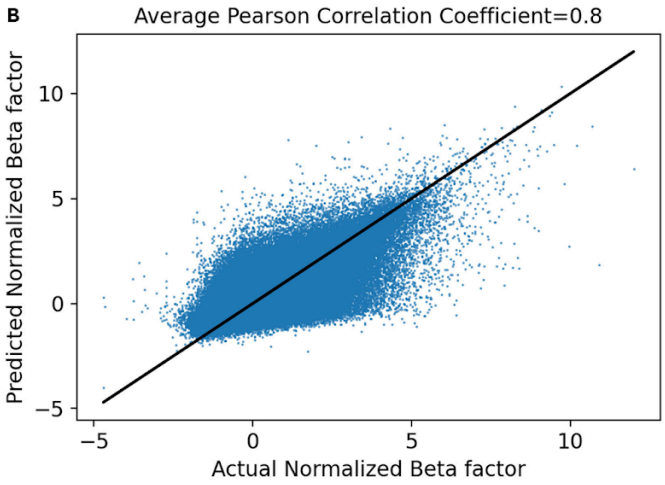

Figure 2: Results from reproducing the LSTM in the mentioned paper: the training dataset was approximately 61,000 sequences and the testing dataset was 2,400. The model reached an average test PCC of 0.8

the authors provide the code and dataset used for training, we both implemented their model to benchmark our experiments and used their dataset as outlined in the Experiments section. Previous approaches to predict protein b-factors have shown varying levels of success due to this problem. It has been reported that the residue-level confidence score produced for structures predicted from sequence by AlphaFold2 is inversely correlated with experimental b-factors (Guo, HB.). Intuitively, this can be understood in that it may be harder to predict positions of atoms with higher average displacements. In addition to uniquely supervised approaches, the prediction of biophysical properties such as b-factors from protein sequence has been shown to be more effective when the input embeddings are learned from various models trained in a self-supervised fashion (see Citation 3).

Methods

As noted earlier, long-range dependencies in protein sequences become an issue especially for classic sequence models like RNNs. The LSTM architecture is a recurrent architecture for sequences that contains cells with an input gate, a forget gate, and an output gate. While this architecture leads to better performance on some tasks when compared to RNNs, it still struggles with increasingly long sequences. As this would suggest, our most successful architectures utilized *scaled dot product attention*, which proved valuable through their ability to effectively learn interactions across the entire input sequence. This operation involved projecting each sequence of amino acids (or amino acid embeddings) into a set of keys, queries and values through learned linear transformations (see Figure 3). While the value vectors represent the information contributed by each token, the query and key vectors are used to weight that contribution. The sum of the value vectors weighted by the dot product of a given token's query vector and the key vectors imbues that token with contextual information from the entire sequence. These dot products are generally normalized by dividing by the square root of the key dimension before being passed to the softmax function to ensure they sum to one, hence the word "scaled."

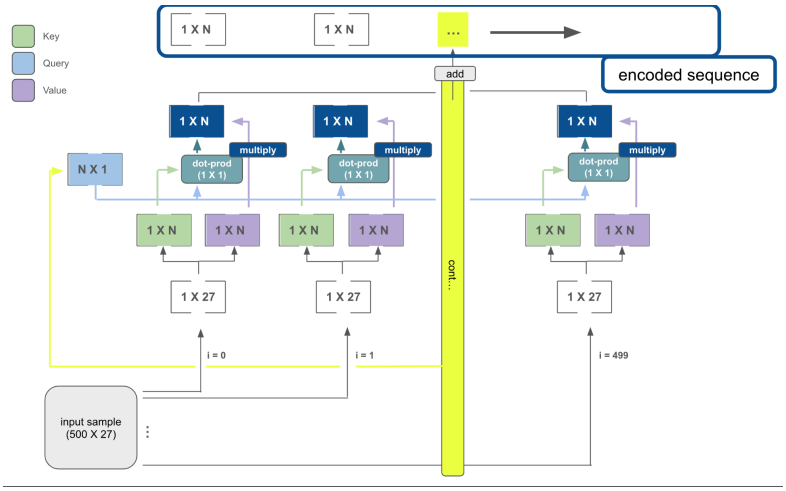

Figure 3: Simplified diagram of the operations performed on input tokens during encoding.

Also relevant to this discussion is the use of learned protein embeddings. As described earlier, we wanted to experiment with pre-training a model in a self-supervised fashion to reconstruct full sequences from corrupted versions.. These embeddings served as a robust starting point for our prediction task as they provided more contextualized representations than the original raw sequences.

Experiments

To conduct our experiments, we used the dataset published by the previously mentioned authors, which they report to be "the widest dataset on which a model for the prediction of the B factor has been tested" (Pandey et al.). This dataset aggregates over 61,000 protein sequences, structures and b-factors from the Protein Data Bank, a publicly accessible biological database, ignoring proteins larger than 500 amino acids. To tokenize sequences, each amino acid's one-hot-encoding is concatenated with a one-hot-encoding for the secondary structure and the cartesian coordinates of its backbone carbon atom. With a dictionary of 21 possible amino acids, 3 possible secondary structures, and 3 coordinates, this yields input tensors of size 500 x 28. Lastly, proteins that contained non-positive b-factors or b-factors greater than 80 were also omitted from this dataset.

The first stage of our experiment involved using the original dataset used to train the LSTM model reporting a state-of-the-art average test PCC of 0.8 (see Figure 2) on 2,442 sequences. After successfully reproducing these results, we decided to train a transformer architecture on this same dataset in order to compare our approaches to a baseline. We tested several configurations, and found that the average test PCC increased as we increased the key dimension of our model. Compared to LSTM, we found that the self-attention mechanism led to more accurate predictions and lower training error (see Figure 8). Our best model achieved a test PCC of 0.815 across our benchmark dataset of 2,442 proteins. We found that the performance improved with the use of the embeddings learned from the auto-encoding task.

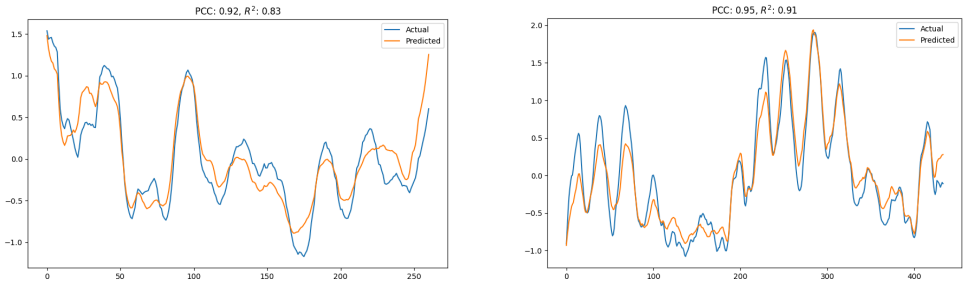

Figure 7: Sampled test predictions from the highest performing architecture

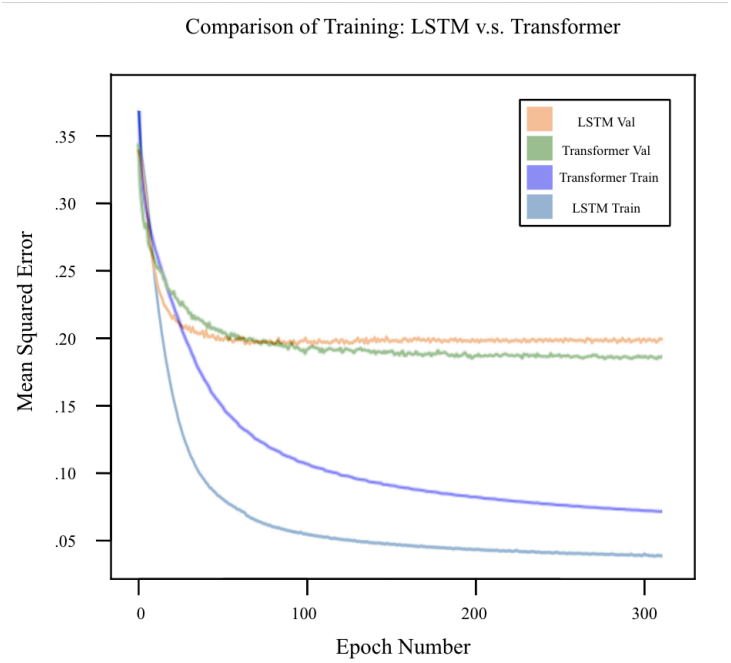

Figure 8: The learned protein embeddings produced faster model convergence and a higher PCC

Discussion

This observation leads us to a few promising future areas of research: first, since our model performs near perfect results on many proteins, we could train a specialized model made to predict proteins in which its performance is high. Furthermore, if there is some commonality between the proteins in which performance is poor—such as inaccurate data or data structured in a different way which is throwing off the model—this could be observed and addressed. Further research is necessary to see if these hypotheses could be true. In the future, we would like to continue investigating this task with more resources. Separately, because the training tokens contain both multi-hot encoded and floating point values, we also note that it is possible that a better representation could be formed by splitting sequence and structure information into two streams.

Conclusion

We demonstrate significant progress in predicting protein B-Factors using deep learning approaches through the implementation of transformer architectures and protein embeddings. Our experiments demonstrated the importance of an encoding architecture in improving prediction accuracy. While some proteins' predicted b-factors were highly accurate, others demonstrated poorer results. Our model ultimately performed on par with other SOTA approaches.

References

Natália Gonçalves Ramos, Gabriel Fonseca Sarmanho, Fernando de Sá Ribeiro, Vanderléa de Souza, Luís Maurício T.R. Lima, The reproducible normality of the crystallographic B-factor, Analytical Biochemistry, Volume 645, 2022, 114594, ISSN 0003-2697, https://doi.org/10.1016/j.ab.2022.114594.

Guo, HB., Perminov, A., Bekele, S. et al. AlphaFold2 models indicate that protein sequence determines both structure and dynamics. Sci Rep 12, 10696 (2022). https://doi.org/10.1038/s41598-022-14382-9

Brandes, N., Ofer, D., et al. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. Bioinformatics, 38(8), 2102-2110.

Chandra, A., Tünnermann, L., et al. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. eLife, 12, e82819.

Jumper, J., Evans, R., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

Pandey, A., Liu, E., et al. (2023). B-factor prediction in proteins using a sequence-based deep learning model. Patterns, 100805.

Smyth, M. S., Martin, J. H. J. (2000). X Ray crystallography. Journal of Clinical Pathology: Molecular Pathology, 53(1), 8-14.

Xu, G., Yang, Y., et al. (2024). OPUS-BFactor: Predicting protein B-factor with sequence and structure information.