Enhancing DR-BERT for accurate classification of intrinsically disordered regions in protein structures

Brendan Wang
Princeton University
byw2@princeton.edu

Viola Chen
Princeton University
yc1709@princeton.edu

Eugene Choi
Princeton University
ec0342@princeton.edu

Abstract

Intrinsically disordered regions (IDRs) of proteins are functionally important and numerous challenges remain in characterizing them. This project builds upon Nambiar et al. [1] by improving the DR-BERT model for predicting IDRs in proteins. We implemented their architecture from scratch and integrated the per-residue confidence scores (pLDDT) from AlphaFold into the model. Our results show that incorporating structural information significantly enhances IDR prediction accuracy. Additionally, our in-silico mutagenesis experiments provide insights into the role of specific residues in IDR formation. We also made an attempt to incorporate more sequence information with Multiple Sequence Alignment (MSA). Overall, our work suggests that a combined approach of utilizing sequence and structural data can yield more accurate IDR predictions.

1 Introduction

The 3-dimensional conformation of a protein is known to be closely linked to its function. However, protein structures are often not rigid. In recent years, there has been increasing research into the dynamical systems of proteins and how a single protein can assume multiple potentially continuous conformations. [2] Certain proteins, such as collagen, can have fairly rigid structures. Such stable, rigid structures enable those proteins to perform functions such as structural support. Other proteins, such as RNA-polymerase, however, do not have such rigid structures. They contain highly flexible, less structurally stable regions known as Intrinsically Disordered Regions (IDRs). Structured domain and IDRs can occur in the same protein or different proteins to enable various proteins to carry out various cellular biochemical functions [3].

Although IDRs are traditionally believed to be passive segments of protein that merely link structured domains together, recent studies have shown that IDRs are mechanically functional, and in fact actively participate in many diverse protein-mediated functions [4]. Therefore, it is important to understand the role played by IDRs in various cellular functions and disease mechanisms.

Accurate identification of IDRs is one of the first steps towards characterizing and understanding IDRs. With recent advancements in machine learning tools for protein structure prediction such as AlphaFold2 [5] and trRosetta [6], accuracy in IDR prediction has been greatly improved. However, due to the high heterogeneity and poor sequence conservation [3] among IDRs, challenges remain in characterizing IDRs and understanding their functions. Firstly, there is often a dynamic continuum between folded and completely disordered states of protein, making it difficult to recognize such proteins from any imaging results [7]. Secondly, due to their similar nature, it is hard to distinguish IDRs from protein switches, which alternate between a limited number of conformations upon specific stimuli [8].

In response to these challenges, the community has come up with Critical Assessment of Protein Intrinsic Disorder Prediction (CAID) to promote computational research in studying IDRs [9]. A challenge involving a curated dataset of disordered proteins, CAID aims to tackle a simpler problem of predicting which protein positions are disordered. Specifically, in the challenge, each predictor is asked to give a probability for each position, which is then converted into a binary classification label by applying a specific cutoff [7].

2 Background

Many attempts have been made to tackle CAID, with the current top performers being flDPnn[10], SPOT-Disorder2[11], DisoPred[12], RawMSA[13]. However, there is still huge room for improvement in terms of performance, as the top performers still have F1 scores below 0.5 [7]. A variety of computational techniques have been explored in those methods. When investigating the landscape of Intrinsically Disordered Region prediction, we narrowed our focus to three main approaches: Ensemble Methods, MSA + CNN Method, and MSA + CNN + RNN Method.

2.1 Ensemble Methods

The most prominent model in the ensemble category is f1DPnn [10], which combines weak learners with a random forest. It operates in three steps:

1. In the first step, f1DPnn generates a sequence profile using existing search tools.
2. Next, it completes feature generation at the protein, window, and residue levels.
3. In the final step, these aggregated features are fed into random forests, where the results are combined to make the final IDR predictions.

This lightweight model performs surprisingly well, with comparable performance to larger models. However, it's important to note that the feature and sequence profile generation is highly tailored to this specific task. Additionally, it can be challenging to iterate on this method, as increasing complexity does not necessarily lead to improved model performance.

2.2 MSA + CNN Method

RawMSA is a method that applies a Convolutional Neural Network (CNN) to embeddings from the Multiple Sequence Alignment (MSA) of the original input sequences [13]. This approach preserves the spatial and locality data of each sequence, as well as its coevolutionary data.

2.3 MSA + CNN + RNN Method

Similar to the RawMSA method, SPOT-Disorder2 employs a CNN over MSA embeddings to extract a feature vector. To capture the sequential nature of the residue data, these outputs are then fed into a Recurrent Neural Network (RNN) [11].

In general, we observed an increase in F1 scores and overall model performance as we progressed from f1DPnn to RawMSA to SPOT-Disorder2 Consequently, the natural progression from RNN implementations is the adaptation of transformer models, as seen in DR-BERT[1].

Similar to all methods discussed above, DR-BERTaims to address the challenge of predicting whether each amino acid position belongs to IDR and assign a probability for each position. Recent advancements in using language models for protein structure prediction, as demonstrated by ESMFold [14], inspired the authors to adapt a protein language model to the task of IDR prediction [1]. DR-BERTmade use of the Bidirectional Encoder Representations from Transformers (BERT) architecture [15], with the training of a token classification head on disordered region labels [1]. Overall, our work aims to extend the work of Nambiar et al. [1] by (1) seeking to enhance the classification of IDRs using DR-BERT by specifically incorporating per-residue confidence scores from AlphaFold into the input, (2) conducting case studies using the model, including in-silico mutagenesis experiments to probe IDR formation, and (3) exploring potential modifications to the model architecture to integrate multiple sequence alignments (MSA) as inputs.

3 Data

The original Uniref90 database is a database of clusters of protein sequences in which each sequence shares at least 90% identity with the longest sequence in the cluster. For pre-training, Nambiar et al. [1] sampled 6,564,742 and 250,000 proteins for training and validation sets, respectively. For fine-tuning, Nambiar et al. [1] acquired 2419 sequences each with ground truth labels from DisProt Version 9.2, June 2022, which is a publicly available database of intrinsically disordered proteins. This dataset was combined with the CAID 1 dataset to generate clusters using the CD-HIT algorithm. Importantly, only clusters without CAID 1 sequences were used for training and validation. The resulting dataset consisted of 1569 training, 176 validation, and 652 test protein sequences each with per-residue labels (0 and 1 for ordered and disordered respectively). For reference, we denote this dataset as "DisProt-Full."

For the customized models, we wanted to compute for each sequence the per-residue pLDDT scores from AlphaFold. While we originally intended to run AlphaFold2 on all the DisProt proteins in the aforementioned dataset, we estimated that the run time would have taken several weeks to get scores for all the proteins. As such, we instead used AlphaFoldDisorder, a publicly available dataset of DisProt protein sequences that had all the associated pLDDT scores. This dataset consisted of 386 training, 43 validation, and 48 test sequences. For reference, we denote this dataset as "DisProt-Sub."

4 Framework

4.1 Model Architecture

The original DR-BERT model consists of the following three components:

Embedding Layers: Based on the Bidirectional Encoder Representations from Transformers (BERT) model Devlin et al. [16], the embedding layer takes in the tokenized input of 1024 tokens and outputs for each token an embedding of size 768. It consists of two different components, namely the word embedding layer and the positional embedding layer. The former maps each amino acid token into a vector of 768 features and the latter learns the positional information of each amino acid relative to other ones in the sequence. In short, the embedding layer aims to learn a higher-dimensional deep representation of each amino acid in a given protein sequence.

Encoder Layers: The encoder layers consist of a series of 6 encoder blocks based on the original Transformer architecture, Vaswani et al. [17]. In this layer, embeddings from the embedding layer go through these blocks, each of which contains 12 self-attention heads; each such attention head learns a different relationship between the tokens in the sequence Vaswani et al. [17]. For a given protein sequence, the final output of the encoder layer is 1026 vectors, each of size 768. The first vector, middle 1024, and last token represent the start [CLS] token, the residues of the sequence, and the last token [SEP] token respectively Nambiar et al. [1].

Classification Head: The classification head is a simple fully-connected neural network that maps any amino acid representation to a vector of size 2, namely the logits for the labels 0 (ordered) and 1 (disordered).

For our custom-modified model, the architecture is the same except the classification head maps from a vector of size 769 to a vector of size 2. More details are described in Section 6.

4.2 Pre-Training Objective

Nambiar et al. [1] uses the masked language modeling objective from the RoBERTa framework Liu et al. [18] to pre-train the model, namely masking tokens in the training data and having the model predict the missing tokens given the context.

4.3 MSA-BERT

While DR-BERT aimed to fully incorporate all coevolutionary data into the model parameters, thus eliminating the need for Multiple Sequence Alignments (MSAs), we hypothesize that the current model's size, comprising approximately 45 million parameters, may not be sufficient to accommodate all coevolutionary data within the parameters. In comparison, other protein language models, as mentioned in [19], required a minimum model size of around 600 million parameters to store all coevolutionary data. We suspect that this limitation might not have been evident in DR-BERT's results due to the relatively small size of the CAID Dataset, consisting of roughly sub-2,000 sequences, which might not have comprehensively stress-tested the model.

To address this and create a comprehensive protein language model while maintaining computational efficiency, we made a modification to DR-BERT's model encoding architecture by introducing MSA inputs. Specifically, we take the 512 MSAs and apply Axial Attention to them, following the approach outlined in [19].

While we encountered implementation challenges and compute constraints in obtaining accurate results, we draw inspiration from [19], which successfully reduced model parameters by a factor of 22 while maintaining similar performance. By supplying coevolutionary data instead of needing to store it in the parameters, we believe we can potentially reduce our MSA-BERT Encoder to 2 million parameters.

5 Evaluation

Nambiar et al. [1] used three main metrics to evaluate how well their model performs for classification of amino acids as ordered or disordered: Area Under the Receiver Operating Characteristic Curve (AU-ROC), F1 score, and Matthew's Correlation Coefficient (MCC). The AU-ROC and F1-score are both methods widely used to assess classification models. The former measures the area under a curve that plots the false-positive and true-positive rate for various classification thresholds; the latter is another measure of classifier accuracy which combines precision and recall. The MCC score is a metric that provides a much more reliable metric on imbalanced datasets. Because the DisProt dataset had much more ordered than disordered labels, Nambiar et al. [1] argued that MCC score was an appropriate evaluation metric.

For our in-silico mutagenesis experiments described under Section 6, we use the mean squared error (MSE) as a measure of what we call "disruption." For a given protein sequence s of length n, let v and v' be the vector of DR-BERT scores for before and after the mutation of deleting the residue at position j (note that the DR-BERT score of a given residue is the probability of that residue being disordered). Additionally, the vector v also has the element corresponding to j omitted to make element-wise subtraction compatible. Then, if we denote s' as the sequence after the deletion, the disruption $D(s)$ is given by:

$$D(s, s') = \frac{1}{n} \sum_{i=1}^{n} (v_i - v_i^{'})^2$$

At a high level, disruption measures the change in prediction scores that result from mutating the sequence, and thus positively correlates with the importance of the deleted residue.

6 Implementation

In this section, we provide details of how we conducted our experiments. For our work, we utilized the Adroit Cluster, PyTorch, and Jupyter notebooks. All training was done using A100 GPUs.

6.1 Baseline Replication

For baseline replication, we fine-tune DR-BERTusing the weights of the pre-trained model provided by Nambiar et al. [1]. In particular, the model was trained for a total of 10 epochs on DisProt-Full using cross-validation and Cross Entropy Loss. The model with the lowest validation loss was chosen as the best fine-tuned model which we call the "fine-tuned reproduced model" or FTR.

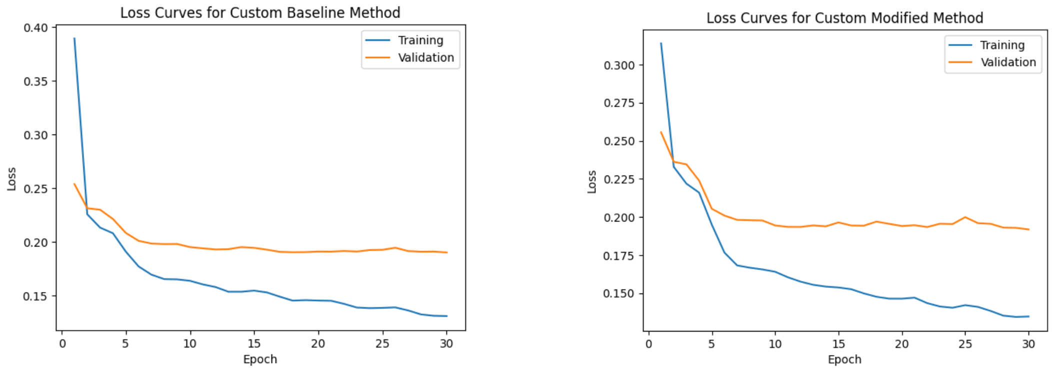

Figure 1: Training and validation curves from training both CBM and CMM model for 30 epochs.

6.2 Incorporating Confidence Scores

After implementing the baseline replication, we tested whether idea of incorporating the per-residue confidence score as an additional input improves classification performance. To begin, we downloaded the raw text file and filtered out all the sequences, labels, and pLDDT scores. Next, because the pLDDT scores were between 0 and 100, we performed normalization to get all values between 0 and 1. Thereafter, we deleted all duplicate sequences and did a random split to create training, validation, and test sets, which are collectively the DisProt-Sub dataset.

Next, we created two custom models. The first is the Custom Baseline Model (CBM), which has the same architecture as DE-BERT, except that we implemented the model from scratch instead of using the implementation from the transformers library. The second is the Custom Modified Model (CMM), which is similar to CBM but has the following modification: after getting the 1026 embeddings from the pre-trained model (each of size 768), we append the normalized pLDDT scores for each residue making the size of each embedding 769. Then, we intput the result into a fully connected layer to generate the logits. We summarize the difference between CBM and CMM as follows, where x, s, P and H are an amino acid token, its pLDDT score, the pre-trained model and the classification head respectively:

$$\text{CBM} : x \in \mathbb{R} \xrightarrow{P} P(x) \in \mathbb{R}^{768} \xrightarrow{H} H(P(x)) \in \mathbb{R}^2$$

$$\text{CMM} : x \in \mathbb{R} \xrightarrow{P} [P(x) + s] \in \mathbb{R}^{769} \xrightarrow{H} H(P(x)) \in \mathbb{R}^2$$

We train both models for a total of 30 epochs on DisProt-Subset using cross-validation and Cross Entropy loss. Similar to the baseline, we define the best models to be ones that achieve the lowest validation loss. In Figure 1, we show the loss curves of training both the CBM and CMM. Importantly, as the validation loss curves of both models plateau off, this suggests that both models have not overfit the data.

6.3 Case Study

We took a closer look at two specific proteins that are known to contain intrinsically disordered regions, the RPB6 protein and tumor P53 protein. Model weights are saved as checkpoints and `pkl` files. To produce the results of the case study, we load the weights from the three models mentioned above and obtain a vector of DR-BERT scores where the score at each position is the predicted score from the various models respectively.

6.4 In-silico Perturbation Experiments

After establishing working models, we sought to further investigate how a protein sequence mechanistically establishes the formation of IDRs. Knowing that the sequence of residues dictates the precise location of disordered regions, we hypothesized that residues in the transition regions from ordered to disordered domains (or vice versa) were most crucial in the formation of these IDRs.

To test this hypothesis, we performed an in-silico mutagenesis experiment using the aforementioned RPB6 protein, which contains 130 amino acids. In particular, suppose we represent the RPB6 protein sequence as $s = \{s_1, ..., s_{130}\}$. For any given position i, we fed s and $s' = s \setminus s_i$ into the FTR model to get two vectors v and v' representing the DR-BERT scores for s and s' respectively. Then, we could calculate the disruption of such a deletion as $D(s, s')$. We repeated this for all the residues. Results from this analysis are shown in Section 7.1.

Finally, we took the residue with the highest position, namely a Proline (P) at position 57, and substituted the residue with a different residue and repeated this for all the other 19 amino acids. Our hypothesis was that substituting amino acids that are more biochemically similar to P would induce less disruption. We use the BLOSUM62 matrix to compute the similarity scores for each amino acid with P (the scores are a measure of similarity with P). We then had the model make predictions on all the 19 mutated sequences and compared the relationship with residue similarity with disruption. Results from this analysis are shown in Section 7.1.

6.5 Encoder Implementation Details

As part of our investigation, we aimed to gain a deeper understanding of the implementation practices of LLMs, specifically BERT. We took a comprehensive approach by coding all the components of the BERT model from scratch, creating a flexible codebase that allowed us to edit each part of the model independently, including the attention block, embedding layer, tokenization layer, and more. All of these components were built off of the standard PyTorch model and the Hugging Face libraries were not used for this implementation. This approach gave us the flexibility to fine-tune input parameters and establish a pipeline for experimenting with different attention models.

In our initial pre-training run, we set the maximum sequence length to 500 tokens and filtered the Uniref90 dataset for sequences with fewer than 500 tokens. This resulted in a subset of the dataset containing approximately 20,000 sequences. However, we encountered a significant challenge with low accuracy, approximately 8%. Initially, we suspected that this low performance was due to the large number of parameters relative to the dataset size. We attempted to address this by reducing the number of encoding layers and scaling down the hidden dimension size of the MLP, resulting in only a marginal increase in accuracy to around 10%. This led us to suspect that the limitations of pre-training performance were inherently tied to the dataset's inability to capture the complexity of the domain. We also observed that, in the context of protein language modeling, any subset containing just thousands of sequences is insufficient to fully saturate the embedding model, particularly when considering that the original DR-BERT was trained on a dataset of 70 million sequences.

Given our time and resource constraints, we were unable to conduct as many diverse experiments as we would have liked. However, the modular structure of our project sets the stage for future work, as researchers can easily maintain the existing structure and make isolated modifications to test hypotheses, without the need to rebuild the entire pipeline for each experiment. Our training and testing pipeline is available at github: https://github.com/eugenechoi2004/COS597N.

7 Results

7.1 Baseline and Extension Results

We show the results of the baseline replication in Table 1. As the table illustrates, we reproduced the results for AU-ROC metric. However, there are discrepancies between the results for F1 and MCC scores. There are a few possible explanations for this. First, the dataset that the authors published did not match the split that they used; in fact, there was only one CAID test set provided despite the authors using two separate CAID test sets in their analysis. Additionally, the authors also included several metrics that involved convolving over windows over different sizes. While we only used the normal metrics (window size of 1), the authors might have reported the best performance across these different metrics. Nonetheless, by being able to reproduce AU-ROC results, we were able to establish a proof-of-concept. Next, we present the results of comparing the performance of the CBM and CMM model in Figure 2. As the table illustrates, the CMM model achieves higher performance than the CBM model in all 5 metrics. Importantly, this supports the idea that incorporating per-residue

Model	AU-ROC	F1	MCC
Original	0.82/0.83	0.55/0.55	0.43/0.43
FT-Reproduced	0.81	0.47	0.37

Table 1: A comparison between the test performance across three metrics of the original DR-BERT model presented in Nambiar et al. [1] and FT-Reproduced, the model from our replication baseline. The two numbers for the original model were the performance on the CAID 1 and 2 test sets respectively.

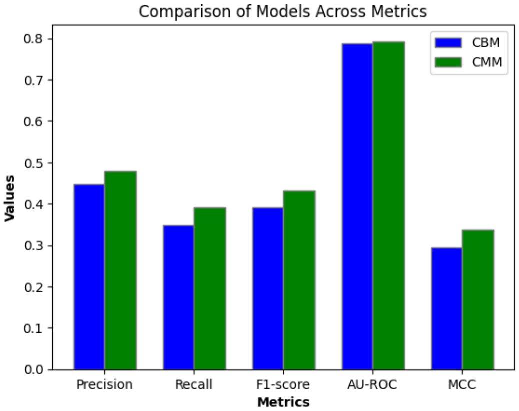

Figure 2: A comparison of the performance of the CBM and CMM models on the DisProt-Sub test set across five different evaluation metrics.

pLDDT scores into our input data and our model improves performance. Additionally, it suggests that there is additional information that the original DR-BERT model could not capture purely from the protein sequence alone. This information is likely to involve the structure of the protein, which is captured by the AlphaFold model. Our results indicate that combining sequence and structural information leads to better prediction of IDRs within protein sequences.

7.2 Case Study and Perturbation Experiment Results

We further examined the result of our model on two proteins with known intrinsically disordered regions. RPB6 is a subunit of RNA-polymerase I, II, and III, and plays important roles in transcription. Part of its chain is known to be intrinsically disordered [20]. In the original paper [1], it is used as a case study to examine the prediction as well as the attention maps. Here, we re-used this protein to show that we are able to reproduce the results presented in the original DR-BERTpaper. This also validated that the custom subset model performs just as well as the fine-tuned DR-BERTmodel. The other protein we looked into is the human tumor p53 protein, which is also known to be disordered. Interestingly, we noticed more discrepancies between the prediction of the three models tested. At position 340, there is a sudden dip for the pre-trained model, a result that is not present for the fine-tuned or custom-modified model. This suggests that the pre-trained model may be lacking in its ability to accurately distinguish disorder. For positions 0 to 100, there seems to be a lot of variations in the prediction score, indicating room for error and improvement. Results are depicted in Figure 3

Finally, we present results from our in-silico mutagenesis study of the RPB6 protein in Figures 4 and 5. Figure 4 contains two subfigures, one showing the distribution of predicted DR-BERT scores across all the sequence positions and the other showing the distribution of disruption (in terms of MSE) that results from deleting the residue at each sequence position. As the left subfigure illustrates, the transition region is the region between the disordered (residues 0 to 60) and the ordered (residues 60 to 130) domain. Strikingly, in the right subfigure, we observe that the residue

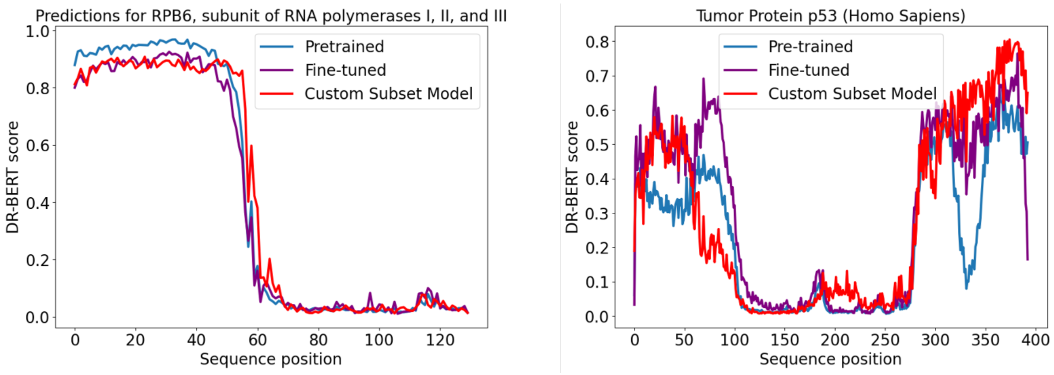

Figure 3: The left and right graphs shows the DR-BERT scores of the RPB6 protein (a subunit of RNA polymerase) and the tumor p53 protein (respectively), as is predicted by pre-trained, fine-tuned, and customized model.

positions that upon deletion lead to the most disruption are precisely the positions in this transition region. Because deleting residues at the transition point was most deleterious, these results suggest that residues at the transitions play a bigger role in establishing IDR structure than residues not in such transitions. In addition to MSE, we also compute the Pearson correlation coefficient (PCC) between the DR-BERT scores before and after deletion. Importantly, lower PCC values indicate more disruption. Results are displayed in Figure 6 of the Appendix. Thereafter, seeing the residue

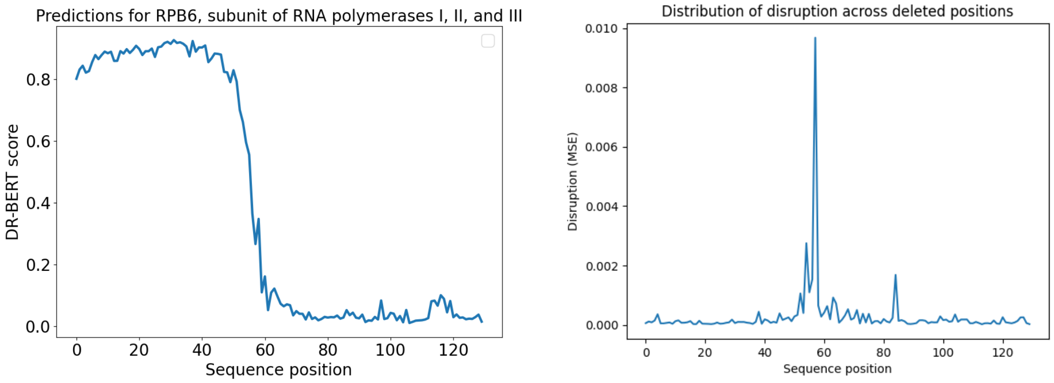

Figure 4: The left graph shows the predicted DR-BERT score (probability of being disordered) for each residue in the RPB6 protein. The right shows the relationship between disruption (measured in terms of MSE) between DR-BERT scores before and after deletion and sequence position. Note that position 57, there is a peak in the disruption.

P at position 57 has the most deleterious effect, we were wondering what would happen if rather than its deletion, we performed an amino acid substitution mutation. From Figure 5, we observe a negative correlation between BLOSUM62 scores and disruption. Because BLOSUM scores provide a measure of the similarity of the current amino acid P and another amino acid (less negative scores indicating greater similarity), our results are suggested to align with our original hypothesis, namely that the more biochemically and structurally similar the substituted amino acid is to P at position 57, the less the disruption. Finally, we also repeated this analysis in another random residue of the sequence which revealed both weaker levels of correlation and disruption, further suggesting that residues in non-transition regions are less crucial. Results from this follow-up analysis are shown in Figure 7 of the Appendix.

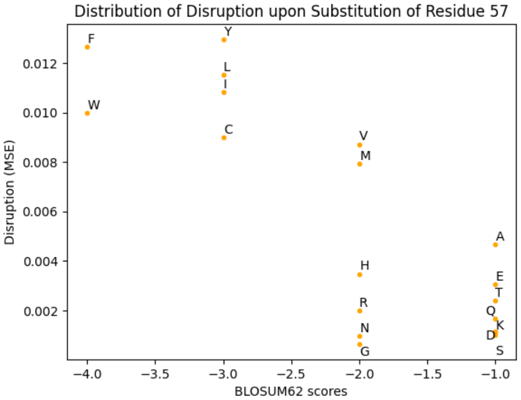

Figure 5: Disruption scores that result from replacing residue 57 in the RPB6 protein with another residue. Scores are plotted across BLOSUM62 scores for each amino acid with respect to proline.

8 Conclusion and Future Work

This work extends the work of Nambiar et al. [1] by investigating ways to adjust the model and data to improve the classification of IDRs in protein structures. After performing baseline replication, we show that incorporating per-residue pLDDT scores from AlphaFold into the input leads to improvements in the predictive power of the classification model. Additionally, we performed in-silico perturbation experiments to demonstrate that deleting residues in the transition regions between disordered and ordered domains led to more deleterious effects than deleting residues in regions outside the transitions. Moreover, we discovered that substitutions of more similar amino acids lead to less disruption. We anticipate these results to provide a foundation for future work that aims to unravel the biological mechanisms behind the formation and interactions of IDRs. Finally, we adapt the original model architecture to be tailored to MSA inputs with the hypothesis that co-evolutionary information within proteins have yet to be fully captured.

We propose many endeavors for future work. First, as our custom models were fine-tuned on DisProt-Sub, a relatively small dataset, we wish to investigate how performance would change if we used a bigger dataset. One way to do this would be to compute pLDDT scores with AlphaFold2 for all proteins available in DisProt. Additionally, we wish to do more similar in-silico perturbations on other proteins to see if we observe similar phenomena. Finally, we hope to extend the preliminary work of our MSA-BERT model to test our conjecture that MSA inputs increase predictive power.

Contributions

Brendan worked on reproducing the baseline replication studies and improving the model by incorporating the pLDDT confidence scores. More specifically, he focused on downloading, preprocessing the DisProt-Sub dataset before designing and training the custom CBM and CMM models to generate quantitative results. Additionally, he conducted and analyzed both the deletion and substitution in-silico mutagenesis experiments before generating the corresponding plots. Eugene worked on the design and implementation of the encoding architecture from scratch. Additionally, he was responsible for the MSA implementation of the BERT model and the modified attention mechanism. Viola worked on reproducing the results for the case study, including getting model parameters and weights from the original paper and the modified model made by Brendan. She did literature research on the background of intrinsically disordered regions and found other potential target proteins as case study. She also worked on subsetting UniRef90 and UniRef50 data in various methods to obtain MSA input for the MSA implementation.

References

[1] Ananthan Nambiar, John Malcolm Forsyth, Simon Liu, and Sergei Maslov. Dr-bert: A protein language model to annotate disordered regions. *bioRxiv*, pages 2023–02, 2023.

[2] Robin Van Der Lee, Marija Buljan, Benjamin Lang, Robert J Weatheritt, Gary W Daughdrill, A Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T Jones, et al. Classification of intrinsically disordered regions and proteins. *Chemical reviews*, 114(13):6589–6631, 2014.

[3] Alex S Holehouse and Birthe B Kragelund. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology*, pages 1–25, 2023.

[4] M Madan Babu. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochemical Society Transactions*, 44(5):1185–1200, 2016.

[5] Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using alphafold2. *Nature communications*, 13(1):1265, 2022.

[6] Zongyang Du, Hong Su, Wenkai Wang, Lisha Ye, Hong Wei, Zhenling Peng, Ivan Anishchenko, David Baker, and Jianyi Yang. The trrosetta server for fast and accurate protein structure prediction. *Nature protocols*, 16(12):5634–5651, 2021.

[7] Alessio Del Conte, Mahta Mehdiabadi, Adel Bouhraoua, Alexander Miguel Monzon, Silvio CE Tosatto, and Damiano Piovesan. Critical assessment of protein intrinsic disorder prediction (caid)-results of round 2. *Proteins: Structure, Function, and Bioinformatics*, 2023.

[8] Peter E Wright and H Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular cell biology*, 16(1):18–29, 2015.

[9] Marco Necci, Damiano Piovesan, and Silvio CE Tosatto. Critical assessment of protein intrinsic disorder prediction. *Nature methods*, 18(5):472–481, 2021.

[10] Gang Hu, Akila Katuwawala, Kui Wang, Zhonghua Wu, Sina Ghadermarzi, Jianzhao Gao, and Lukasz Kurgan. fldpnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nature Communications*, 2022. doi: 10.1038/s41467-021-24773-7.

[11] He Zhou, Yifei Yuan, Jing Li, Yixue Yang, Yuxuan Deng, Yunhu Zhang, Yuanyuan Huang, and Jianzhu Fan. Spot-disorder2: Improved protein intrinsic disorder prediction by ensembled deep learning. *Journal of Chemical Information and Modeling*, 60(8):3733–3745, 2020. doi: 10.1021/acs.jcim.20b00243. URL `https://www.sciencedirect.com/science/article/pii/S1672022920300243`.

[12] David T Jones and Domenico Cozzetto. Disopred3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, 31(6):857–863, 2015.

[13] Claudio Mirabello and Björn Wallner. rawmsa: End-to-end deep learning using raw multiple sequence alignments. *PLOS ONE*, 14(8):e0220182, 2019. doi: 10.1371/journal.pone.0220182.

[14] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[15] 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.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

[19] Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer: A protein language model for efficiently learning from multiple sequence alignments. *bioRxiv*, 2021. doi: 10.1101/2021.02.12.430858. URL `https://www.biorxiv.org/content/10.1101/2021.02.12.430858v1.full`.

[20] Bi Zhao, Akila Katuwawala, Christopher J Oldfield, Gang Hu, Zhonghua Wu, Vladimir N Uversky, and Lukasz Kurgan. Intrinsic disorder in human rna-binding proteins. *Journal of Molecular Biology*, 433(21):167229, 2021.

A Appendix

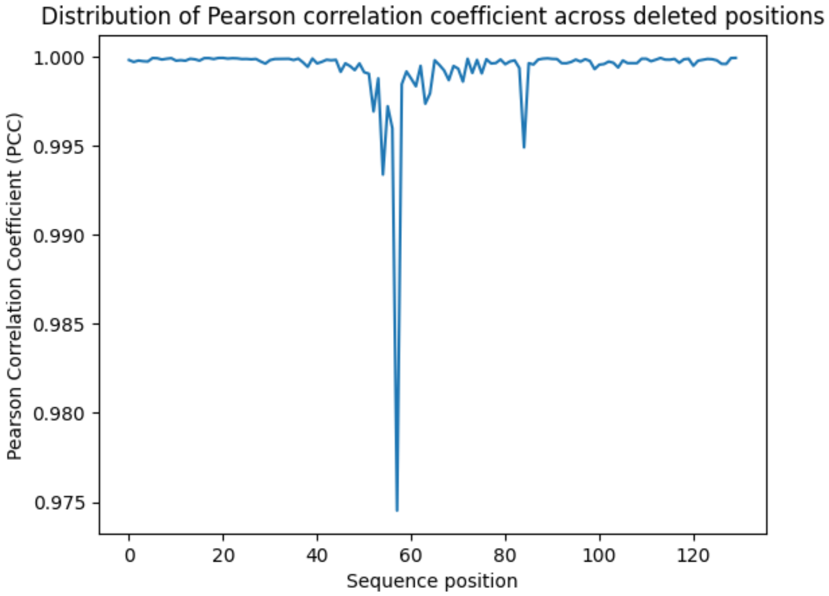

Figure 6: Distribution of the Pearson correlation coefficients between DR-BERT scores before and after deletion and sequence position. Note that position 57, there is a noticable dip in Pearson Correlation Coefficient, suggesting more disruption.

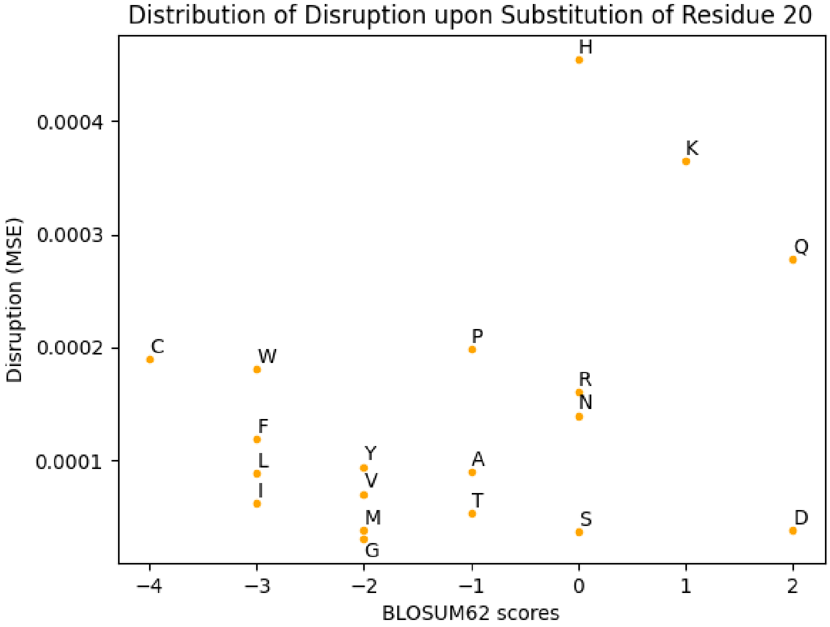

Figure 7: Disruption scores that result from replacing residue 20 (a random residue) in the RPB6 protein with another residue. Scores are plotted across BLOSUM62 scores for each amino acid with respect to proline.