Multi-DeepProtGraphGO: Integrating GCN on PPI Networks with Sequence-Driven Convolutional Bi-LSTM and Attention for Protein Function Prediction

Supplemental Document S1

Balaiah Kukkala, Akshay Deepak, Vikash Kumar, and Aravind Prakash

In this supplementary document, we have added three sections (Sections S.I to S.IV). A set of quantitative evaluation metrics, as defined in Section S.I, was employed for performance analysis of the proposed model. The performance analysis of the proposed model for protein sequence *DeepAttnSeq* on varying embedding dimensions, choosing different CNN and attention parameters in Section S.II. And, also defined the model variants for the ablation analysis in Section S.III. Lastly, model interpretability and it's biological insights in the Section S.IV.

S.I. Evaluation metrics

Different metrics are used to evaluate the performance or quality of our model. By analyzing these measures, we can determine how well our model performed on the acquired data. The metrics we have chosen to analyze our model are the F_{max} score and the AUPR. These metrics are defined as follows:

- **F1-Score:** Harmonic mean of average precision and average recall at a specified threshold υ.

$$F_{max} = \max_{\upsilon} \left\{ \frac{2 * pr\left(\upsilon\right) * re\left(\upsilon\right)}{pr\left(\upsilon\right) + re\left(\upsilon\right)} \right\} \tag{1}$$

where $pr\left(\upsilon\right)$ and $re\left(\upsilon\right)$ represent the average precision and average recall calculated at a specific threshold value υ.

$$pr\left(\upsilon\right) = \frac{1}{k\left(t\right)} \sum_{i=1}^{k(t)} \frac{\sum_j 1\left(Pred\left[G_j, P_i\right] \geq \upsilon\right) \cdot Act\left[G_j, P_i\right]}{\sum_j 1\left(Pred\left[G_j, P_i\right] \geq \upsilon\right)} \tag{2}$$

$$re\left(\upsilon\right) = \frac{1}{n\left(t\right)} \sum_{i=1}^{n(t)} \frac{\sum_j 1\left(Pred\left[G_j, P_i\right] \geq \upsilon\right) \cdot Act\left[G_j, P_i\right]}{\sum_j Act\left[G_j, P_i\right]} \tag{3}$$

where $k(t)$ represents the total number of proteins with at least one predicted positive GO term and "$n(t)$ represents the total number of proteins in the test set. $1(.)$ is an indicator function. $Pred[G_j, P_i]$ represents the predicted probability of the model for a specific protein P_i with the GO term G_j. $Act[G_j, P_i]$ represents the actual label value for protein P_i with the GO term G_j.

- **AUPR:** It is an area under the $precision - recall$ curve, where each point on the curve represents precision and recall values at different thresholds.

$$AUPR = \int_{-\infty}^{\infty} pr\left(\upsilon\right) - re\left(\upsilon\right) d\upsilon \tag{4}$$

S.II. Performance analysis of the proposed *DeepAttnSeq*

A. Effect of varying Embedding dimension

In this sub-section, we evaluate the *DeepAttnSeq* of the proposed model with respect to various embedding dimensions such as 64, 128, 256, and 512 with a kernel size of 128 and the $no.of heads = 8$ in multi-head self-attention mechanism as constant. The results are shown in Table I. The performance of the *DeepAttnSeq* improves as the embedding dimension increases up to 256. However, when the embedding dimension is further increased to 512, a slight decrease in performance is observed across all sub-ontologies, indicating that excessively large embeddings may lead to overfitting or redundant feature representations.

TABLE I
COMPARATIVE ANALYSIS OF PROTEIN SEQUENCE DATA FOR DIFFERENT SUB-ONTOLOGIES AGAINST PERFORMANCE METRICS ON VARIOUS EMBEDDING DIMENSIONS

Sub-Ontology	Emb dim	3-mer				4-mer			
		Re_{Avg}	Pre_{Avg}	F_{max}	AUPR	Re_{Avg}	Pre_{Avg}	F_{max}	AUPR
BP	64	0.4737	0.4101	0.4396	0.44	0.4872	0.4028	0.4410	0.42
	128	0.4858	0.4079	0.4435	0.44	0.4908	0.4100	0.4468	0.43
	256	0.4878	**0.4107**	0.4460	**0.45**	**0.4916**	0.4099	**0.4470**	0.44
	512	0.4858	0.4028	0.4404	0.43	0.4839	0.4027	0.4396	0.43
CC	64	0.4862	0.5433	0.5131	**0.49**	0.4979	0.5146	0.5061	**0.49**
	128	0.4896	0.5434	0.5151	0.48	0.5093	0.5152	0.5122	0.48
	256	0.4905	**0.5582**	**0.5221**	**0.49**	**0.5105**	0.5179	0.5142	**0.49**
	512	0.4935	0.5535	0.5218	**0.49**	0.5031	0.5148	0.5089	0.48
MF	64	0.4660	0.5146	0.4891	0.49	0.4774	0.5140	0.4950	**0.50**
	128	0.4662	0.5208	0.4920	0.48	0.4833	0.5096	0.4961	**0.50**
	256	0.4924	**0.5222**	**0.5068**	**0.50**	**0.5092**	0.4914	0.5002	0.49
	512	0.5005	0.5024	0.5014	0.49	0.4906	0.4979	0.4942	0.49

TABLE II
COMPARATIVE ANALYSIS OF PROTEIN SEQUENCE DATA FOR DIFFERENT SUB-ONTOLOGIES AGAINST PERFORMANCE METRICS ON VARIOUS KERNEL SIZES OF CNN USING EMBEDDING DIMENSION OF 256 AND HAVING STRIDE 32

Sub-Ontology	Kernel Size	3-mer				4-mer			
		Re_{Avg}	Pre_{Avg}	F_{max}	AUPR	Re_{Avg}	Pre_{Avg}	F_{max}	AUPR
BP	32	0.5033	0.3967	0.4437	0.44	0.4858	0.4049	0.4417	**0.45**
	64	**0.5042**	0.3947	0.4428	0.43	0.4987	0.3962	0.4416	0.43
	128	0.4878	**0.4107**	**0.4460**	**0.45**	0.4912	0.4029	0.4427	**0.45**
CC	32	0.4865	0.5829	0.5304	**0.51**	0.4923	0.5165	0.5041	0.47
	64	0.4806	**0.5961**	**0.5322**	**0.51**	**0.5295**	0.5158	0.5225	0.50
	128	0.4905	0.5582	0.5221	0.49	0.5069	0.5541	0.5294	**0.51**
MF	32	0.4586	0.4457	0.4521	0.44	0.4377	0.4845	0.4599	0.45
	64	0.4651	0.5075	0.4854	0.48	0.4632	0.4718	0.4675	0.47
	128	**0.4924**	**0.5222**	**0.5068**	0.50	0.4952	0.4927	0.4939	**0.51**

TABLE III
COMPARATIVE ANALYSIS OF PROTEIN SEQUENCE DATA FOR DIFFERENT SUB-ONTOLOGIES AGAINST PERFORMANCE METRICS USING A FIXED EMBEDDING DIMENSION OF 256 AND VARYING ATTENTION HEADS

Sub-Ontology	Attention Heads	3-mer				4-mer			
		Re_{Avg}	Pre_{Avg}	F_{max}	AUPR	Re_{Avg}	Pre_{Avg}	F_{max}	AUPR
BP	2	0.4684	0.4119	0.4383	0.43	**0.4944**	0.4076	0.4468	**0.45**
	4	0.4808	0.4079	0.4414	0.44	0.4916	0.4099	**0.4470**	0.43
	8	0.4878	0.4107	0.4460	**0.45**	0.4862	**0.4136**	0.4470	0.44
CC	2	0.4907	0.5525	0.5198	**0.49**	0.4890	0.5484	0.5170	0.48
	4	0.5006	0.5034	0.5020	0.48	**0.5093**	0.4852	0.4970	0.48
	8	0.4905	**0.5582**	**0.5221**	**0.49**	0.4918	0.5030	0.4973	0.47
MF	2	0.4660	0.5199	0.4915	0.48	0.4887	0.5021	0.4953	**0.50**
	4	0.4662	0.5208	0.4920	0.48	0.4833	0.5096	0.4961	**0.50**
	8	**0.4924**	**0.5222**	**0.5068**	**0.50**	0.4709	0.5163	0.4925	0.49

B. Performance analysis on various CNN parameters

In this sub-section, we discuss the performance of *DeepAttnSeq* of the proposed model by adjusting the parameters of the Convolutional Neural Network (CNN). For this experiment, we vary the kernel size to 32, 64, and 128 with a constant stride of 32, while the $embdim = 256$ and the $no.of heads = 8$ in multi-head self-attention were fixed. In order to test our model, we fit our *DeepAttnSeq* on the Uniprot dataset that was discussed in the main manuscript, in Section 3.1.

As shown in Table II, the *DeepAttnSeq* achieves best results on 3-mer representation of protein sequence. The *DeepAttnSeq* performs better with a kernel size of 128 for both BP and MF sub-ontology. Subsequently, for the CC sub-ontology, the *DeepAttnSeq* achieves highest Pre_{Avg}, Re_{Avg}, and F_{max} with kernel size of 64.

C. Effect of varying the number of heads in the Multi-Head Self-Attention Mechanism

To determine the impact of the number of attention heads in the multi-head self-attention mechanism of the *DeepAttnSeq* of the proposed architecture, we first ensure that the kernel size of the convolution layer is 128 and embed dim is 256 before fitting our *DeepAttnSeq* to the Uniprot dataset, as discussed in Section **??**.

As shown in Table III, for the BP sub-ontology, the *DeepAttnSeq* behaves inconsistent with different number of heads. For the CC sub-ontology, the *DeepAttnSeq* achieves highest Re_{Avg}, F_{max} and AUPR values with 8 multi-heads on 3-mer representation. However, for the MF sub-ontology as well, the *DeepAttnSeq* produces better results with 8 multi-heads on 3-mer representation.

D. Section Summary

As shown in Tables I, II, and III, the 3-mer representation consistently outperforms the 4-mer representation across different embedding sizes, kernel sizes, and attention heads. Overall, with embedding dim = 256, kernal size = 128, and attention head = 8 the proposed method produces superior results as shown in Tables I, II and III. Hence, proposed model with these configurations are used for the experiments in the remaining part of the paper.

S.III. Evaluated Variants in Ablation Study

We analyzed the isolated effects of removing key components from the *DeepAttnSeq* and *Graph-GCN*$_{\text{mean}}$ modules, and further evaluated several hybrid and combination variants to examine their collaborative impact on overall model performance. The following variants were considered:

- **DeepAttnSeq**: The baseline sequence model comprising an embedding layer, CNN layer, BiLSTM layer, and an attention mechanism, applied to protein sequences as described in main script, Section 2.1.2.
- **DeepAttnSeq (-BiLSTM)**: A variant in which the BiLSTM layer is removed, and the attention mechanism operates directly on the features extracted by a CNN layer from the input protein sequences. This setup allows us to assess the role of contextual modeling introduced by the BiLSTM, isolating the contribution of convolutional features and attention alone.
- **DeepAttnSeq (-Attention)**: A variant where the attention mechanism is removed, and the BiLSTM outputs are aggregated using simple mean pooling across the sequence length. This setup allows us to assess the importance of the attention mechanism in identifying and emphasizing informative residues, compared to uniform aggregation. The performance of this variant reveals how much the model benefits from selectively weighting features versus treating all positions equally.
- **Graph-GCN$_{\text{mean}}$**: A standalone graph-based model leveraging mean aggregation over neighboring nodes to generate smooth and balanced embeddings from protein–protein interaction (PPI) data, as detailed in main script, Section 2.1.3.
- **DeepAttnSeq (-BiLSTM) + Graph-GCN$_{\text{mean}}$**: A multimodal variant combining the ablated DeepAttnSeq model (without BiLSTM) with Graph-GCN outputs via concatenation.
- **DeepAttnSeq (-Attention) + Graph-GCN$_{\text{mean}}$**: Similar to the above, but the attention layer is ablated instead of BiLSTM.
- **Multi-DeepProtGraphGO (Full Model)**: As outlined in main script, Section 2.1, the complete proposed architecture combining DeepAttnSeq (with both BiLSTM and attention) and Graph-GCN$_{\text{mean}}$, followed by a fusion layer for final prediction.

S.IV. Model Interpretability and it's Biological Insights

In this section, we cover the proposed model interpritability and it's characteristics on the Uniprot dataset (explained in the main manuscript, in Section 3.1).

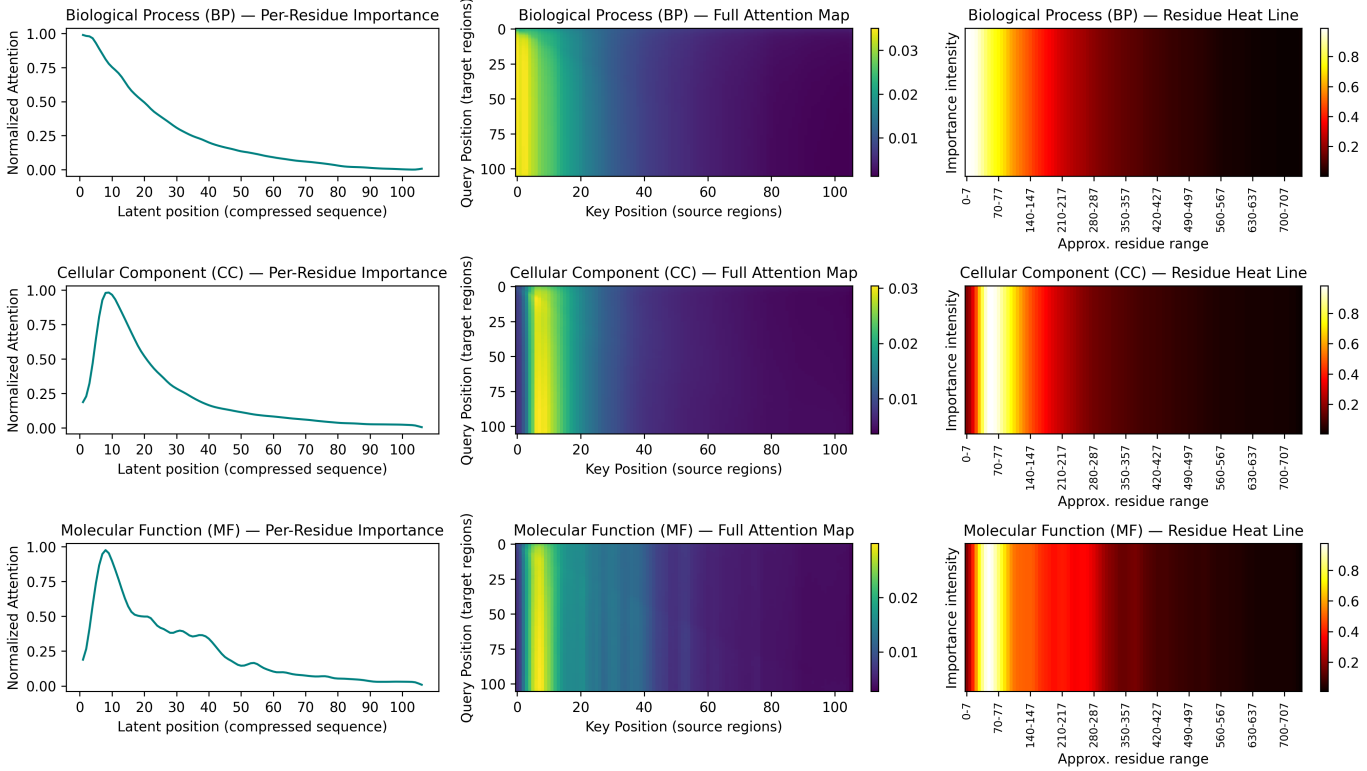

Fig. S1. Attention Analysis of the Proposed Architecture *DeepAttnSeq* to Extract the Protein Sequence Features

A. Attention / Feature Importance Maps (Sequence)

DeepAttnSeq exhibits pronounced attention toward specific residues, reflecting its ability to isolate functionally salient regions, departing from uniform sequence processing. This biologically guided focus is reinforced by alignment between high-attention residues and conserved motif regions.

As illustrated in the figure 1, *DeepAttnSeq*'s BP attention profile exhibits a sharp decline in importance across latent positions, suggesting that the initial compressed representations exert the greatest influence. The corresponding attention heatmap reinforces this observation, displaying a dense concentration of activation in the early regions—indicating localized dependencies and strong intra-segment focus. The residue activation trend further supports this pattern, with pronounced responses within the first 100–200 residues that gradually diminish along the sequence. Collectively, these findings indicate that BP-related information is predominantly encoded within early conserved motifs, underscoring the model's capacity to selectively emphasize biologically relevant sequence regions.

For the Cellular Component (CC) category, the attention profile reveals a pronounced and sharply localized peak around latent positions 10–20. This high-intensity region in the per-residue importance curve indicates that the model leverages a compact set of representations that are pivotal for localization-related prediction. The corresponding attention heatmap and residue activation trace further emphasize this trend, exhibiting strong focus in the early latent dimensions with negligible activation elsewhere. Such selective concentration is consistent with biological expectations, as cellular component information is typically encoded within short, localized sequence motifs rather than dispersed across the entire sequence.

The MF attention profile displays several moderate peaks across latent positions, indicating that functionally relevant signals originate from multiple, non-contiguous regions of the sequence. This dispersed pattern reflects the model's ability to integrate local and mid-range dependencies, consistent with the biological observation that critical residues, though separated in sequence, collectively determine molecular activity. Overall, MF prediction relies on an ensemble of informative regions rather than a single dominant subsequence, highlighting the model's capacity to capture distributed functional information.

B. Graph Interpretability (for PPI)

Interpretation of the GCN using edge weights reveals which protein–protein interactions are most critical for model predictions. Interactions with weights exceeding the average combined score threshold are deemed highly influential, reflecting their

prominent role in propagating information across the network. For instance, edges connecting protein A and protein B surpass this threshold, indicating functional interdependence in the associated biological process. Highlighting these top-weighted edges allows the model to pinpoint key subgraphs and interaction patterns, providing a biologically meaningful view of network-level contributions to prediction.

C. Fusion-level Interpretability (Sequence + PPI)

The integration of sequence and PPI-derived embeddings reveals complementary contributions to protein function. Sequence features capture intrinsic biochemical activities, while PPI embeddings reflect context-dependent roles within interaction modules. Functional clusters in the PPI network correspond to shared GO terms, and nodes with multiple annotations highlight multifunctional proteins. These patterns demonstrate that combining sequence and network information not only improves predictive performance but also provides biologically interpretable insights into protein function.