IT499 - Major Project II

# OntoPred: An efficient attention-based approach for Protein Function Prediction using skip-gram features


Team Members (Batch 23):

- Suyash Chintawar - 191IT109
- Rakshit Kulkarni - 191IT245

# Introduction

- The presence of proteins in an organism enables various molecular and biological processes that are essential for smooth operation of different biological components.
- Understanding the behavior of biological components additionally requires knowledge of the proteins' functions.
- Generally, function can be thought of as, "anything that happens to or through a protein". The GO Consortium provides a classification of functions, based on a dictionary of well-defined terms divided into three main categories of molecular function, biological process and cellular component.
- This is a particularly challenging task because proportionally small number of unique GO terms account for a large proportion of the annotations. This leads to skewness in the distribution of the GO terms.

# Literature Survey

| S. No. | Authors | Methodology | Merits | Demerits |
|---|---|---|---|---|
| 1. | Hakala et al. (2020) | Ensemble of multiple models used and best subset is derived | Use of multiple features to find protein similarity | Only NN and RF used. Did not capture sequence information |
| 2. | Giri et al. (2020) | Use of multi-modal features using structure, sequence and interaction features. | Performed well on Cellular Component aspect | Comparatively poor performance on Biological process aspect. |
| 3. | Ranjan et al. (2019) | Proposed a hybrid framework of two different models namely ProtVecGen-Plus and MLDA approaches. | Proposed a simple yet powerful method of segmentation of sequences | The standalone proposed model performs slightly poor for short and mid-length sequences. |

# Literature Survey

| S. No. | Authors | Methodology | Merits | Demerits |
|---|---|---|---|---|
| 4. | Cao et al. (2017) | Used Neural Machine Translation technique along with RNNs | Proposed a completely new method to solve the task | Could not beat homologous based approaches. |
| 5. | Kulmanov et al. (2018) | Used both the protein sequences, and protein-protein interaction network features. | Has a potential to predict any class given training data. | Computationally expensive and requires lot of training data. |
| 6. | Ranjan et al. (2021) | Used variants of tf-idf descriptors that are length and log normalized tf-idf. | Proposed method captures small pattern regions efficiently. | Slight decrease in F1-scores for long protein sequences. |

# Outcome of Literature Survey

- Attention mechanism can be used to focus on those segments in the sequences which contribute in the exhibition of GO terms.
- Skip grams captures positional information of amino acids on a broader level compared to the traditional n-gram features.
- It can be also observed that although using multi-modal approaches does improve the performance of the task, the trade-off between the metric measures and the difficulty of acquiring these multi-modal features must also be considered.

# Motivation

- Traditionally, experimental procedures are used for estimation of protein functions, but they have a major drawback of being slow and expensive.
- As the volume of newly sequenced proteins is increasing over the past decade due to the use of cutting-edge high-throughput sequencing techniques, it has become necessary to automate the process of protein function prediction.
- Machine learning, and deep learning has been actively used as a solution to automate this annotation process and this area of bioinformatics is still under research because the task of function prediction is very challenging due to its nature. (Sparsity, Variable length sequences, etc).

# Problem Statement

To perform the task of protein function prediction using segmentation of variable length protein sequences and analyse the performance on three aspects of gene ontology (MF, BP and CC).

# Objectives

- To prepare datasets corresponding to the three ontology aspects, Molecular function (MF), Biological Process (BP) and Cellular Components (CC) from the UniProtKB.
- To use a feature based approach which transforms raw protein sequences into their corresponding feature vectors.
- To build a multi-label classifier to predict the functions of proteins, i.e. their related Gene Ontology (GO) annotations.
- To check the performance of built models with respect to the three different aspects namely, MF, BP, and CC to predict GO terms and enhance overall accuracy measures.
- To compare the performance of the proposed architecture with state-of-the-art methods and improve performance as required.

# Proposed Methodology

- The proposed methodology can be divided into three major steps namely,
  - Sequence segmentation
  - Feature Vector Generation
  - Thresholding for Multi-label classification
- Let’s head over into each of these steps in detail…

# Proposed Methodology

**Step 1** :

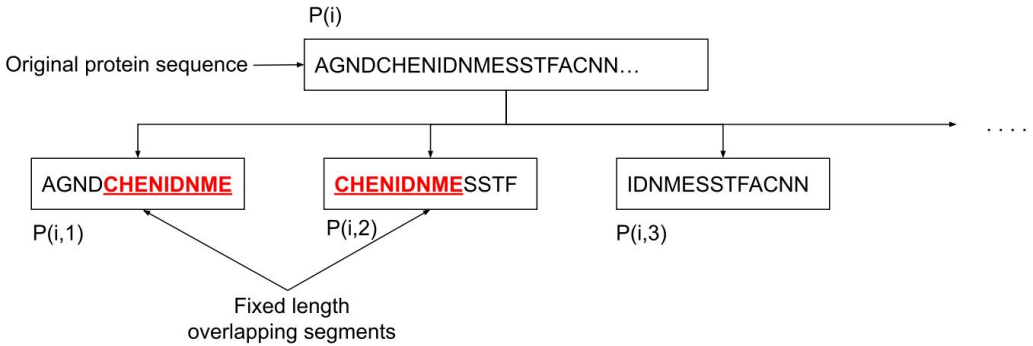



Fig 2. Converting protein sequences to smaller segments

- This technique is used to bring all sequences to a uniform length.
- These segments have overlapping regions to preserve the order of amino acids in the original sequence.
- Each of these segments is considered as an individual sample to the deep learning model having the same gene ontology annotations as that of the original sequence.

# Proposed Methodology

**Step 2** :

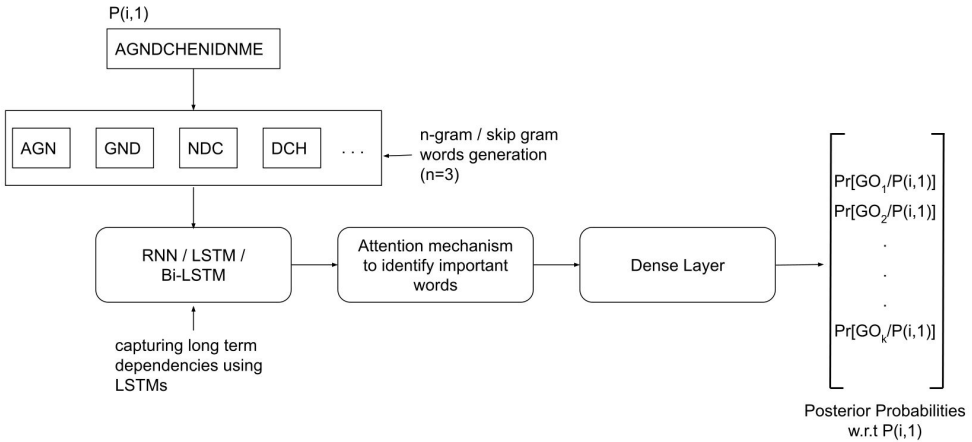



Fig 3. Deriving probabilities of GO terms for a segment

- The sequence-to-feature conversion is achieved by forming the words of the segment by using n-grams/skip-grams (protein words).
- The embeddings of these so called protein words is fed to the Bi-LSTM to get the sequence output on which attention mechanism is applied.
- Attention mechanism is used to give higher weights to the output features of the Bi-LSTM model which have more influence on the output. That is, if a particular part of the sequence is responsible for the presence of a GO term, the attention weights for that particular input will be higher.

# Proposed Methodology

**Step 3** :

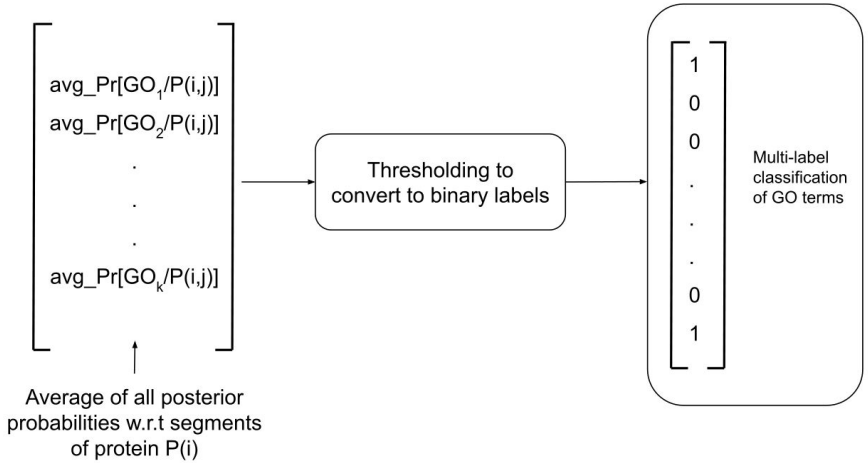



Fig 4. Thresholding for multi-label classification

- The posterior probabilities of the original sequence are obtained by computing the average posterior probability for each GO term across all protein segments.
- A fixed threshold of 0.5 is used to compute the F1-score, precision and recall metrics. While computing the maximal F1-score, 100 equally spaced numbers between [0,1] are used as the threshold values.

# Sequence Similarity Scores

- Diamond is a sequence alignment technique used for pairwise alignment of proteins.
- DiamondScore makes use of the similarity between proteins in the train set to annotate new proteins by transferring labels between that of the most similar ones in train proteins.
- Using the diamond tool we obtain the pairwise bitscores for the test protein sequences of each aspect which are then normalized to get the DiamondScores (predictions).
- In OntoPredPlus, the average of the predictions obtained from OntoPred and the predictions obtained from the diamond scores (which is basically the sequence similarity scores) for each GO term of a protein sequence is considered.

# Experimental Results and Analysis

# Datasets

**UniProt-SP dataset**

- To create this dataset, the Swiss-Prot entries from the UniProtKB database have been used as it is experimentally verified and reviewed.
- To prepare the individual datasets of each aspect, 100K samples which have sequence lengths between 100 and 2000 both inclusive are randomly chosen from the total 568K samples in the UniProtKB database.
- For each aspect, the GO terms which annotate less than 200 sequences are removed.

# Datasets

Table 1 : Specifications of the UniProt-SP dataset.

| Ontology Aspect | No. of Samples | No. of GO Terms |
|---|---|---|
| Molecular Function (MF) | 76741 | 149 |
| Biological Process (BP) | 72470 | 201 |
| Cellular Component (CC) | 95317 | 105 |

# Datasets

**CAFA3 Evaluation Benchmark**

- To centralize and establish a common evaluation method for this task, the Critical Assessment for Functional Annotation (CAFA), has set up common rules, guidelines, and evaluation measures which are extensively accepted by all researchers in this domain.
- The CAFA challenge releases a set of target protein sequences onto which the participants computationally annotate them with the corresponding GO terms or Human Phenotype Ontology (HPO) terms.
- CAFA3 has released its evaluation benchmarks in which the test data consists of experimentally annotated proteins that were used to evaluate the predictions submitted by participants.

# Datasets

Table 2: Specifications of the CAFA3 dataset

| Ontology Aspect | Training | Testing | No. of GO Terms |
|---|---|---|---|
| Molecular Function (MF) | 36110 | 1137 | 677 |
| Biological Process (BP) | 53500 | 2392 | 3992 |
| Cellular Component (CC) | 50596 | 1265 | 551 |
| **Total** | 66841 | 3328 | 5220 |

# Datasets

**GOLabeler dataset**

| Ontology Aspect | Training | Testing | No. of GO Terms |
|---|---|---|---|
| Molecular Function (MF) | 34488 | 679 | 652 |
| Biological Process (BP) | 51716 | 1434 | 3904 |
| Cellular Component (CC) | 49346 | 1148 | 545 |
| **Total** | 65028 | 1788 | 5101 |

Table 3: Specifications of the GOLabeler dataset

# Evaluation Metrics

- Precision: Portion of predicted GO terms that are actually correctly classified.

$$Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i|}$$

- Recall: Portion of actual GO terms that have been found and predicted correctly.

$$Recall = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i|}$$

- F1-score: It is the harmonic mean of precision and recall. Let '$P_{avg}$' be the precision and '$R_{avg}$' be the recall.

$$F1 - score = \frac{2 * P_{avg} * R_{avg}}{P_{avg} + R_{avg}}$$

# Evaluation Metrics

- Fmax: It is computed as the maximum F1-score obtained when the threshold for converting posterior probabilities to binary labels is varied from values in the range [0,1]

$$F_{max} = \frac{2 * P_{avg}^{t_{best}} * R_{avg}^{t_{best}}}{P_{avg}^{t_{best}} + R_{avg}^{t_{best}}}$$

- Smin: The semantic distance measures the distance between the actual and predicted labels based on the value of their information contents.

$$S_{min} = \min_t \sqrt{ru(t)^2 + mi(t)^2}$$

$$ru(t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{c \in Y_i - \hat{Y}_i^t} I(c)$$

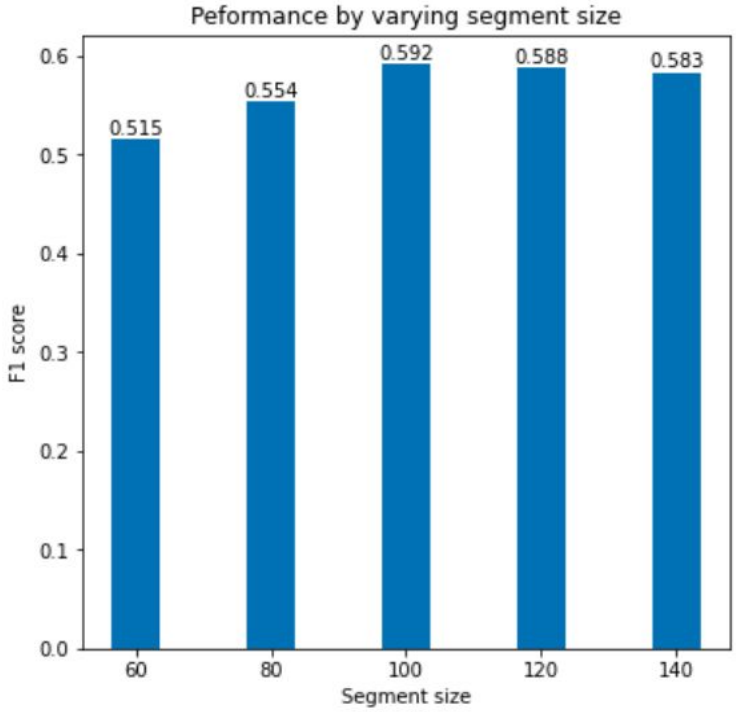
$$mi(t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{c \in \hat{Y}_i^t - Y_i} I(c)$$

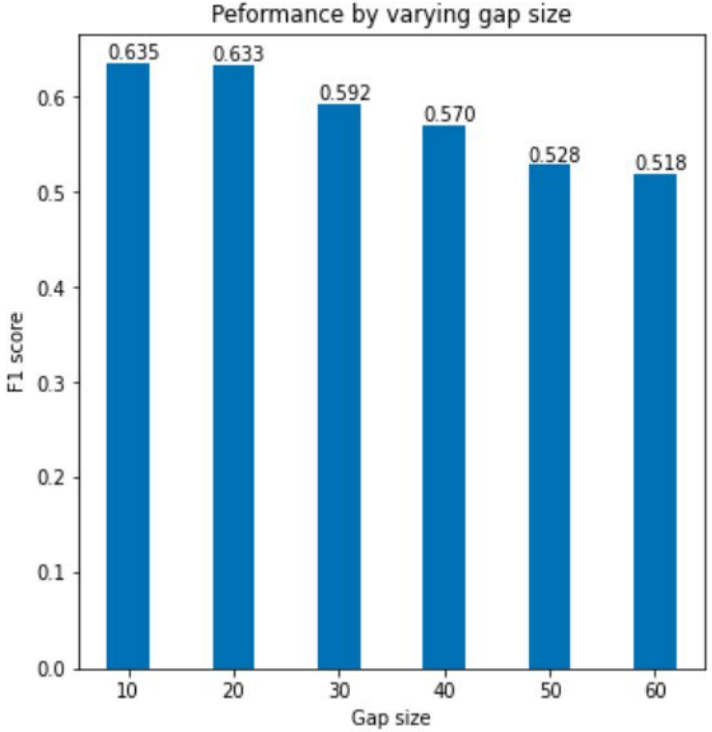
$$I(c) = -log(Pr(c|P(c)))$$

# Training Setup

- All the experiments and model training was performed in the environment of NVIDIA Tesla V100 GPUs with 32GB memory and 256GB RAM.
- In model training, the binary cross entropy loss is used with the Adam optimizer. Direct use of the binary cross entropy loss results in a high precision and low recall model. This happens due to the high imbalance of the labeled data towards the negative class.
- To curb this, we boost the recall by giving higher penalties to the false negatives compared to the false positives. So the model will learn to suppress predicting false negatives. After a lot of experiments, it was concluded that a 6:1 penalty ratio to false negatives as compared to the false positives gave the best performance. A batch size of 32 was used in training the model.

# Hyperparameter tuning



# Hyperparameter Tuning

- The two most influential hyperparameters are the segment size and the gap between two segments.
- Other basic hyperparameters consists of varying the number of epochs, n-grams with n=3,4,5, etc. The model was also trained by using 5-grams on the CAFA3 dataset but no significant improvement was observed as compared to that by using 4-grams in the proposed model.
- Moreover, the training time when 5-grams are used is approximately 1.5 times more as compared to the training time when using 4-grams.

# Results on UniProt-SP dataset

Table 7: Accuracy of OntoPred using 3-grams on UniProt-SP dataset

| Ontology Aspect | Recall | Precision | F1-Score |
|---|---|---|---|
| Molecular Function (MF) | 0.6203 | 0.5654 | 0.5915 |
| Biological Process (BP) | 0.5012 | 0.4984 | 0.4998 |
| Cellular Component (CC) | 0.7157 | 0.6188 | 0.6637 |

# Results on UniProt-SP dataset

Table 8: Accuracy of OntoPred using 4-grams on UniProt-SP dataset

| Ontology Aspect | Recall | Precision | F1-Score |
|---|---|---|---|
| Molecular Function (MF) | 0.6931 | 0.7137 | 0.7032 |
| Biological Process (BP) | 0.6484 | 0.6547 | 0.6515 |
| Cellular Component (CC) | 0.7247 | 0.7076 | 0.7160 |

# Results on CAFA3 Benchmark

Table 9: The performance comparison on the CAFA3 challenge dataset

| | $F_{max}$ | | | $S_{min}$ | | | AUPR | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | MF | BP | CC | MF | BP | CC | MF | BP | CC |
| Naive | 0.290 | 0.357 | 0.562 | 10.733 | 25.028 | 8.465 | 0.130 | 0.254 | 0.456 |
| DiamondBLAST (2015) | 0.431 | 0.399 | 0.506 | 10.233 | 25.320 | 8.800 | 0.178 | 0.116 | 0.142 |
| DiamondScore (2015) | 0.509 | 0.427 | 0.557 | 9.031 | 22.860 | 8.198 | 0.340 | 0.267 | 0.335 |
| DeepGO (2018) | 0.393 | 0.435 | 0.565 | 9.635 | 24.181 | 9.199 | 0.303 | 0.385 | 0.579 |
| DeepGOCNN (2020) | 0.420 | 0.378 | 0.607 | 9.711 | 24.234 | 8.153 | 0.355 | 0.323 | 0.616 |
| DeepGOPlus (2020) | 0.544 | 0.469 | 0.623 | 8.724 | 22.573 | 7.823 | 0.487 | 0.404 | 0.627 |
| OntoPred[a] | 0.494 | 0.480 | 0.637 | 7.878 | 21.053 | 7.870 | 0.481 | 0.467 | 0.632 |
| OntoPredPlus[a] | **0.566** | **0.545** | **0.652** | **7.276** | **20.072** | **7.611** | **0.577** | **0.536** | **0.652** |

[a]Proposed approach

# Results on GoLabeler dataset

Table 10: The performance comparison on the GOLabeler dataset

| | $F_{max}$ | | | $S_{min}$ | | | AUPR | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | MF | BP | CC | MF | BP | CC | MF | BP | CC |
| Naive | 0.306 | 0.318 | 0.605 | 12.105 | 38.890 | 9.646 | 0.150 | 0.219 | 0.512 |
| DiamondBLAST (2015) | 0.525 | 0.436 | 0.591 | 9.291 | 39.544 | 8.721 | 0.101 | 0.070 | 0.089 |
| DiamondScore (2015) | 0.548 | 0.439 | 0.621 | 8.736 | 34.060 | 7.997 | 0.362 | 0.240 | 0.363 |
| DeepGO (2018) | 0.449 | 0.398 | 0.667 | 10.722 | 35.085 | 7.861 | 0.409 | 0.328 | 0.696 |
| DeepGOCNN (2020) | 0.409 | 0.383 | 0.663 | 11.296 | 36.451 | 8.642 | 0.350 | 0.316 | 0.688 |
| DeepText2GO (2018a) | 0.627 | 0.441 | 0.694 | 5.240 | 17.713 | **4.531** | 0.605 | 0.336 | 0.729 |
| GOLabeler (2018b) | 0.580 | 0.370 | 0.687 | **5.077** | **15.177** | 5.518 | 0.546 | 0.225 | 0.700 |
| DeepGOPlus (2020) | 0.585 | **0.474** | 0.699 | 8.824 | 33.576 | 7.693 | 0.536 | 0.407 | 0.726 |
| OntoPred[a] | 0.569 | 0.418 | 0.700 | 7.978 | 32.093 | 7.345 | 0.536 | 0.391 | 0.710 |
| OntoPredPlus[a] | **0.644** | 0.463 | **0.716** | 6.984 | 30.703 | 7.084 | **0.616** | **0.443** | **0.732** |

[a]Proposed approach

# Ablation Study

Three major models are built depending on the types of features used in it,

- N-gram based features
- Skip-gram based features
- N-gram + Skip-gram based features (OntoPred)

These models are varied by using n=3 and n=4 and skip value being 1 on molecular function aspect of the UniProt-SP dataset.

# Ablation Study

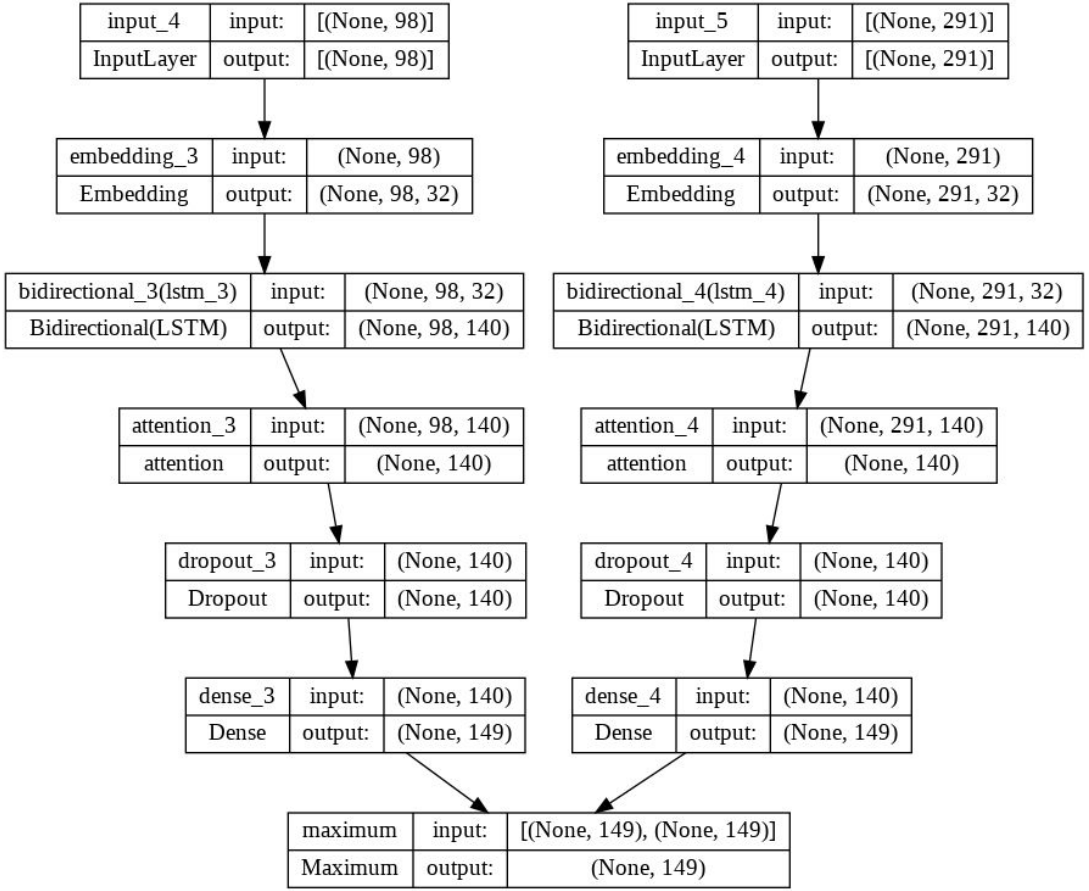



Fig. Proposed n-gram + skip-gram based model with n=3.

# Ablation Study

Table 4: Accuracy of models without attention mechanism using 3-grams.

| Model | Recall | Precision | F1-Score |
|---|---|---|---|
| N-gram features | **0.5787** | 0.4425 | 0.5015 |
| Skip-gram features | 0.5606 | 0.4060 | 0.4709 |
| N-gram + Skip-gram features | 0.5745 | **0.4693** | **0.5166** |

- It is observed that the model using both types of features performed better than being used alone.

# Ablation Study

Table 5: Accuracy of models with attention mechanism using 3-grams.

| Model | Recall | Precision | F1-Score |
|---|---|---|---|
| N-gram features | 0.6051 | 0.5573 | 0.5802 |
| Skip-gram features | 0.5861 | 0.5366 | 0.5602 |
| N-gram + Skip-gram features | **0.6203** | **0.5654** | **0.5915** |

- Use of attention mechanism has boosted the overall performance of each of the models.

# Ablation Study

Table 6: Accuracy of models with attention mechanism using 4-grams.

| Model | Recall | Precision | F1-Score |
|---|---|---|---|
| N-gram features | 0.6909 | **0.7140** | 0.7022 |
| Skip-gram features | 0.6856 | 0.7066 | 0.6960 |
| N-gram + Skip-gram features | **0.6931** | 0.7137 | **0.7032** |

- When 4-grams are used instead of 3-grams, there is a significant performance boost in terms of all the metrics in all cases

# Conclusions and Future Work

- The proposed model OntoPred was evaluated on three different datasets, UniProt-SP, CAFA3 benchmark, and the GOLabeler datasets.
- The OntoPred model outperformed the state-of-the-art techniques on the CAFA3 benchmarks. Moreover, results showed that OntoPred gave a competing performance with some methods on the GOLabeler dataset.
- It was concluded that the use of the attention mechanism improves the overall model accuracy. It was also observed from the results that 4-gram features give better performance than the use of 3-grams but not on 5-grams.
- The future work for this research may include increasing the number of annotations per sample by propagating up the hierarchical graph of GO, using protein-protein interaction features, structure information, etc.

# THANK YOU