Amalgamation of *protein sequence*, *structure* and *textual information* for improving protein-protein interaction identification

Pratik Dutta, Sriparna Saha
Department of Computer Science & Engineering
Indian Institute of Technology Patna
(pratik.pcs16, sriparna)@iitp.ac.in

Abstract

An in-depth exploration of protein-protein interactions (PPI) is essential to understand the metabolism in addition to the regulations of biological entities like proteins, carbohydrates, and many more. Most of the recent PPI tasks in BioNLP domain have been carried out solely using textual data. In this paper, we argue that incorporation of multimodal cues can improve the automatic identification of PPI. As a first step towards enabling the development of multimodal approaches for PPI identification, we have developed two multi-modal datasets which are extensions and multi-modal versions of two popular benchmark PPI corpora (BioInfer and HRPD50). Besides, existing textual modalities, two new modalities, 3D protein structure and underlying genomic sequence, are also added to each instance. Further, a novel deep multi-modal architecture is also implemented to efficiently predict the protein interactions from the developed datasets. A detailed experimental analysis reveals the superiority of the multi-modal approach in comparison to the strong baselines including uni-modal approaches and state-of the-art methods over both the generated multi-modal datasets. The developed multi-modal datasets are available for use at https://github.com/sduttap16/MM_PPI_NLP.

1 Introduction

Understanding protein-protein interactions (PPI) is indispensable to comprehend different biological processes such as translation, protein functions (Kulmanov et al., 2017), gene functions (Dutta and Saha, 2017; Dutta et al., 2019b), metabolic pathways, etc. The PPI information helps researchers to discover disease mechanisms and plays seminal role in designing the therapeutic drugs (Goncearenco et al., 2017). Over the years, a significant amount of protein-protein interaction information has been published in scientific articles in unstructured text formats. However, in recent years, there has been an exponential rise in the number of biomedical publications (Khare et al., 2014). Therefore, it becomes imperative, urgent and of extreme interest to develop an intelligent information extraction system to assist biologists in curating and maintaining PPI databases.

This pressing need has motivated Biomedical Natural Language Processing (BioNLP) researchers to automatically extract PPI information by exploring various AI techniques. Recent advancements in deep learning (LeCun et al., 2015)(Bengio et al., 2007) have opened up new avenues in solving different well-known problems ranging from computational biology (Alipanahi et al., 2015; Dutta et al., 2019a), machine translations (Cho et al., 2014), image captioning (Chen et al., 2017). Subsequently, there is a notable trend in using deep learning for solving different natural language processing (NLP) tasks in the biomedical and clinical domains (Asada et al., 2018; Alimova and Tutubalina, 2019) including the identification of protein-protein interactions from biomedical corpora (Yadav et al., 2019; Peng and Lu, 2017). Multi-modal deep learning models, combining information from multiple sources/modalities, show promising results compared to the conventional single modal-based models while solving various NLP tasks like sentiment and emotion recognition (Qureshi et al., 2019, 2020), natural language generation, machine translation (Poria et al., 2018; Zhang et al., 2019; Qiao et al., 2019; Fan et al., 2019) etc. There exist few popular multi-modal datasets which are extensively used in solving various problems in NLP like emotion recognition from conversations (Poria et al., 2018; Chen et al., 2018), image captioning (Lin et al., 2014), sentiment analysis (Zadeh et al., 2016), etc. Compared to single modal-based approaches, multi-modal techniques provide a more comprehensive perspective of the

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6396–6407
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

dataset under consideration.

Despite the popularity of multi-modal approaches in solving traditional NLP tasks, there is a dearth of multi-modal datasets in BioNLP domain especially for the PPI identification task. The available PPI benchmark datasets contain solely the textual knowledge of different protein pairs, which do not help in anticipating the molecular properties of the proteins. Hence, along with the textual information, incorporation of molecular structure or underlying genomic sequence can aid in understanding the regulations of the protein interactions. The integration of multi-modal features can help in obtaining deeper insights but the concept of multi-modal architecture, for textual and biological aspects, has not been cultivated much in the BioNLP domain (Peissig et al., 2012; Jin et al., 2018).

1.1 Motivation and Contribution

The main motivation for this research work is to generate multi-modal datasets for PPI identification task, where along with the textual information present in the biomedical literature, we did explore the genetic and structure information of the proteins. The biomedical and clinical text database is an important resource for learning about physical interactions amongst protein molecules; however, it may not be adequate for exploring biological aspects of these interactions. In the field of Bioinformatics, there are various web-based enriched archives[1][2] that contain multi-omics biological information regarding protein interactions. The integration of multi-omics information from these aforementioned databases helps in understanding the various physiological characteristics (Sun et al., 2019; Ray et al., 2014; Amemiya et al., 2019; Hsieh et al., 2017; Dutta et al., 2020). Hence, in our current work, along with the textual information from biomedical corpora, we have also incorporated structural properties of protein molecules as biological information for solving PPI task. For structural information of proteins, we have considered the atomic structure (3D PDB structure) and underlying nucleotide sequence (FASTA sequence) of protein molecules. In the BioNLP domain, collection of biological data (muti-omics information) from the text corpus is little difficult. To obtain the aforementioned information about other modalities, we need to exploit different web-based archives that are meant for biological structures.

Drawing inspirations from these findings, we have generated a protein-protein interaction-based multi-modal dataset which includes not only textual information, but also the structural counterparts of the proteins. Finally, a novel deep multi-modal architecture is developed to efficiently predict the protein-protein interactions by considering all modalities. The main contributions of this study are summarized as follows:

1. For this study, we extend and further improve two biomedical corpora containing PPI information for multi-modal scenario by manually annotating and web-crawling two different bio-enriched archives.

2. Our proposed multi-modal architecture uses self-attention mechanism to integrate the extracted features of different modalities.

3. This work is a step towards integrating multi-omics information with text-mining from biomedical articles for enhancing PPI identification. To the best of our knowledge, this is the first attempt in this direction.

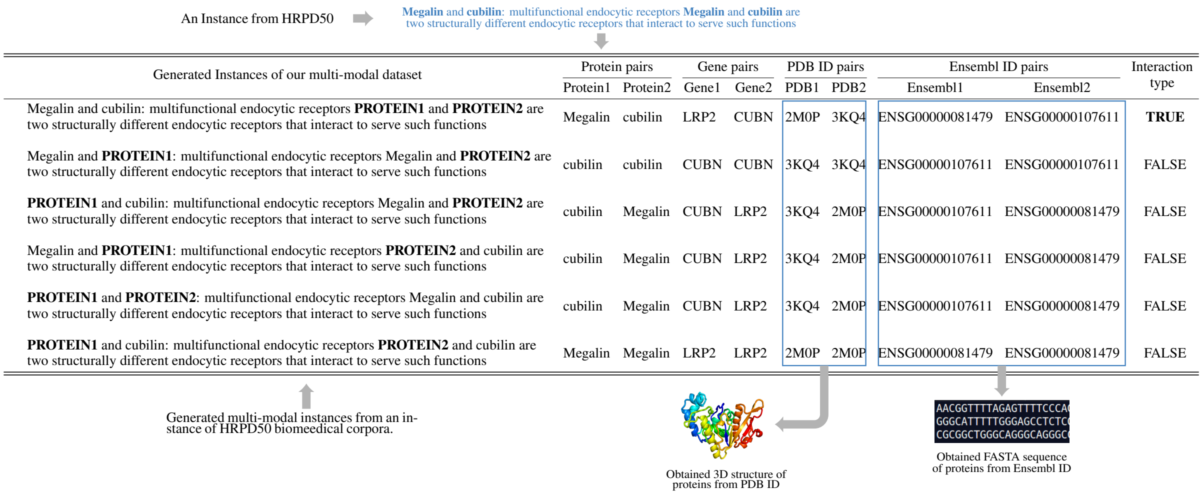
4. The results and the comparative study prove the effectiveness of our developed multi-modal datasets along with proposed multi-modal architecture.

2 Related Works

There are few works (Ono et al., 2001; Blaschke et al., 1999; Huang et al., 2004) which focus on rule-based PPI information extraction method such as co-occurrence rules (Stapley and Benoit, 1999) from the biomedical texts. In (Giuliano et al., 2006), relation is extracted from entire sentence by considering the shallow syntactic information. (Erkan et al., 2007) utilize semi-supervised learning and cosine similarity to find the shortest dependency path (SDP) between protein entities. Some important kernel-based methods for PPI extraction task are graph kernel (Airola et al., 2008a), bag-of-word (BoW) kernel (Sætre et al., 2007), edit-distance kernel (Erkan et al., 2007) and all-path kernel (Airola et al., 2008b). (Yadav et al., 2019) presented an attention-based bidirectional long short-term memory networks (BiLSTM) model that uses SDP between protein pairs, latent PoS and position

[1] https://www.cancer.gov

[2] https://www.ncbi.nlm.nih.gov/


An Instance from HRPD50
Megalin and **cubilin**: multifunctional endocytic receptors **Megalin** and **cubilin** are two structurally different endocytic receptors that interact to serve such functions

Generated Instances of our multi-modal dataset	Protein pairs		Gene pairs		PDB ID pairs		Ensembl ID pairs		Interaction type
	Protein1	Protein2	Gene1	Gene2	PDB1	PDB2	Ensembl1	Ensembl2	
Megalin and cubilin: multifunctional endocytic receptors **PROTEIN1** and **PROTEIN2** are two structurally different endocytic receptors that interact to serve such functions	Megalin	cubilin	LRP2	CUBN	2M0P	3KQ4	ENSG00000081479	ENSG00000107611	**TRUE**
Megalin and **PROTEIN1**: multifunctional endocytic receptors Megalin and **PROTEIN2** are two structurally different endocytic receptors that interact to serve such functions	cubilin	cubilin	CUBN	CUBN	3KQ4	3KQ4	ENSG00000107611	ENSG00000107611	FALSE
PROTEIN1 and cubilin: multifunctional endocytic receptors Megalin and **PROTEIN2** are two structurally different endocytic receptors that interact to serve such functions	cubilin	Megalin	CUBN	LRP2	3KQ4	2M0P	ENSG00000107611	ENSG00000081479	FALSE
Megalin and **PROTEIN1**: multifunctional endocytic receptors **PROTEIN2** and cubilin are two structurally different endocytic receptors that interact to serve such functions	cubilin	Megalin	CUBN	LRP2	3KQ4	2M0P	ENSG00000107611	ENSG00000081479	FALSE
PROTEIN1 and **PROTEIN2**: multifunctional endocytic receptors Megalin and cubilin are two structurally different endocytic receptors that interact to serve such functions	cubilin	Megalin	CUBN	LRP2	3KQ4	2M0P	ENSG00000107611	ENSG00000081479	FALSE
PROTEIN1 and cubilin: multifunctional endocytic receptors **PROTEIN2** and cubilin are two structurally different endocytic receptors that interact to serve such functions	Megalin	Megalin	LRP2	LRP2	2M0P	2M0P	ENSG00000081479	ENSG00000081479	FALSE

Generated multi-modal instances from an instance of HRPD50 biomeedical corpora.
Obtained 3D structure of proteins from PDB ID
AACGGTTTTAGAGTTTTCCCA
GGGCATTTTTGGGAGCCTCTC
CGCGGCTGGGCAGGGCAGGGC
Obtained FASTA sequence of proteins from Ensembl ID

Figure 1: An example of generating instances along with the structural and sequence counterparts of our multi-modal dataset from HRPD50 dataset. PDB ID and Ensembl ID are utilized for obtaining protein 3D atomic structure and underlying FASTA sequence, respectively.

embeddings for PPI extraction. Some of the popular deep learning based PPI extraction techniques are reported by (Shweta et al., 2016; Zhao et al., 2016; Hua and Quan, 2016; Hsieh et al., 2017).

3 Dataset Formation and Preprocessing

In this study, we have extended, improved, and further developed two popular benchmark PPI corpora, namely BioInfer[3] and HRPD50[4] dataset for the multi-modal scenario. Along with the textual information, these enhanced multi-modal datasets contain the biological counterparts of the interacting or non-interacting protein pairs. Biological information comes from the underlying FASTA sequence and the atomic structures of interacting protein pairs.

[3] http://corpora.informatik.hu-berlin.de/
[4] https://goo.gl/M5tEJj

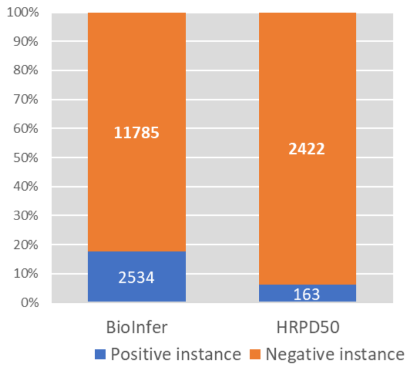


Figure 2: Statistics of positive and negative instances across our developed multi-modal datasets.

3.1 Dataset Preparation

Firstly, we have extracted data, primarily consisting of two and more protein entities, from the XML representations of two PPI corpora mentioned earlier. To simplify this complex relations among multiple protein entities, we have considered only a single protein pair at a time and found out if they are interacting or not. Among these relations, we have considered positive instances that are directly mentioned in the dataset. The other interactions are considered as non-interacting proteins, i.e., negative instances.

Consider an instance of HRPD50 dataset, ”***Megalin** and **cubilin**: multifunctional endocytic receptors **Megalin** and **cubilin** are two structurally different endocytic receptors that interact to serve such functions*”(Figure 1). In this particular example, we have four protein entities but we have considered the interactions between two proteins at a time and arrived at six possible relations (shown in table of Figure 1). Among these relations, only one pair (Megalin, cubilin) is denoted as ***interacting proteins*** in the HRPD50 dataset. Hence, the number of instances in our dataset is much higher than those in BioInfer and HRPD50 datasets.

After generating both positive and negative instances, next we have downloaded other two modalities. To download the genomic sequence and the 3d structure of proteins, the ensemble ID and PDB ID of the proteins are required to be known. But all the biological archives contain the relationships between gene and PDB ID or Ensemble ID instead

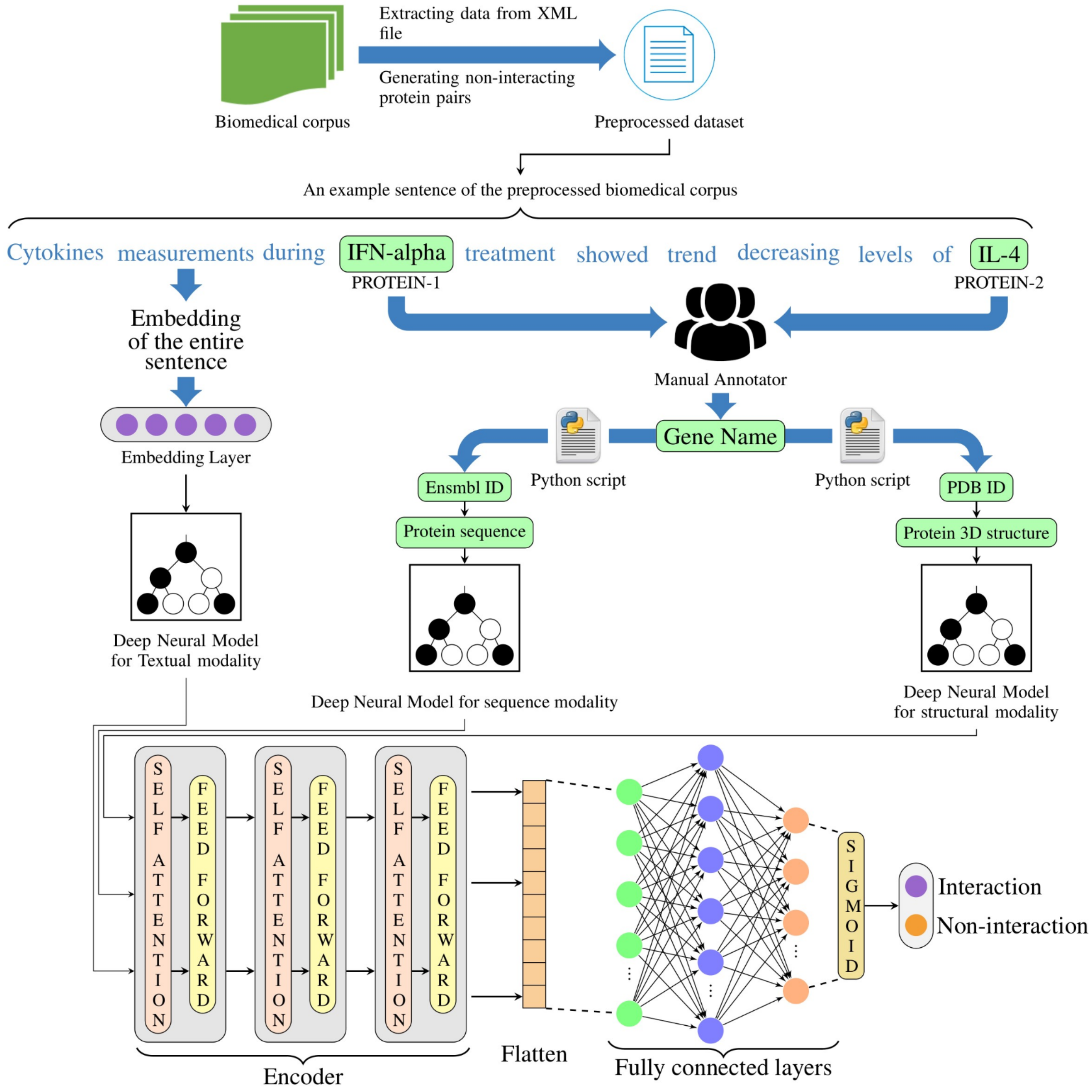

Extracting data from XML file
Generating non-interacting protein pairs
Biomedical corpus
Preprocessed dataset
An example sentence of the preprocessed biomedical corpus
Cytokines measurements during IFN-alpha treatment showed trend decreasing levels of IL-4
PROTEIN-1
PROTEIN-2
Embedding of the entire sentence
Embedding Layer
Deep Neural Model for Textual modality
Manual Annotator
Gene Name
Python script
Python script
Ensmbl ID
PDB ID
Protein sequence
Protein 3D structure
Deep Neural Model for sequence modality
Deep Neural Model for structural modality
SELF ATTENTION
FEED FORWARD
SIGMOID
Interaction
Non-interaction
Encoder
Flatten
Fully connected layers

Figure 3: An overview of the proposed deep multi-modal architecture for predicting protein-protein interactions. For each modality, we have designed different deep learning based models which are finally integrated using self-attention mechanism.

of any relationship between the proteins and aforementioned IDs. Hence, we have used manual annotation to find out the respective gene names of each protein name and then python based methodologies to find out Ensembl ID and PDB ID of each of these genes. These IDs help us in downloading the underlying genomic sequence (FASTA sequence) from [5] and structures of these proteins (3D PDB structure) from the RCSB Protein Data Bank [6] archive. The pre-processing and generation of the multimodal datasets from the biomedical corpora are pictorially depicted in Figure 1. The complete exemplified multi-modal datasets are available at the provided GitHub link.

3.2 Dataset Annotation and Statistics

A major challenge in creating the dataset is to manually encode the relationships between genes and proteins, a many to many mapping for biological reasons. Hence, to find out the genes which are more related to a particular protein, we asked three annotators who have strong biological knowledge. The disagreement between the annotators was less than 1% and the disagreement is solved by the ma-

[5] https://useast.ensembl.org/index.html
[6] http://www.rcsb.org/

jority voting. The total number of instances of the developed multi-modal datasets are shown in Figure 2.

4 Problem Formalization

Our goal is to develop a deep multi-modal architecture that can efficiently predict whether two proteins are interact with each other or not from the developed multi-modal datasets. Formally, consider the multi-modal dataset $\mathbb{D} = \{S^i\}_{i=1}^N = \{(I^i_{Text}, I^i_{Struc}, I^i_{Seq})\}_{i=1}^N$ consisting of N instances. $\forall i \in \{1, 2, \ldots, N\}, I^i_{Text}, I^i_{Struc}$ and I^i_{Seq} represent the textual, structural and sequence modality of S^i sentence/instance, respectively. The proposed PPI task for an instance S^i is mathematically formulated as

$$f_{act}\Big(f_{sa}\big(\mathbb{M}_1(I^i_{Text}), \mathbb{M}_2(I^i_{Struc}), \mathbb{M}_3(I^i_{Seq})\big)\Big)$$

Here $\mathbb{M}_1, \mathbb{M}_2, \mathbb{M}_3$ are three different deep learning based models for text, structure and sequence modality, respectively. The extracted features are fused by *self attention mechanism* (f_{sa}) which is finally fed to an *activation function*(f_{act}) for predicting protein interactions.

5 Proposed Methodology

The major steps of our proposed multi-modal architecture are shown in Figure 3.

5.1 Feature Extraction from Textual Modality

The proposed deep learning model ($\mathbb{M}_1$) for extracting features from textual modality is described in Figure 4. Firstly, we use BioBERT v1.1(Lee et al., 2019) model to provide a vector representation ($u^i \in \mathbb{R}^d$) of the textual instance (I^i_{Text}). With almost same architecture of BERT (Bidirectional Encoder Representation from Transformers) model (Devlin et al., 2018), BioBERT v1.1 is pre-trained on 1M PubMed abstracts. Here, each sentence is embedded as a unique vector of size 768 (i.e., d=768) by averaging the last four transformer layers of the first token ([CLS]) of BioBERT model. Inspired by the efficient usage of stacked Bidirectional long short term memory (BiLSTM)(Yadav et al., 2019), we use this to encode the embedded representation (u^i). In stacked BiLSTM, the l^{th} level BiLSTM computes the forward ($\overrightarrow{h^l_{u^i}}$) and backward hidden states ($\overleftarrow{h^l_{u^i}}$) which are then concatenated and fed to the next $(l+1)^{th}$ level of BilSTM layer. Therefore, the final representation (F^i_{Text}) of I^i_{Text} is obtained from the last layer (L) of the stacked BiLSTM model as

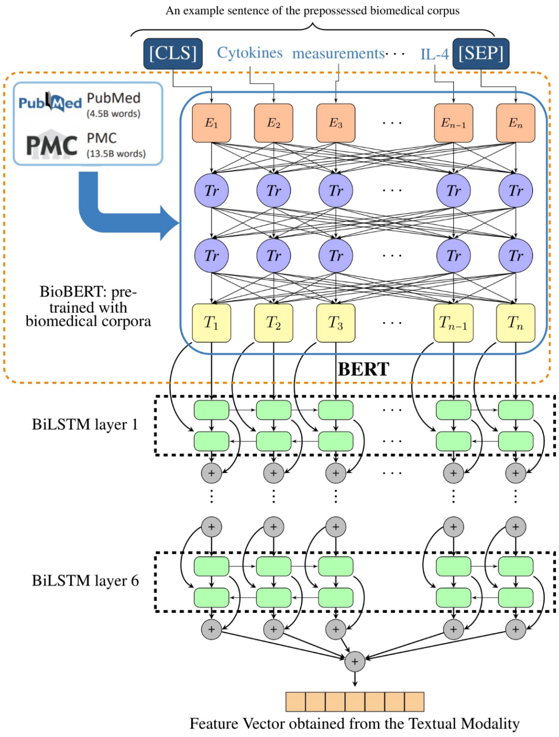


Figure 4: Proposed hybrid model combining BioBERT and stacked BiLSTM for the Textual modality.

$$F^i_{Text} = \mathbb{M}_1(I^i_{Text}) = [\overrightarrow{h^L_{u^i}} \bigoplus \overleftarrow{h^L_{u^i}}] \qquad (1)$$

5.2 Sequence Feature Extraction

Firstly, we have downloaded the FASTA sequence of protein pairs of an instance (S^i) from Ensembl genome browser. In this modality, each protein (I^i_{Seq}) is represented as string of four nucleotides, i.e., $I^i_{Seq} = \{A, T, G, C\}^+$. The underlying genomic sequence is considered as a separate channel of the text modality. Since molecular properties of protein molecules are heavily dependend on the sequence of nucleotides, we apply capsule network (Sabour et al., 2017) to capture the spatial information between the nucleotides. In this regard, firstly, we have converted all four nucleotides into one-hot vector representation, i.e., the protein is represented as a 2D matrix, $\mathbb{O} = \{0, 1\}^{4\times m}$ where m is the number of nucleotides in the sequence. Now, three convolutional layers (f_{conv}) are applied on $\mathbb{O}$ where the output of the third layer is fed to the primary capsule. Finally, the output of the primary capsule is fed to secondary capsule which

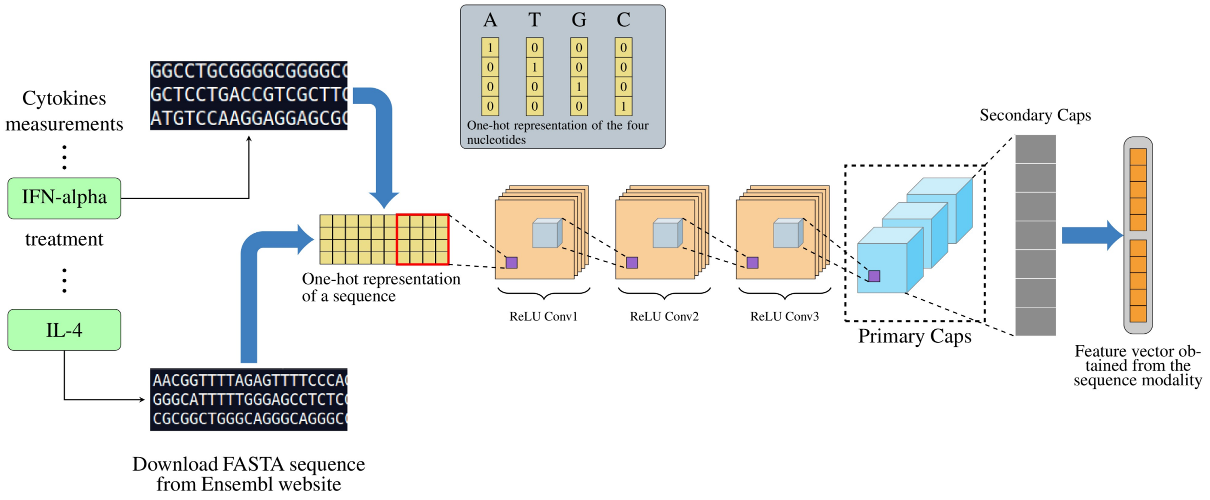


Figure 5: Capsule network-based deep model for extracting features from underlying genomic sequence of proteins.

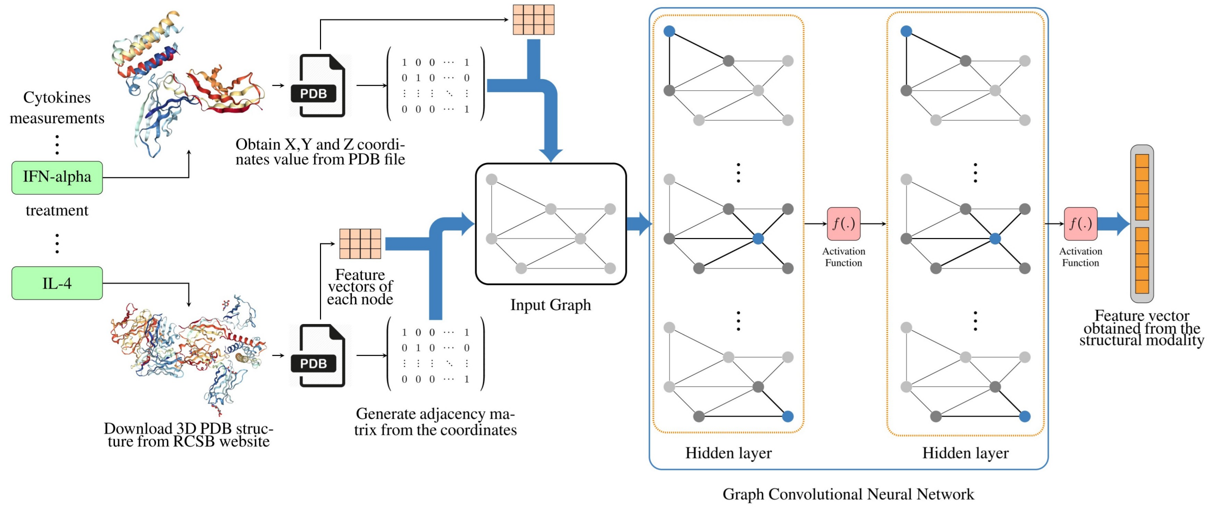

Cytokines measurements
IFN-alpha
treatment
IL-4
PDB
Obtain X,Y and Z coordinates value from PDB file
Feature vectors of each node
PDB
Download 3D PDB structure from RCSB website
Generate adjacency matrix from the coordinates
Input Graph
f(.)
Activation Function
f(.)
Activation Function
Hidden layer
Hidden layer
Graph Convolutional Neural Network
Feature vector obtained from the structural modality

Figure 6: Graph convolutional neural network-based deep model for extracting features from molecular structure of proteins.

provides the final representation (F^i_{Seq}) of the sequence modality. The final feature vector obtained from the developed deep architecture ($\mathbb{M}_2$) is

$$F^i_{Seq} = \mathbb{M}_2(I^i_{Seq}) = f_{capsule}\Big(f_{CONV}\big(\mathbb{O}\big)\Big)$$

5.3 Structural Feature Extraction

For the structure modality, firstly we have downloaded protein 3D structure from RCSB protein data bank website and obtained the atomic coordinates from the PDB file. Among all the modalities, structural modality is the most relevant modality for inferring biological information. In this modality, we have considered the atomic structure of the proteins. Inspired by the inherent capabilities of graph convolutional neural network (Kipf and Welling, 2016; Zamora-Resendiz and Crivelli, 2019) for understanding the effective latent representation of the graph, we have used it to learn a local neighborhood representation around each atom of the proteins. For this structural modality, the developed model (Figure 6) learns the chemical bonding information from the atomic structure of the proteins rather from its corresponding image. Each protein, which consists of a set of atoms $\{a_1, a_2, \ldots, a_n\}$, has an adjacency matrix, $A \in \{0,1\}^{n \times n}$, and a node feature matrix, $X \in \mathbb{R}^{n \times d_v}$. In this study, we have considered two proteins (P_1, P_2) in an instance and extracted the insightful features (y_1, y_2) using GCNN and then concatenated them for the final representation (F^i_{Struc}). The GCNN takes A

and X as inputs of the proteins and the structural feature represented as

$$F^{i}_{Struc} = \mathbb{M}(I^{i}_{Struc}) = [y_1 \bigoplus y_2] \tag{2}$$

$$y_{j|j\in\{1,2\}} = f(H^{i}_{j}, A_j) = \sigma(A_j, H^{i}_{j}, W^{i}_{j}) \tag{3}$$

Here, $\oplus, f, \sigma$ are the concatenation operator, a non-linear activation function and the propagation rule, respectively. W^{i}_{j} is the weight matrix of layer i of protein P_j and H^{i}_{j} is defined as $f(H^{i-1}_{j}, A)$ where $H^{0}_{j} = X_j$.

5.4 Attention-based Multi-modal Integration

After extracting the features of three modalities (textual, protein sequence and protein structure), we have fused the features using attention mechanism. Attention mechanism has the ability to focus on the features which are the most relevant to a context specific task. In this study, we have used self-attention mechanism of the transformer model which concatenates the final integrated feature representations ($\mathbb{F}$) of i^{th} instance (S_i) using the following formula.

$$\mathbb{F} = [W^{i}_{Text}F^{i}_{Text} \bigoplus W^{i}_{Seq}F^{i}_{Seq} \bigoplus W^{i}_{Strc}F^{i}_{Strc}] \tag{4}$$

Here, W_i represents the attention weight of ith modality. Finally, this final representation ($\mathbb{F}$) is fed to *softmax* layer for final classification.

6 Experimental Results and Analysis

In this section, we have briefly described the details of the hyper-parameters and the comparative analysis of the proposed deep multi-modal architecture. To explore the role of developed multi-modal datasets along with the proposed multi-modal architecture for predicting the protein interactions, several experiments are conducted for evaluating each modality and also different combinations of the modalities. Additionally, we have compared the performance of our multi-modal approach with various state-of-the-art methods.

6.1 Details of Hyper-parameters

In our proposed multi-modal architecture, for the final classification we have used *softmax*. *Adam* optimizer is used through out the multi-modal architecture. In stacked BiLSTM model for textual modality, 6 (i.e., L=6) layers of BiLSTM are used. In case of *structural features*, graph convolutional neural network with *two hidden layers* is used. For *sequence modality*, capsule network followed by three *ReLU convolutional* layers are used. In the developed capsule network, the number of *primary capsules* are eight along with two *secondary capsules*. Finally, *self-attention* of transformer model is utilized for integrating the features of different modalities. For self-attention, we have used three encoders which are followed by a *fully connected network* with two hidden layers. The output of the fully connected network is then fed to *softmax* for final classification.

6.2 Comparative analysis with baselines

For baselines, we have compared our multi-modal approach with three uni-modal, three bi-modal and two other multi-modal architectures.

- ***Textual modality*** BioBERT and stacked BiLSTM are utilized for this model.
- ***Protein sequence modality*** Capsule network is utilized to understand the underlying features extracted from the protein sequences.
- ***Protein structural modality*** Inspired by the effective performance of GCNN in understanding the graph representation, GCNN is applied on atomic structure of proteins.
- ***3D structural + sequence modality*** In this bi-modal architecture, GCNN and capsule network are used for structural and sequence modality, respectively. Finally, self-attention is utilized to understand the integrated features of these two modalities.
- ***Textual + sequence modality*** In this model, self-attention is applied on the extracted features of textual and sequence modality.
- ***Textual + 3D structure modality***: To learn the different attributes discussed in the text and protein structural modality, self-attention mechanism is applied to fuse them.
- ***Multi-modal approach 1*** This architecture of this baseline is the same as the proposed multi-modal approach, except the learned features of each modality are simply concatenated instead of using any attention mechanism.
- ***Multi-modal approach 2*** In this model, attention mechanism is applied for integrating

		Textual modality	Protein sequence modality	Protein structural modality	Textual + sequence modality	Textual + 3D structure modality	3D structural + sequence modality	Multi-modal approach 1	Multi-modal approach 2	Proposed approach
BioInfer	Precision	54.42	50.63	59.34	64.51	69.04	68.15	79.16	83.77	**86.81**
	Recall	87.45	83.68	91.63	87.45	88.49	89.53	87.44	86.40	**89.53**
	F-measure	67.09	63.09	72.04	74.25	77.54	77.39	83.11	85.07	**88.15**
HRPD50	Precision	90.44	86.95	91.75	91.01	94.79	93.57	96.51	96.61	**96.93**
	Recall	58.67	41.32	69.01	62.81	75.21	75.21	74.38	76.44	**78.51**
	F-measure	71.17	56.02	78.77	74.32	83.87	83.39	84.01	85.35	**86.75**

Table 1: Comparative study of our proposed deep multi-modal approach with several baselines in terms of *precision*, *recall*, *F-measure*

the features of textual, protein sequence and structural modalities. For extracting the features from textual, protein sequence and protein structure, we use BioBERT, BiLSTM and CNN, respectively.

The results reported in Table 1 illustrate the supremacy of the proposed multi-modal approach over other baselines.

6.3 Comparison with State-of-the-art

Additionally, along with the baselines, we have compared the performance of our multi-modal approach with several existing works reported in the literature. For BioInfer dataset, we have compared our proposed method with nine state-of-the-art models. These existing methods are based on different techniques like kernel-based (Choi and Myaeng, 2010; Tikk et al., 2010; Qian and Zhou, 2012; Li et al., 2015), deep neural network-based (Zhao et al., 2016), multi-channel dependency-based convolutional neural network model (Peng and Lu, 2017), semantic feature embedding (Choi, 2018) and shortest dependency path (Hua and Quan, 2016). Along with the aforementioned methods, we have also compared our approach with a recent deep learning-based approach proposed by (Yadav et al., 2019). The comparative performance analysis for BioInfer dataset is tabulated in Table 2. We have also compared our approach with nine existing approaches for HRPD50 dataset. The comparative results for HRPD50 dataset are presented in Table 3.

6.4 Discussion

By analyzing the above comparative study, we can infer that the overall performance of our proposed multi-modal approach surpasses other *baselines* and *existing methods*. Among the baseline models, proposed multi-modal approach outperforms its unimodal and bimodal counterparts. Among the uni-modal architecture, *structural modality* outperforms other two modalities which suggests the importance of *structural* modality over *textual* and *sequence* modalities. The *sequence modality* performs poorly because of its huge length (length of most of the sequences is approx 10,000 nucleotides).

Among the bimodal architectures, (*textual* + *structural*) model surpasses other bimodal and unimodal counterparts. This fusion shows improvements of 5.1% and 5.5% F-score values over the best unimodal architecture for HRPD50 and BioInfer data sets, respectively. Similarly, our proposed multi-modal architecture shows an improvement over bi-modal counterparts. Also, the proposed multi-modal architecture shows an aver-

		Precision	Recall	F-score
Proposed Model		**86.81**	**89.53**	**88.15**
(Yadav et al., 2019)		80.81	82.57	81.68
(Hua and Quan, 2016)		73.40	77.00	75.20
(Choi, 2018)		72.05	77.51	74.68
(Qian and Zhou, 2012)	63.61	61.24	62.40	
(Peng and Lu, 2017)		62.70	68.2	65.30
(Zhao et al., 2016)		53.90	72.9	61.60
(Tikk et al., 2010)		53.30	70.10	60.00
(Li et al., 2015)		72.33	74.94	73.61
(Choi and Myaeng, 2010)		74.50	70.90	72.60

Table 2: Comparative analysis of the proposed multi-modal approach with state-of-the-art techniques for BioInfer dataset.

	Precision	Recall	F-score
Proposed Model	**96.93**	**78.51**	**86.75**
(Yadav et al., 2019)	79.92	77.58	78.73
(Tikk et al., 2010)	68.20	69.80	67.80
(Tikk et al., 2010)(with SVM)	68.20	69.80	67.80
(Palaga, 2009)	66.70	80.20	70.90
(Airola et al., 2008a)(APG)	64.30	65.80	63.40
(Van Landeghem et al., 2008)	60.00	51.00	55.00
(Miwa et al., 2009)	68.50	76.10	70.90
(Airola et al., 2008a)(Co-occ)	38.90	100	55.40
(Pyysalo et al., 2008)	76.00	64.00	69.00

Table 3: Comparative analysis of the proposed multi-modal approach with other state-of-the-art approaches for HRPD50 dataset.

age improvement of 3.87% and 2.24% F-scores over multi-modal approach1 and multi-modal approach2, respectively. This improvement indicates that in addition to multiple modalities, underlying deep learning models and fusion technique contribute significantly in improving the performance of the overall architecture.

In addition, Table 2 and Table 3 indicate that the proposed multi-modal architecture outperforms the best and recent existing methods for both BioInfer and HRPD50 dataset, respectively. We have performed Welch's t-test to show that obtained improvements by the proposed approach are statistically significant. From the above comparative study, it is evident that our proposed multi-modal approach identifies the protein interactions in an efficient way and can be further improved in different ways.

6.5 Error Analysis

After thoroughly analyzing false positive and false negative instances, it can be inferred that following are the possible reasons of errors:

1. The instances which contain huge number of protein entities lead to misclassification. The maximum number of proteins in an instance of HRPD50 and BioInfer are 26 and 24, respectively; this has a huge chance of misclassification. For example: "*Mutations in Saccharomyces cerevisiae* ***RFC5, DPB11, MEC1, DDC2, MEC3, PDS1, CHK1, PDS1,*** *and* ***DUN1*** *have increased the rate of genome rearrangements up to 200-fold whereas mutations in* ***RAD9, RAD17, RAD24, BUB3,*** *and* ***MAD3*** *have little effect.*"

2. Repetitive mentions of the same protein entity adds noise that leads to loose contextual information. For example "*Here we demonstrate ...* ***CLIP-170*** *and* ***LIS1*** *Overexpression of* ***CLIP-170*** *results ... phospho-****LIS1*** *... that* ***CLIP-170*** *and* ***LIS1*** *regulate ... that* ***LIS1*** *is a regulated adapter between* ***CLIP-170*** *... MT dynamics*".

3. For sequence modality, we consider underlying FASTA sequence of proteins. The length of the sequence varies from 100 to 10000 nucleotides. This increased protein length leads to misclassification as the deep learning-based model is unable to possess this long chain of nucleotides.

7 Conclusion and Future Work

In this work, we have generated some multi-modal protein-protein interaction databases by amalgamating protein structures and sequences with existing text information available in the biomedical literature. The process of generating multi-modal datasets from PPI corpora is illustrated with some examples. Besides, we have proposed a novel deep multi-modal architecture for managing the multi-modal scenario for PPIs. For each modality (textual, protein sequence and protein atomic structure), we have developed different deep learning models for efficient feature extractions. A detailed comparative analysis proves that the proposed multi-modal architecture outperforms other strong baselines and existing models. Future work aims at enhancing sequence feature extraction methods to improve the classification performance as those suffer from low accuracy. Further there are plenty of options for improving the fusion technique to enhance the overall performance of the model.

Acknowledgements

Pratik Dutta acknowledges Visvesvaraya PhD Scheme for Electronics and IT, an initiative of Ministry of Electronics and Information Technology (MeitY), Government of India for fellowship support. Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008a. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC bioinformatics*, 9(11):S2.

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008b. A graph kernel for protein-protein interaction extraction. In *Proceedings of the workshop on current trends in biomedical natural language processing*, pages 1–9. Association for Computational Linguistics.

Ilseyar Alimova and Elena Tutubalina. 2019. Detecting adverse drug reactions from biomedical texts with neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 415–421.

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

Takayuki Amemiya, M Michael Gromiha, Katsuhisa Horimoto, and Kazuhiko Fukui. 2019. Drug repositioning for dengue haemorrhagic fever by integrating multiple omics analyses. *Scientific reports*, 9(1):523.

Masaki Asada, Makoto Miwa, and Yutaka Sasaki. 2018. Enhancing drug-drug interaction extraction from texts by molecular structure information. *arXiv preprint arXiv:1805.05593*.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2007. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160.

Christian Blaschke, Miguel A Andrade, Christos A Ouzounis, and Alfonso Valencia. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Ismb*, volume 7, pages 60–67.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Sung-Pil Choi. 2018. Extraction of protein–protein interactions (ppis) from the literature by deep convolutional neural networks with various feature embeddings. *Journal of Information Science*, 44(1):60–73.

Sung-Pil Choi and Sung-Hyon Myaeng. 2010. Simplicity is better: revisiting single kernel ppi extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 206–214. Association for Computational Linguistics.

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

Pratik Dutta and Sriparna Saha. 2017. Fusion of expression values and protein interaction information using multi-objective optimization for improving gene clustering. *Computers in biology and medicine*, 89:31–43.

Pratik Dutta, Sriparna Saha, Saraansh Chopra, and Varnika Miglani. 2019a. Ensembling of gene clusters utilizing deep learning and protein-protein interaction information. *IEEE/ACM transactions on computational biology and bioinformatics*.

Pratik Dutta, Sriparna Saha, and Saurabh Gulati. 2019b. Graph-based hub gene selection technique using protein interaction information: Application to sample classification. *IEEE journal of biomedical and health informatics*, 23(6):2670–2676.

Pratik Dutta, Sriparna Saha, Sanket Pai, and Aviral Kumar. 2020. A protein interaction information-based generative model for enhancing gene clustering. *Scientific Reports (Nature Publisher Group)*, 10(1).

Gunes Erkan, Arzucan Ozgur, and Dragomir R Radev. 2007. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1999–2007.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Alexander Goncearenco, Minghui Li, Franco L Simonetti, Benjamin A Shoemaker, and Anna R Panchenko. 2017. Exploring protein-protein interactions as drug targets for anti-cancer therapy with in silico workflows. In *Proteomics for Drug Discovery*, pages 221–236. Springer.

Yu-Lun Hsieh, Yung-Chun Chang, Nai-Wen Chang, and Wen-Lian Hsu. 2017. Identifying protein-protein interactions in biomedical literature using recurrent neural networks with long short-term memory. In *Proceedings of the eighth international joint conference on natural language processing (volume 2: short papers)*, pages 240–245.

Lei Hua and Chanqin Quan. 2016. A shortest dependency path based convolutional neural network for protein-protein relation extraction. *BioMed research international*, 2016.

Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G Payan, Kunbin Qu, and Ming Li. 2004. Discovering patterns to extract protein–protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612.

Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, et al. 2018. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*.

Ritu Khare, Robert Leaman, and Zhiyong Lu. 2014. Accessing biomedical literature in the current information landscape. In *Biomedical Literature Mining*, pages 11–31. Springer.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Maxat Kulmanov, Mohammed Asif Khan, and Robert Hoehndorf. 2017. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Lishuang Li, Rui Guo, Zhenchao Jiang, and Degen Huang. 2015. An approach to improve kernel-based protein–protein interaction extraction by learning from large-scale network data. *Methods*, 83:44–50.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. 2009. Protein–protein interaction extraction by leveraging multiple kernels and parsers. *International journal of medical informatics*, 78(12):e39–e46.

Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami, and Toshihisa Takagi. 2001. Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161.

Peter Palaga. 2009. Extracting relations from biomedical texts using syntactic information. *Mémoire de DEA, Technische Universität Berlin*, 138.

Peggy L Peissig, Luke V Rasmussen, Richard L Berg, James G Linneman, Catherine A McCarty, Carol Waudby, Lin Chen, Joshua C Denny, Russell A Wilke, Jyotishman Pathak, et al. 2012. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *Journal of the American Medical Informatics Association*, 19(2):225–234.

Yifan Peng and Zhiyong Lu. 2017. Deep learning for extracting protein-protein interactions from biomedical literature. *arXiv preprint arXiv:1706.01556*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. In *BMC bioinformatics*, volume 9, page S6. BioMed Central.

Longhua Qian and Guodong Zhou. 2012. Tree kernel-based protein–protein interaction extraction from biomedical literature. *Journal of biomedical informatics*, 45(3):535–543.

Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. 2019. Mnn: multimodal attentional neural networks for diagnosis prediction. *Extraction*, 1:A1.

Syed Arbaaz Qureshi, Gaël Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*.

Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52.

Bisakha Ray, Mikael Henaff, Sisi Ma, Efstratios Efstathiadis, Eric R Peskin, Marco Picone, Tito Poli, Constantin F Aliferis, and Alexander Statnikov. 2014. Information content and analysis methods for multi-modal high-throughput biomedical data. *Scientific reports*, 4:4411.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866.

Rune Sætre, Kenji Sagae, and Jun'ichi Tsujii. 2007. Syntactic features for protein-protein interaction extraction. *LBM (Short Papers)*, 319.

Shweta, A. Ekbal, S. Saha, and P. Bhattacharyya. 2016. A deep learning architecture for protein-protein interaction article identification. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3128–3133.

Benjamin J Stapley and Gerry Benoit. 1999. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. In *Biocomputing 2000*, pages 529–540. World Scientific.

Dongdong Sun, Minghui Wang, and Ao Li. 2019. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 16(3):841–850.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS computational biology*, 6(7):e1000837.

Sofie Van Landeghem, Yvan Saeys, Bernard De Baets, and Yves Van de Peer. 2008. Extracting protein-protein interactions from text using rich feature vectors and feature selection. In *3rd International symposium on Semantic Mining in Biomedicine (SMBM 2008)*, pages 77–84. Turku Centre for Computer Sciences (TUCS).

Shweta Yadav, Asif Ekbal, Sriparna Saha, Ankit Kumar, and Pushpak Bhattacharyya. 2019. Feature assisted stacked attentive shortest dependency path based bi-lstm model for protein–protein interaction. *Knowledge-Based Systems*, 166:18–29.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Rafael Zamora-Resendiz and Silvia Crivelli. 2019. Structural learning of proteins using graph convolutional neural networks. *bioRxiv*, page 610444.

Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. 2019. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 919–928.

Zhehuan Zhao, Zhihao Yang, Hongfei Lin, Jian Wang, and Song Gao. 2016. A protein-protein interaction extraction approach based on deep neural network. *International Journal of Data Mining and Bioinformatics*, 15(2):145–164.