MidZPPI: A Sequence-Based Approach for Predicting Multi-Label Protein-Protein Interactions in Midnight-Zone of Protein Sequence Alignments

By

Farhan Haseen Prantor
21301536

Musarrat Tasnim Roja
22101276

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
October 2025

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Farhan Haseen Prantor

21301536

Musarrat Tasnim Roja

22101276

Approval

The thesis titled "MidZPPI: A Sequence-Based Approach for Predicting Multi-Label Protein-Protein Interactions in Midnight-Zone of Protein Sequence Alignments" submitted by

1. Farhan Haseen Prantor (21301536)
2. Musarrat Tasnim Roja (22101276)

Of Summer, 2025 has been accepted as satisfactory in partial fulfilment of the requirement for the degree of B.Sc. in Computer Science on October 10, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Muhammad Iqbal Hossain

Associate Professor
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

The essential building blocks of life are known as protein, which performs a wide range of tasks in our bodies while the most important role of the proteins is interacting with each other and performing cellular activities. However, the challenging part is to predict the Protein-Protein Interactions (PPI) as they hold a complicated nature and in order to predict these interactions accurately, advanced computational approaches are required. This thesis inspects sequence-based approaches for the prediction of MultiLabel Protein-Protein Interaction (PPI) types amidst the Midnight Zone of protein sequence alignments. These proteins in Midnight Zone are hard to explore, has similar structures rarely with lack of enough features which makes it computationally challenging. By using advanced Deep Learning (DL) methods and Deep Neural Networks (DNN) on protein sequences, our study determines to perform the prediction and classification of these Protein-Protein Interactions (PPI) in the Midnight Zone of protein sequence alignments. This paper proposes a novel **MidZPPI** (**Mid**night **Z**one **P**rotein-**P**rotein **I**nteraction) prediction model, aims to label Protein-Protein Interaction (PPI) types properly on our prepared **MHS62k** dataset which consists of proteins of Midnight Zone sequence alignments. This model uses a subgraph approach based on the Protein Protein Interaction (PPI) types and uses Graph Neural Networks (GNN), Graph Convolutional Networks (GCN) and Autoencoders (AE) on this. This paper also provides a wide range of knowledge about the functionalities of other zones in protein sequence alignments. Our thesis promotes the field of bioinformatics, which is expanding every day by a variety of methods and techniques that not only aims to predict Protein Protein Interaction (PPI) types, but also contributing in discovering new ways to produce new drugs and medicines, and also the exploration of critical diseases.

Keywords: Protein-Protein Interaction (PPI); Midnight Zone; MultiLabel Classification; Deep Learning (DL); Deep Neural Networks (DNN); Graph Neural Networks (GNN); Graph Convolutional Networks (GCN); Autoencoders (AE)

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption. Secondly, to our thesis supervisor Professor Muhammad Iqbal Hossain sir for his kind support and advice in our work. He helped us whenever we needed help. Finally, to our parents without their support, it may not be possible. With their kind support and prayer, we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AE Autoencoders

BFS Breadth-First Search

$BiGRU$ Bidirectional Gated Recurrent Unit

$CBRVAE$ Convolutional Bidirectional Recurrent Variational Autoencoder

DFS Depth-First Search

DL Deep Learning

DNN Deep Neural Networks

GAT Graph Attention Networks

GCN Graph Convolutional Networks

GNN Graph Neural Networks

GO Gene Ontology

$MLPAE$ Multi-Layer Perceptron Autoencoder

$PIDE$ Pairwise Sequence Identity

PPI Protein-Protein Interaction

RNN Recurrent Neural Networks

$SCAGCN$ Subgraph Context-Aware Graph Convolutional Network

VAE Variational Autoencoder

Chapter 1

Introduction

1.1 Background

Proteins are known as vital biomolecules which perform cellular activities. Proteins act as the building block of life which performs various activities like metabolism reflex, foundational support in cells as well as molecules are also transported by them. Proteins are structured with a long chain of amino acids, they perform the functions of enzymes, construct the cell structure and transmit messages within the body, assisting in the regulation of almost all biological functions. Protein sequences play a very fundamental role in the study of the role of proteins in biological systems since their arrangement of amino acids determine their shape and role.

Proteins are formed with long chains of the amino acids. Amino acids are available in 20 different varieties, and can be merged in many various ways to form a variety of proteins. These amino acids are joined to create some chain referred to as a polypeptide. These amino acids are organized in a given order in the chain and therefore determine the shape that the protein will assume and therefore its functions within the body. Proteins perform a number of very important functions: enzymes make the chemical processes in the body faster, antibodies help us to fight the infections, hormones as well as insulin make the body, structural, and body activities well and strong, and structural proteins, like collagen, make our skin, bones, and tissues strong and supportive.

Protein-protein interactions form the basis of almost all processes that take place in our cells [16]. As individuals who come together as a team apply their individual aptitudes to accomplish a shared objective, so also do proteins collaborate with each other to execute essential duties that make our bodies operate efficiently. PPI is very important to each and every component of our cells as those interactions help our cells to work properly as well as keep us healthy.

PPIs are used in virtually all the processes within the cell [17]. As an example,

proteins are messengers, which assist in signaling between the exterior of the cell and its interior to enable the cell to react to its surroundings.Most of the biological processes are handled by PPI for example growth, response of our immune system as well as regulation of our body. These protein interactions may malfunction resulting in such serious illnesses as cancer or diabetes [7].

Consider the immune system, e.g. The antibodies are proteins that assist the body to combat infections by attaching to dangerous invaders such as bacteria and viruses. These antibodies do not, however, work alone. They must be able to communicate with the other proteins in the immune system to effectively counter these threats. Our resistance to disease is increased by this collaboration of the proteins.

There are some problems with a study of PPIs, however. Proteins are modeled, dimension, and shapes and this is based on the role and locality in the body. It is a complex nature that is not easily learned as to how the proteins interact with each other. Nevertheless, scholars have come up with superior methods of investigating these interactions. The study of shapes of proteins would enable them to know how two proteins can interact with each other, such as a key into a lock.The research on PPIs promises to promote medicine and provide new hope for those who are afflicted with complex diseases.

1.2 Problem Statement

PPI, is known as the foundation of each and every process that are biological and these process includes synchrozation of genes , reactions of immune etc.Nonetheless, such experimental procedures are not only time and labor intensive but also susceptible to errors which may include false positives and false negatives and this affects the accuracy and reliability of findings. Moreover, these approaches are not scalable over large numbers of interactions, which are present in the differing protein sequences, and notably so those interactions with proteins are yet to be experimentally characterised.

It has been essential that computational methods, especially machine learning and sequence-based methods, have evolved in order to handle these challenges. In order to overcome these types of challenges computational methods for example, sequence based methods as well as machine learning techniques can be involved [26] .These techniques use biological sequence information, such as protein sequences and protein structures, to rather efficiently and accurately predict PPIs [33]. Various features, including sequence motifs, evolutionary information and sequence alignments, are used as sequence-based methods of modeling protein interactions. Yet a major issue in the area is that it is quite difficult to predict multi-label PPIs, in which each protein can be involved in a number of interactions at the same time, and interactions can be of different label or type.

Most of the existing techniques in the given context of PPI prediction are based on single-label classifications which are not capable of reflecting the nature of multi-label interactions where proteins may interact with other functional outcomes in multiple ways. Moreover, current models tend to ignore the context where interactions take place, e.g. by taking into account sequence alignments between different protein domains, or by investigating new protein regions that might be part of possible interactions.

One of the gaps in the present PPI prediction scene is in the realm of the so-called Midnight Zone of protein sequence alignments. Protein sequence regions that are poorly characterized or understudied in terms of the biological functions and interaction are known as the Midnight Zone. Such areas have high and overly diversified sequence motifs, and thus it is not easily predictable what their effects are on protein interactions. Conserved and comprehensively studied segments of protein sequences are generally used to predict PPI, making much of the proteome uninvestigated. But, these not only explored areas hold a potential of containing critical information on the new forms of interaction that are yet to be learnt.

Although this issue has been noted, predicting PPIs in these midnight zones alignments has not been explored much. Although a lot of study has been done so far on enhancing the accuracy and scalability of the PPI prediction based on the conventional sequence motifs and conserved regions, the midnight zone is not very much explored and this holds a significant lacking. Indicatively, certain proteins can have disordered, flexible regions which are hard to obtain using conventional alignment based techniques, yet they may have an important role in the interaction process.

No special study has been done on the midnight zone, which is worsened by the difficulty of multi-label PPI forecasting. Proteins are not likely to have one interaction; instead they have multiple interactions and are able to do more complex biological processes. The conventional approaches to predicting PPI with binary or single-label interactions cannot predict the complexity of multi-label interactions, whereby a protein may be a subject of multiple different interactions. The latter is especially problematic with such areas as the midnight zone where interaction is less predictable because of the lack of conserved motifs.

This thesis aims on developing MidZPPI, a sequence-based methodology that will be used to predict multi-label protein-protein interactions in the midnight zone of protein sequence alignments. The conceptual space is called the midnight zone in the protein sequence alignments that is composed of poorly characterized or poorly explored regions of proteins, which are not sufficiently represented by the current prediction models. The regions are usually neglected since they may not have a strong sequence conservation and thus it becomes difficult to predict the interaction of these regions using the conventional methods. The model that we will develop is going to be lightweight and efficient and less computationally intensive as well as appropriate to be used in large scale data sets on average machines. It will generate numerous predictions of every protein with the consideration of the complexity of

biological systems where proteins are not engaged in a solitary interrelation.

1.3 Research Objectives

Protein-Protein Interaction prediction is one of the vital factors in the study of biological systems as PPIs have a significant role in cellular events like signal transmission, immune reaction and metabolic control. As computational biology developed, sequence-based techniques of PPI prediction have become efficient options, which are highly scalable and faster than experimental methods. Nevertheless, the existing models have serious limitations, especially on the background of the prediction of multi-label PPIs in the under-explored areas of the proteome, so-called Midnight-Zone. This thesis will approach these challenges by developing MidZPPI, a sequence-based model of predicting multi-label PPIs, in case of proteins that have very limited or no functional annotations.

Conventional approaches in order to forecast protein-protein interactions include such techniques as tandem affinity purification and yeast two-hybrid. There are however some challenges with these methods. They are very time consuming and take much effort to complete. These experimental methods also tend to generate both false positives (incorrectly detecting interactions) and false negatives (no interaction exists) thus decreasing the accuracy and reliability of the data [27].

The ultimate goal of the study is to come up with and assess a sequence-based approach to the prediction of multi-label PPIs. Conventional techniques, e.g. application of the single-label techniques cannot reflect the complexity of PPIs, in which several interactions between a group of proteins are existing concurrently. This classification model will consider predicting the Multi-Label Protein-Protein Interactions in Midnight-Zone of Protein Sequence Alignments. The goal of this thesis is to examine the importance of the Midnight Zone of protein sequences, which has generally low-sequence similarity and is weakly annotated, and hard to categorize and investigate. The model MidZPPI that we will develop is going to be lightweight and efficient and less computationally intensive as well as appropriate to be used in large scale data sets on average machines The study will reveal how sequence based approaches can help to enhance the prediction of PPIs. The main objectives of our research are mentioned below:

1. Providing knowledge about the functionalities of different ranges of protein sequence alignments.

2. Predicting Multi-Label Protein-Protein Interactions of protein sequence alignments amidst Midnight-Zone by following a sequence based approach.

3. Developing a novel prediction model MidZPPI (Midnight Zone Protein-Protein Interaction), which aims to label Protein-Protein Interaction types properly on our prepared MHS62k dataset, which consists of proteins of Midnight Zone sequence alignments.

This paper aims to develop a effective computational model which is known as MidZPPI, which enhance the prediction of protein-protein interactions as well as strengthen our knowledge of the biological functions of proteins in the Midnight Zone.

Chapter 2

Literature Review

2.1 Preliminaries

2.1.1 Protein

Protein is a fundamental element for all living things. They perform various biological functions,they help in rebuilding tissues, constructing the cells etc.The interaction between proteins is not simple but they work together to safeguard essential processes like metabolism and gene expression [8]. Such interactions are also significant to know how to identify new drug targets and how diseases develop and spread. Nonetheless, the experiments of studying these interactions are costly and time-intensive within the laboratory, prompting the development of the computer-based techniques to forecast the interactions [4][31].

Proteins are formed with long chains of the amino acids. They are known as important macromolecules which are found in every living being. Proteins play a vital role in every biological activity and process. The amino acids work as organic compounds and these compounds form a spectacular sequence. In this way, the structure of the protein is found and those specific sequences of the amino acids work as a foundation of the protein. These sequences also help to determine the protein interaction of the cell and also how the proteins do its tasks inside the cell. These sequences are mostly called the primary structure of a protein [8].

Proteins are deeply involved with our lives as they play a vital role in working as enzymes.Proteins also catalyzes the reactions of the elements like biochemicals which is very essential for our cellular life and metabolism. The components that are structural, for example collagen in the connective tissue, proteins help to form them

as they are the parts of cell membrane which are integral. Proteins also works as hormones for example antibodies helps to function the human body for defending the infections successfully.Regulating metabolic processes which is done by insulin also is an example of functioning proteins as hormones.Proteins also works in controlling the signal of the human cells and in gene regulation as well [17].

2.1.2 Protein-Protein Interaction

Proteins interact with other proteins are known as protein-protein Interactions, which is the fundament and base of every biological processes, such as transduction of signal, regulation of genes and responses of immune [31]. Proteins bind to other proteins (so-called protein-protein interactions), and this is essential to mostly cell processes, such as transmission of signal, metabolic regulation, as well as the assembly of complex molecules [18]. PPI research can be vital in diseases including cancer, where tumor development may be caused by an abnormal signaling pathway. These interactions are very difficult to predict given the large diversity of proteins and the complexity of interactions between proteins [31]. Recent developments in computational predictive methods, especially sequence-based methods have contributed significantly to PPI prediction. Nevertheless, such models are not easily explainable, and it is a hindrance to their usage in real biology [35].

Protein interaction prediction has grown into a significant field of study to establish the mechanism of cell functioning, novel drug discovery, and disease genetics. Although the traditional methods of identifying PPIs have been experimental, including yeast two-hybrid also protein microarrays, these techniques are time-consuming and resource intensive [28]. In the past, experimental techniques that included yeast two-hybrid or mass spectrometry were utilized in order to predict the protein protein interactions.

They are useful in information and may be expensive and hard to apply when dealing with large amounts of data. To deal with these types of challenges, researchers came to sequence based methods in which the interactions are approximated depending on the amino acid sequences of proteins. These are cheaper and can be applied on large datasets but they also possess some difficulties. Proteins are complex structures, and there are numerous proteins to study, and one cannot possibly capture all the possible interactions properly [25] [31].

Protein-Protein Interactions are very vital in the mechanism of our body. Proteins will also assemble themselves into groups to accomplish significant functions in the body like a group of individuals collaborating to reach a destination. These interactions make the cells communicate with each other, control the energy production, and they even fight infections. It is hardly possible that some process in the organism does not require PPI [4][31].

Protein-Protein interaction can be studied to learn a lot about how our bodies work, and how diseases develop. Problems of protein interactions like such are linked to numerous diseases including cancer or Alzheimer disease. By getting to know about PPIs, scientists can learn how to treat such diseases [31]. PPIs are used in virtually all the processes within the cell [17].

The capability to estimate protein interaction is in reality highly useful in medicine. Through the assistance of identifying the collaborating proteins, we acquire a clearer picture of the cause of certain diseases. This information can be of immense assistance during the drug development because the scientists can utilize this information to identify new drug targets. One such area is whereby in the case of a disease caused by the interaction of proteins, a drug can be tailored to inhibit such interaction and potentially treat the ailment [25].

PPI prediction in drug development also helps the researchers to design better drugs. They are not required to conjecture on which treatments they can apply since they are able to focus on specific proteins that trigger the disease. This lowers the cost and speed of the drug development process. It also helps in the individualization of treatment i.e. it can be tailored to every individual patient [31][4].

PPI can also be used to diagnose diseases. Protein markers are found in a variety of diseases and thus may help the doctor in diagnosing the disease at an early stage. There is also one method to include more biological data in order if we want accurate PPI predictions.This can be accomplished through one method and that is through the use of knowledge graphs which are diagrams that demonstrate the relationship of proteins and also how they interact among themselves in the body. These graphs hold valuable data such as functions of proteins, their positions in the cell and the other proteins with which the proteins interact [31]. With this extra information, models can make better predictions and give more information about protein interactions [31].

One of the areas of science in which predicting protein-protein interactions can be relevant is in computational biology, to discover the mechanism by which cells function and to understand the pathogenesis of disease. Prediction of these interactions traditionally was based on sequence based methodologies, in which the researchers compare the amino acid sequences of proteins through various algorithms. Nevertheless, these techniques are not sufficiently described to comprehensively describe the nature of protein-protein interactions. More recently, machine learning techniques in spectacular graph-based models such as Graph Neural Networks have significantly enhanced PPI predictions, especially of complex proteins in extreme environments such as the Midnight Zone [32].

2.1.3 Daylight Zone

Daylight zone pairs of protein sequences hold on high similarity.Basically daylight zone pairs of proteins hold more than 40 percent sequence identity (high sequence identity), the protein sequence pairs hold between them [3]. Within this zone, the conventional methods of alignment e.g. pairwise sequence alignments work remarkably well since proteins sharing high sequence identity may share similar functions and structures. Much of the initial research in protein sequence aligned and homology detection has been on the Daylight Zone, which is a part of protein-protein relationships that are more readily detected and argued [5].

Although studies in this field have produced some significant information about the proteins that are closely related, they provide very minimal information about the proteins that are far apart or those that have evolved along divergent paths. The Daylight Zone, therefore, has been thoroughly investigated, and the techniques of the analysis of these alignments are strong and highly developed.

2.1.4 Twilight Zone

Protein sequences whose identity falls between 20 percent to 35 per cent are known as the Twilight Zone [3]. At this point, methods of sequence alignment will start to fail and the precision of observing evolutionary relationships and functional similarities will begin to fade away. Nonetheless, much work has been done in the twilight zone to enrich sensitivity of alignment algorithms and techniques, with Hidden Markov Models (HMMs) and profile alignments being one example of the technologies capable of recognizing remote homologous relationships [1].

Twilight zones sequence identity is moderate and the protein sequences hold identity from 20 to 35 percent [3]. The Twilight Zone is a currently studied region as despite moderate sequence identity, proteins in this spectrum can have significant biological functions in common.The twilight zone is the range of protein sequence pairs that are 20 to 35 percent similar, and here in the twilight zone, sequence-based alignment algorithms are not as accurate [3]. Although major studies have been undertaken in this area such as the development of highly sophisticated alignment methods and models to identify remote homologs, the Midnight Zone goes beyond that, to include even lower sequence identities.Twilight zone holds functional similarity as well as structural in proteins. These similar characteristics cannot be detected with the help of sequence alignment tools or methods. Researchers have shown that proteins that hold less than 15 percent sequence identity which can also hold the same and similar 3D functions as well 3D structures which is called ”hidden homology”. Moreover, in order to detect such kinds of relationships, we need more complicated methods for predicting patterns in sequence data [2].

2.1.5 Existing Models

Deep learning models are black-box, meaning they cannot be used in important domains like drug discovery or personalized medicine, where one needs to understand the underlying mechanisms behind the model.Explainable AI (XAI) methods are currently being studied to resolve this problem, which provides methods to understand and visualise how models reach certain PPI predictions.It is especially vital with regard to testing the validity of models in biological studies where scientists need to believe that the models predicting are valid and biologically meaningful. Through improving the explainability of models such as GOProteinGNN, one can more readily follow the connections between particular sequence features to the predicted interactions, which is essential to downstream biological uses [31][35].

Although the development of sequence-based and knowledge-enhanced PPI prediction models have achieved progress, there are still some challenges. A large number of models, even those that do not rely on representations using protein language, have issues with scalability, especially when attempting to recreate large PPI networks. Also, the combination of knowledge graphs enhances the accuracy of prediction, but, again, it is not a silver bullet. Several models do not do a great job at maintaining the structural and functional characteristics of PPI networks especially with large-scale datasets [31] [35].

Researchers have shown that knowledge graph based machine learning models lead to better predictions in PPI. Indicatively, a classifier called GOProteinGNN, integrating biological data with protein sequence data has become much more effective in prediction of PPIs.GOProteinGNN is constructed around a type of neural network called Graph Neural Networks (GNN), which is designed to process more complex and interconnected data (i.e. a knowledge graph). In this model, the protein sequence is supposed to be a sentence, and the knowledge graph serves to acquire an idea of the greater picture [31].

Hybrid knowledge graphs containing the Gene Ontology (GO) have recently become important for enhancing sequence-based model prediction. For example, GOProteinGNN integrates protein functional knowledge as a GKI layer to guide the learning [31]. This improved representation can therefore allow a better understanding of biological roles and binding linkages of proteins, which are indispensable for accurate PPI prediction. Such external sources of knowledge significantly outperform models depending only on sequence data.Such external sources of knowledge greatly out-perform models that use just sequence data .Although sequence-based methods have achieved impressive progress in prediction of PPIs, they remain not transparent [31].

In order to comprehend the predictions of these models, such tools as SHAP (Shapley Additive Explanations), and LIME (Local Interpretable Model-agnostic Explanations), are applied. Such tools assist to describe the significance of every feature

in the prediction. In addition, other models which are built around significant segments of the protein sequence, such as transformer networks, emphasize the areas of the protein which are of interest in both determining interactions [34].Most recently, explainable AI (XAI) methods including SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explainings) have been developed to enhance the interpretability of DL models [11] [13]. Such techniques shed light on the features that make a model most useful in predicting, and it is simpler to interpret and justify the predictions made by a model within a biological setting [28].

Also is the issue of cross-species generalization where models that are trained on one species tend to perform poorly when used with other species.Whereas much has been done to date in predicting Protein-Protein Interactions (PPIs) with sequence-based approaches, protein language models, and knowledge graphs have demonstrated enormous potential in terms of accuracy. Nevertheless, there is still the issue of bringing these models to light. To continue with the advancement of protein-protein interactions (PPIs) in the future, models that can not only predict the interactions but also shed light into the mechanism would be required [31][35].

Adding protein structural data , i.e. showing how proteins fold and complement each other on a molecular scale, can be one of the improvements. Protein structure in 3D is essential to the mechanism by which a protein interacts with other proteins, and the inclusion of such data may assist the model in making even more accurate predictions about PPIs [31].

2.2 Related Work

2.2.1 Midnight Zone

The Midnight Zone proteins are a special group of proteins which exist under extreme conditions, i.e., environments that are deep into the body or other organisms that are very remote. The proteins are also known as the midnight zone in the ocean where the natural light fails to penetrate, symbolizing the role of these proteins in the unsociable or less known conditions. However, they are not simple to study because most of the time they are complexly interacting with other proteins and it is difficult to determine how they act on one another. The interaction of these proteins is critical to learning how our bodies respond in the presence of stress or other extreme conditions, such as during diseases or extreme conditions.

Midnight zone proteins holds low sequence identity, mainly less than 15 percent for a sequence alignment with greater than 250 amino acids [21]. Sequences of proteins

found on the Midnight Zone are potentially quite different to those of organisms found on the surface, and it would be more challenging to predict interactions with these sequences by most conventional sequence-based modeling techniques. The protein of these organisms have special structural and functional changes and seems to be an exciting object of investigation. The experimental data on PPIs in Midnight Zone proteins, however, are limited and, therefore, computational prediction techniques promise to be a substitute [28]. Furthermore, these proteins have an evolutionary history that is found to be underrepresented in databases thus making it equally challenging to predict. The challenges can be addressed with sequence-based methods, especially methods that integrate deep learning, to learn based on existing protein data and predict interactions of proteins that are understudied [28].

PPI Prediction in Midnight Zone Proteins is not easy.The main problem with predicting PPIs of Midnight Zone proteins is that few annotated datasets are available. Such proteins undergo evolution in extreme conditions, so they can develop some unusual sequences that are not well-adapted in traditional models. This leads to the demand to have particular models that can model these unique sequences and accurately predict interactions. [28].In case of of 2 proteins,there is highly possibility of founding those proteins in midnight zones [21]. Conventional approaches to predicting PPI have been structural based or based on established interaction databases, including BioGRID and STRING. But, sequence-based approaches do not require experimental structures, which can be both expensive and time-intensive. [34].

We have choosed to cover the Midnight Zone of protein sequence alignments in our study since it is a blank frontier in the field of protein bioinformatics. Though much research has already been done in the Twilight Zone, the Midnight Zone has a vast potential to be discovered. The reason behind the interest in this zone is that low-sequence identity proteins can still have conserved structural motifs or be carrying out a similar function even though sequence similarity is no longer evident. Even though both the Twilight Zone and Daylight Zone have already been extensively studied and numerous studies have been done on them, the Midnight Zone provides a novel chance to study the mostly unexplored fields of protein sequence alignments. Through the Midnight Zone, our study will help reveal the concealed functional and structural correlations among proteins with low sequence identities that will contribute to our learning of protein evolution and make new biological findings.

2.2.2 Machine Learning Models

Predictive algorithms PPI prediction by sequence usually relies on machine learning algorithms, trained on large datasets of protein sequences and known interactions. The techniques are such as, feature extraction, model training, representation learning. Extraction of useful features of protein sequences, including amino acids composition, physicochemical characteristics, and evolutionary profiles. Then, predicting with the help of classifiers (SVM, Random Forest) or neural networks (CNNs,

RNNs) using these features. Recent developments include deep learning methods such as sequence embeddings with models like BERT, which have brought a disruptive change to the world of protein sequence representation and processing [34]. PPI prediction has become more accurate and informative because large quantities of sequence data can be analyzed using the deep learning models. Models such as BERT, which are transformer based or specialized such as ProtBERT have achieved new benchmarks in PPI prediction [34].

The initial methods of PPI prediction were based on simple characteristics of sequences, amino acid composition, secondary structure, and hydrophobicity [22]. Such techniques used the classical machine learning models like Support Vector Machines (SVM) and Random Forests (RF) to determine protein pairs as either interactors or non-interactors [28]. Although these methods provide informative data, they are prone to overlooking more complex interactions of proteins, in particular when the sequences are evolutionally divergent [10].

Conventionally, the models of PPI prediction that are based on sequences address the similarities between protein sequences, assuming that similar protein sequences have a higher probability of interaction [35]. Such techniques are usually computationally efficient, but can have difficulties with new or less well characterised proteins. Recently, deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been used to extract patterns in raw sequence data. These models can be used to capture more complex associations between proteins though they still have a problem with generalizing to proteins associations that are not seen [31].

Sequence based techniques operate by deriving properties out of protein sequences in order to determine the possible interaction of proteins. The machine learning models such as the Support Vector Machines (SVM) and the random forest find it easy to analyze the data [32]. Researchers are embracing computational techniques, particularly, machine learning to increase the accuracy of prediction and minimize costs [26] [33]. The typical ML algorithms used in the prediction of Protein-Protein Interaction (PPI) are support vectors [12], random forests [14], logistic regression[9], and k-nearest neighbors[19]. The majority of PPI prediction models use manually chosen protein sequence, structural, and evolutional characteristics.

Deep learning models are also powerful at predictions, however, they can be black boxes. This is also not good in the biological sphere where it has been discovered that it is important to know the reason behind predictions [34]. Protein-Protein Interactions can be predicted using deep learning models using protein sequences since they are able to locate complex patterns in biological data. The reason behind the use of convolutional neural networks (CNNs) is that they are able to identify significant patterns in the protein sequences and in the smaller ones, in particular, which are critical in predicting protein-protein interactions. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, however, operate by sequentially looking at the protein data assisting them to learn the interaction

of the amino acids with each other at longer distances [34].

The GNN-PPI model predicts novel protein-protein interactions using graph-based representations, and provides impressive performance by considering network topology and sequence data [32]. Graph Attention Networks (GATs) make PPI prediction more accurate with the help of attention mechanisms. These models enable the system to pay more attention to the more relevant interactions in the graph, which enhances the accuracy of the predictions in the complex case. The GATs have proven to perform better in comparison to traditional GCNs in diverse datasets through dynamic adjustment of the significance of neighboring nodes depending on the nature of the network [32].

Besides, transformer-based models, including Graph-BERT, have also been used to predict PPI, where protein sequence data of large scale has been used to amplify the performance of interaction prediction tasks [32]. Large-scale pre-trained models have become a successful technology in natural language processing and have been leveraged to create Protein Language Models (PLMs) as an effective technology in PPI prediction. GOProteinGNN is a novel graph-based PLM approach, which integrates protein knowledge graphs with sequence information to provide a more biologically meaningful interpretation of PPIs. The integration approach has the merit of enriching protein representation with more biological context, and further provides better predictions of PPI identification and interpretation [31].

The techniques which are computational have brought a result in PPI prediction. These computational techniques are known as deep learning as well as machine learning. The utilization of protein sequence data has brought a revolutionary change in large scaled predictions [22] [10]. To determine the protein interactions, the amino acid sequences work as input and all these things are totally dependent on sequence based methods.Also these methods or techniques can be separated into traditional sequence based techniques [28]. Proteins are made up of linear arrays of amino acids and their structure and functional properties Using classifiers (SVM, Random Forest) or neural networks to make predictions based on these features [34].

Chapter 3

Methodology

3.1 Data Collection

According to our study needs, we firstly needed to collect protein-protein interaction data from reliable sources. STRING [24] database has been one of the vast and most reliable and sources of analyzing interactions among proteins and visualizing interaction networks of protein nodes. STRING [24] database contains interaction networks for multiple species to work on. For our study, we have decided to work with human proteins for the prediction of protein protein interactions. We collected protein-protein interactions for human from STRING [24] database under the organism ID 9606, used for indicating humans, and collected the protein-protein interaction files as protein actions file. Each row of our collected actions file contains a pair of proteins with their names as IDs which indicates the interaction instance for that protein pair. The mode indicated what the interaction of that protein pair is labeled as, which also determines the functionality of that protein pair.

After that, we needed to collect sequences for those interacting proteins regarding our purposes on sequence based protein-protein approach. UniProt [30] database has been one of the most well organized websites with various protein related toolkits. UniProt [30] database stores protein details of many species alongside many noteable tools like BLAST, Align etc. which are popular in protein analysis. From UniProtKB's [30] database, we collected human protein sequences using advanced search options for our thesis purposes. We have further stored the suquences data in an individual file. We only stored sequences that are cross-referenced with STRING [24] database as our protein-protein interaction data has been collected from that website.

3.2 Dataset Creation

There are primarily 3 types of ranges of protein sequence alingments which are Daylight Zone, Twilight Zone and Midnight Zone. Midnight Zone of protein sequence alignments are defined as any aligned protein pairs having HVAL less than -5, which means a pairwise sequence identity (PIDE) less than 15% for aligned sequences having more than 250 amino acids [21]. So, proteins in Midnight Zone has less chances of having similar structures [21]. Our thesis aims to work with the proteins in Midnight Zone of protein sequence alignments. So, we proceeded in creating the dataset accordingly.

We firstly decided to use PIDE as the measure for filtering protein pairs in Midnight Zone of protein sequence alignments from our collected data. For a pair of proteins, pairwise sequence identity is defined as the percentage of the number of identical amino acids in both proteins divided by their alingment length [6]. This alignment length contains both the number of residues and internal gaps inside the alignment [6]. The formula for pairwise sequence identity is such [6]:

$$\text{PIDE} = \left(\frac{\text{Identical Amino Acids in Alignment}}{\text{Amino Acids in Alignment} + \text{Internal Gaps in Alignment}} \right) \times 100\%$$

As we are dealing with sequence alignments of a protein pair, there are no clear and strict rules mentioned on the alignment types for pairwise sequence identity formula. The alignment can be either a global alignment or a local alignment.For a global alignment, it considers the entire length of the alignment for both the proteins for finding identical residues. But, a local alignment only considers the length of the dense location in the alignment for both the proteins for finding amino acid matches. We decided to use global alignment as our definition for the PIDE calculation.

Our selection of global alignment utilizes the full alignment length for better filtering protein pairs in Midnight Zone of protein sequence alignments and increases the probability of extracting more meaningful features in a zone with very low protein structure similarity. After that, we have filtered protein pairs having lower than 15% pairwise sequence identity for the creation of our dataset consisting of protein pairs in Midnight Zone of protein sequence alignments.

Our thesis focuses on a sequence-based approach for predicting protein-protein interactions. For pairwise sequence identity formula, we consider proteins sequences having greater than 250 amino acids in it. So, we filter out protein sequences which are lower than the lower threshold of 250 amino acid. We also set an upper threshold of 2000 amino acids. This upper threshold is used to ensure a consistency in sequence length to train our model properly.

Lastly, we conbine data from both of our collected sources after applying filtering process. We combine protein sequences collected from UniProt [30] database to their respective proteins listed in protein-protein interaction data which is collected from STRING [24] database. After applying filter to the combined data above, we create our **MHS62k** dataset consisting of around 62,000 protein-protein interactions in it.

3.3 Dataset Description

Our prepared **MHS62k** dataset contains 62,920 protein-protein interaction (PPI) rows in it. These protein-protein interactions (PPI) are of protein pairs in Midnight Zone of protein sequence alignment. A number of 3,143 proteins have been filtered in this process. These proteins have a pairwise sequence identity (PIDE) lower than 15% and sequence length greater than 250 amino acids. Also, the maximum length for protein sequences is 2000 amino acids.

MHS62k dataset consists of 9 rows, those are sequence_a, sequence_b, item_id_a, item_id_b, mode, action, is_directional, a_is_acting, score [24]. Here, sequence_a and item_id_a corresponds to 1st protein's sequence and it's name ID in STRING [24] database for the interaction pair. Similarly, sequence_b and item_id_b corresponds to 2nd protein's sequence and it's name ID in STRING [24] database for the interaction pair. The "mode" corresponds to the type the interaction of the protein pairs. There are 7 types of interactions which are as such: reaction, binding, activasion, inhibition, catalysis, expression and ptmod (post-translational modifications). MHS62k dataset requires a Multi-Label Classification approach to predict the protein labels as interacting protein pairs might have more than one label for them. Also, it is a suitable dataset for constructing a Graph Dataset as each protein-protein interaction in the dataset can be considered as an edge alongside both proteins in the interaction pair can be denoted as nodes.

3.4 Exploratory Data Analysis

Our prepared dataset MHS62k contains 62,920 protein-protein interaction rows. It has 9 columns in total which are sequence_a, sequence_b, item_id_a, item_id_b, mode, action, is_directional, a_is_acting, score [24]. The score column is only numerical, all the other columns are categorical. After exploring the dataset, we have seen that the column "action" contains the most missing values. The rest of the columns have no missing values in them. The "action" column has a total of 58,112 missing values in it among 62,920 rows. So, the "action" value is missing for the 92.36% of rows in the dataset. We will be considering to remove the "action" columns for this reason.

In MHS62k dataset, there are 7 types of interactions which are as such: reaction, binding, activation, inhibition, catalysis, expression and ptmod (post-translational modifications). These interaction types are labeled in "mode" column. Also, the "is_directional" and "a_is_acting" columns have "t" and "f" in it for denoting true/false. The distributions of columns are shown in Figure 3.1.

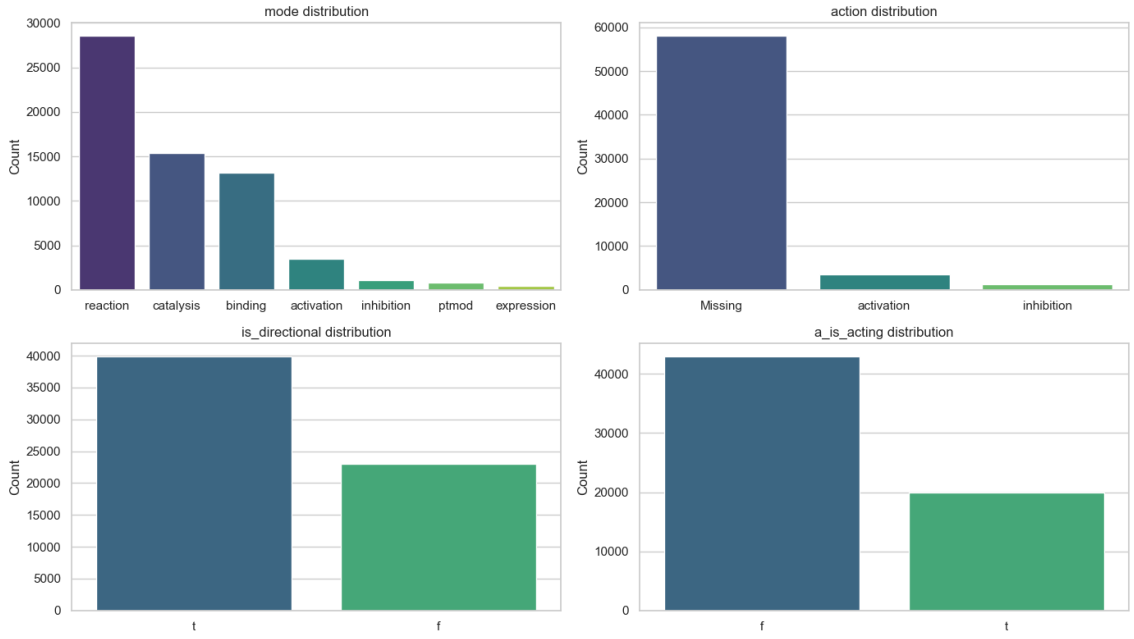

Figure 3.1: Sample Distributions

In MHS62k dataset, Among the 7 protein-protein interaction types, "reaction" has the highest number of samples and "expression" has the lowest number of samples. The numbers of protein protein interactions are shown in Table 3.1.

Interaction Mode	**Interactions**	**Percentage (%)**
Reaction	28,582	45.43
Catalysis	15,396	24.47
Binding	13,124	20.86
Activation	3,500	5.56
Inhibition	1,094	1.74
PTMod	768	1.22
Expression	456	0.72

Table 3.1: Protein-Protein Interaction Labels

Now, for "action" column, 3502 protein-protein interactions are labeled as "activation" as interaction action, which belongs to the 5.57% of interactions in the dataset. And, 1306 protein-protein interactions are labeled as "inhibition" as interaction action, which belongs to the 2.08% of interactions in the dataset.For "is_directional" column, 39904 protein-protein interactions have "t" as the value for it indicating directional, which belongs to the 63.42% of interactions in the dataset. And, 23016 protein-protein interactions have "f" as the value for it indicating not directional, which belongs to the 36.58% of interactions in the dataset. For "a_is_acting" column, 42968 protein-protein interactions have "f" as the value for it indicating the

first protein denoted as A is not acting, which belongs to the 68.29% of interactions in the dataset. And, 19952 protein-protein interactions have "t" as the value for it indicating the first protein denoted as A is acting, which belongs to the 31.71% of interactions in the dataset.

Protein sequences for the interacting proteins are stored in "sequence_a" and "sequence_b" columns. Both of these columns have a mean length of 933.21 amino acids alongside a standard deviation length of 646.72 amino acids. The minimum length of the sequences, as mentioned earlier, is 250 amino acids with a maximum length of 2000 amino acids. 10%, 25%, 50%, 75%, 90%, 95% percentile lengths for sequences are as follows 269, 305, 557, 1564, 1828, 1939 amino acids. Distribution of protein sequences based on length is shown in Figure 3.2.

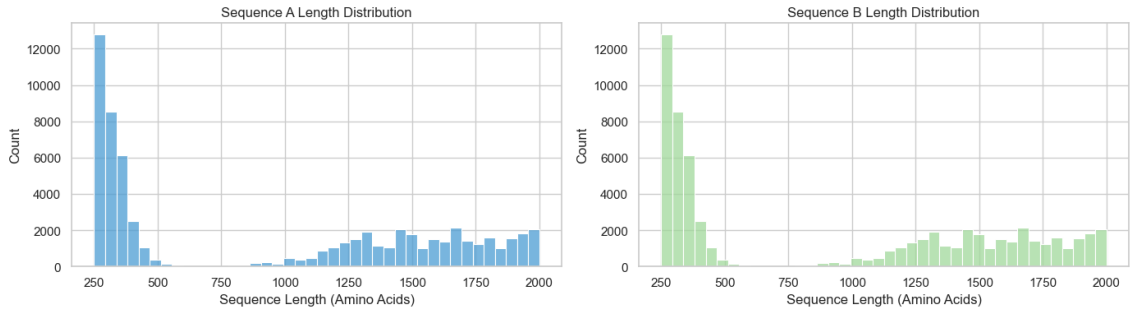

Figure 3.2: Sequence Length Distributions

3.5 Dataset Pre-Processing

For pre-processing MHS62k dataset, firstly we will be going through checking if any duplicate rows exists or not. Based on that, we would be dropping duplicate rows as it would introduce bias in our model's training phase. Among the 62,920 protein-protein interaction rows, there were no duplicate rows as we had filtered rows before creating our dataset. This filtering resulted in droping duplicate rows before procceding to preprocessing our dataset. So, no duplicate rows were found and dropped because of it.

MHS62k dataset contains sequence_a, sequence_b, item_id_a, item_id_b, mode, action, is_directional, a_is_acting, score columns. The "action" column consists of a total of 58,112 missing values in it among 62,920 rows. So, the "action" value is missing for the 92.36% of rows in the dataset.This large amount of value is quite difficult to reproduce. Applying imputation techniques for imputing values for this column would result in large amount of misinformation which will result in improper prediction. So, we have also decided to drop this "action" column based on the above mentioned reasons.

After that, "is_directional", "a_is_acting" and "score" columns are derived from our

collected protein actions data from STRING's [24] database. "is_directional" and "a_is_acting" columns of STRING [24] database corresponds to provide better visualization of their generated interaction networks which is useful to understand for users on how the protein nodes are associated with each other. Also, the column "score" is used to measure the evidence of an interaction pair based on their collected sources. So, these 3 columns "is_directional", "a_is_acting" and "score" are not meaningful for our study purpose on sequence-based approach on predicting protein-protein interactions in Midnight Zone of protein sequence alignments. So, we will be dropping these columns due to their irrelevance to our study.

After the above mentioned pre-processing steps, we will be working with sequence_a, sequence_b, item_id_a, item_id_b, mode columns alongside 62,920 rows. The column "mode" contains the labels for the interaction pairs which are reaction, binding, activasion, inhibition, catalysis, expression and ptmod (post-translational modifications). So, for predicting the labels in mode, we need to apply encoding for proper train of our model. We encoded the "mode" column using single-labeled encoding. But for the other columns, neihter the single-labeled encoding or the one-hot encoding is suitable, especially the sequence_a and sequence_b columns.

In MHS62k dataset, both sequence_a and sequence_b are in string format containing characters which represents amino acids for the sequence. For this, any one of single-labled encoding or one-hot encoding do not provide significace in sequence feature extraction. So, we proceeded with amino acid embeddings. We adopted amino acid embedding techniques of PIPR [15] model. Similar to the amino acid embedding techniques applied in the PIPR [15] model's study, we have produced a 5-dimensional vector based on co-occurrence similarity of amino acids in protein sequences. This vector is achieved by pretraining Skip-Gram model with our dataset's 3,143 proteins. This 5-dimensional vector was achieved by using Word2Vec model with context window of size 7 and negative samples of size 5. After that, we have constructed another 8-dimensional vector based on hydrophobicity and electrostaticity for the amino acids [15]. This 8-dimensional vector was collected from PIPR model's study. For amino acids, after combining both of these 5-dimensional vectors of co-occurrence similarity and 8-dimensional vectors of hydrophobicity and electrostaticity, we construct our amino acid embeddings for our protein sequences. This would be used to encode protein sequences for extracting meaningful sequence features. This process of protein sequence embedding approach has been observed to be more effective compared to usual one-hot encoding [15].

After that, we proceeded to construct a graph from **MHS62k** dataset. **MHS62k** dataset has 62,920 interaction rows. Each protein-protein interaction (PPI) rows can be considered and represented as an edge of a graph. Also, each proteins in an iteracting pair can be considered as a node of a graph. So, we proceeded to construct a graph from extracting edges and nodes for the graph from protein-protein interaction rows. We have considered the graph as bidirectional graph as in protein-protein interaction, both proteins are interacting with each other.

A large graph will not be effective enough in storing interaction type specific features. So, we adopted a subgraph approch where the subgraphs are the each labels in the interaction mode. We stored an edge index for a protein-protein interaction row both in the main graph and in their label-wise subgraph. This approach not only helps in extracting graph-wise features for protein-protein interactions, but also help in understanding and extracting subgraph-wise features which helps in enhancing interaction type prediction.

Our MHS62k protein-protein interaction dataset is a graph dataset. Also it is an imbalanced dataset with "reation" being the largest and "expression" being the smallest based on interaction type samples. So, conventional train test splitting is not enough for evaluation. So, we need split the dataset in other ways too. Breadth-First Search (BFS) and Depth-First Search (DFS) methods are well known as popular graph traversing methods [20]. As our dataset itself is a graph, so we can utilize both BFS and DFS methods for train test split [20]. Also, we will be applying random search alongside breadth-first search and depth-first search methods [20]. Additionally, we will be splitting our graph dataset into 0.7, 0.1 and 0.2 ratios for training, validation and testing to better train our model.

Chapter 4

Model Architecture

4.1 CBRVAE Model

The **CBRVAE** (Convolutional Bidirectional Recurrent Variational Autoencoder) model is a part of our model's pretrain model for our proposed **MidZPPI** (Midnight Zone Protein Protein Interction) model. CBRVAE model is used to encode protein sequence embeddings. The CBRVAE model has convolutional layers, recurrent layers, and a variational autoencoder. This model can be called an unsupervised learning model which tries to learn from a meaningful latent representation of protein sequence embeddings and tries to reconstruct that protein sequence embeddings from its latent space. This model helps MidZPPI with the protein sequence data.

This CBRVAE model has three parts in it: an Encoder, a Decoder and the Variational Autoencoder (VAE). For the CBRVAE model, the Encoder part of this model takes protein sequence embeddings as its input and then it provides the encoded protein sequence embeddings. After that, The Decoder part of this model takes the encoded protein sequence embeddings as its input and then it provides the reconstructed protein sequence embeddings. This reconstructed protein sequence embeddings tries to be similar to the original protein sequence embeddings. The variation aspect of this model outputs mean layer and log-variance layer which introduces the uncertainty in encoding.

4.1.1 Encoder

The Encoder takes protein sequence embeddings as its input feature. Then provides an encoded latent representation of protein sequences as its output feature. The

Encoder's architecture is shown in Figure 4.1. The first layer of the Encoder is an 1-dimensional convolutional layer with a kernal size of 3. This convolutional layer applies a filter of size 3 with stride 1 to convert input protein sequence embeddings into feature map. This layer extracts local features from the input protein sequence embeddings. The following layer is an 1-dimensional batch normalization layer. This layer stabilizes training by normalizing the activations of the convolution layer. This layer ensures that the data that will be passed to following layers is normalized for making training stable. The next layer is an 1-dimensional max pooling layer with kernal size of 3 and a stride of 3. This layer is used for downsampling the feature map and reducing complexity in computation. This layer enables the model to focus on most important features.

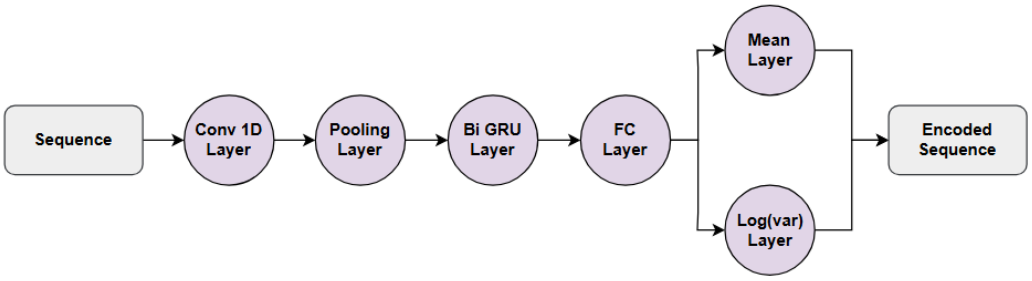

Figure 4.1: CBRVAE Encoder Architecture

The next layer is a bidirectional gated recurrent unit (BiGRU) layer. This layer helps capturing long-range dependencies of the protein sequence embeddings. As this is bidirectional, it can process the the protein sequence embeddings in both forward and backward directions. The following layer is an 1-dimensional adaptive average pooling layer. This layer pools the previous recurrent layer's output into a fixed length. This layer ensures that the pooled output is consistent in length which is necessary for the next layer. The following layer is the fully connected layer. This layer transforms the pooled output of previous layer into a vector with size 512. This layer reduces the dimension and provides the latent space.

After the previous fully caonnected layer, two other fully connected layers are connected to that which outputs mean, μ and log-variance, $log(\sigma^2)$ of the latent space. The mean layer produces the mean and the log-variance layer produces the logarithm of the variance of the latent variables in vectors with sizes of 1024. This helps the encoder's output z to be a latent space representation by applying the parameteriztion trick:

$$z = \mu + \sigma \cdot \epsilon$$

here the μ is the mean, σ is the standard deviation derived from log variance and ϵ is a random noise.

4.1.2 Decoder

The Decoder takes the latent variable z from the encoder and reconstructs a protein sequence embedding which tries to be similar as the original protein sequence embedding. The Decoder's architecture is shown in Figure 4.2. The first layer of the decoder is a fully connected layer. It takes z and maps to a vector of size 1024. This layer transforms the latent space to a higher dimension. Also, there is a *ReLU* activation function after the output of the first layer before the second layer. The second layer is a fully connected reconstruction layer. This layer reconstructs the output to match the shape of the original. After the second layer, the output is reshaped to match the shape of the original protein sequence embedding.

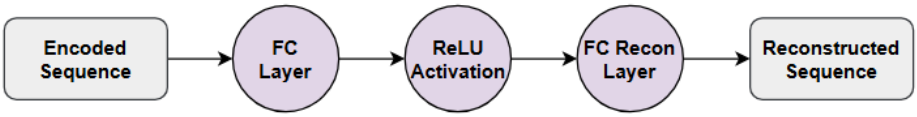

Figure 4.2: CBRVAE Decoder Architecture

4.1.3 Loss Function

The CBRVAE model connects both the Encoder and the Decoder. The Encoder produces latent variables which are mean, μ and log-variance, $log(\sigma^2)$ from the input protein sequence embeddings. The Decoder uses those latent variables to reconstruct the input protein sequence embeddings, which is close to the original one. The Encoder uses the reparameterization trick to sample from latent space. The reparameterization trick is as follows where the mean is μ, standart deviation is σ and a random noise is ϵ:

$$z = \mu + \sigma \cdot \epsilon$$

The CBRVAE is based on two loss function components. The first loss component

is a "Reconstruction Loss" which is $Loss_{recon}$. Reconstruction loss is a MSE loss, it measures the difference between the reconstructed protein sequence embeddings and the original protein sequence embeddings. The Reconstruction loss $Loss_{recon}$ is calculated in this way if x is the original input and x_{recon} is the reconstructed output:

$$Loss_{recon} = \frac{1}{N}\sum(x - x_{recon})^2$$

The second loss component is "KL Divergence Loss" which is $Loss_{KL}$. KL Divergence Loss is measured by the divergence between the learned distribution and a standard normal distribution. The KL Divergence Loss fomula as follows:

$$Loss_{KL} = -\frac{1}{2}(1 + log(\sigma^2) - \mu^2 - \sigma^2)$$

So, the loss function $Loss_{CBRVAE}$ formula for CBRVAE model is:

$$Loss_{CBRVAE} = Loss_{recon} + Loss_{KL}$$

4.2 MLPAE Model

The **MLPAE** (Multi-Layer Perceptron Autoencoder) model is a part of our model's pretrain model for our proposed **MidZPPI** (Midnight Zone Protein Protein Interction) model. MLPAE model is used for extracting protein sequence features from encoded protein sequence embeddings from the CBRVAE model. The MLPAE model is composed of fully connected layers alongside, Tanh and ReLU activations, and an autoencoder. This model is considered as an unsupervised learning model which tries to learn from a meaningful latent representation of encoded protein sequence embeddings and tries to reconstruct that encoded protein sequence embeddings from its latent space. This model helps MidZPPI with extracting protein sequence features.

This MLPAE model has three parts in it: an Encoder, a Decoder and the Autoencoder (AE). For the MLPAE model, the Encoder part of this model takes encoded protein sequence embeddings of CBRVAE model as its input and then it extracts the protein sequence features. After that, The Decoder part of this model takes the extracted protein sequence features as its input and then it provides the reconstructed

encoded protein sequence embeddings.

4.2.1 Encoder

The Encoder takes encoded protein sequence embeddings as its input feature. Then provides a latent representation of protein sequence features. The Encoder's architecture is shown in Figure 4.3. The first layer is a fully connected layer. This layer takes a vector size of 512 and transforms it to a smaller vector with size 256. This layer reduces the dimensionality of the encoded protein sequence embeddings. After the previous layer, *Tanh* activation function is applied on the previous fully connected layer. This activation function introduces non-linearity to the data which allows learning more complex relationships among the encoded protein sequence embeddings.

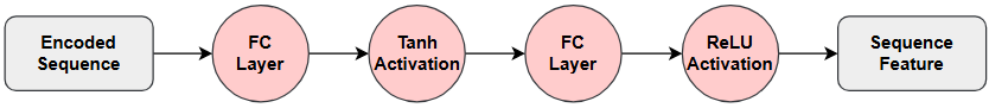

Figure 4.3: MLPAE Encoder Architecture

The next layer is another fully connected layer. This layer further reduces the dimensionality to a smaller vector of size 256. This layer captures most important features in the latent space and forms the final latent representation. Then a *ReLU* activation is applied on the output of the previous fully connected layer. This activation is used to introduce non-linearity to the data for better learning.

4.2.2 Decoder

The Decoder takes the latent variable z from the encoder and reconstructs encoded protein sequence embeddings which tries to be similar as the original encoded protein sequence embeddings. The Decoder's architecture is shown in Figure 4.4. The first layer is a fully connected reconstruction layer. This layer takes the latent representation and transforms to a higher dimension. This layer tries to reverse the compression performed by encoder. Then a *ReLU* activation function is applied after the previous fully connected reconstruction layer. This activation function introduces non-linearity to the data which allows to learn more complex relationships.

The final layer is a fully connected reconstruction layer. This layer reconstructs the vectors to the original encoded protein sequence embedding size of 512.

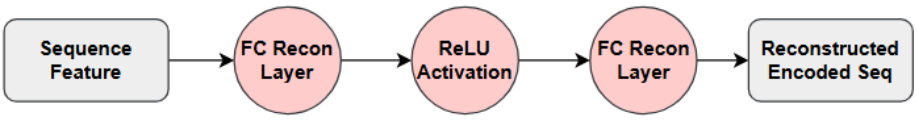

Figure 4.4: MLPAE Decoder Architecture

4.2.3 Loss Function

The MLPAE model connects both the Encoder and the Decoder. The Encoder produces a latent representation from the input encoded protein sequence embeddings. The Decoder uses the latent representation to reconstruct the input encoded protein sequence embeddings, which is close to the original one. The MLPAE uses MSE as it loss function. MSE loss measures the difference between the reconstructed encoded protein sequence embeddings and the original encoded protein sequence embeddings. The Reconstruction loss $Loss_{MLPAE}$ is calculated in this way if x is the original input and x_{recon} is the reconstructed output:

$$Loss_{MLPAE} = \frac{1}{N}\sum(x - x_{recon})^2$$

4.3 SCAGCN Model

The SCAGCN (Subgraph Context-Aware Graph Convolutional Network) model is a part of our proposed **MidZPPI** (Midnight Zone Protein Protein Interction) model. This model is designed for in-depth understanding of subgraph representations and provide node representation to capture both local and global context. This model has three parts in it: GNN module, Subgraph Encoder and Context Encoder. The SCAGCN's architecture is shown in Figure 4.5.

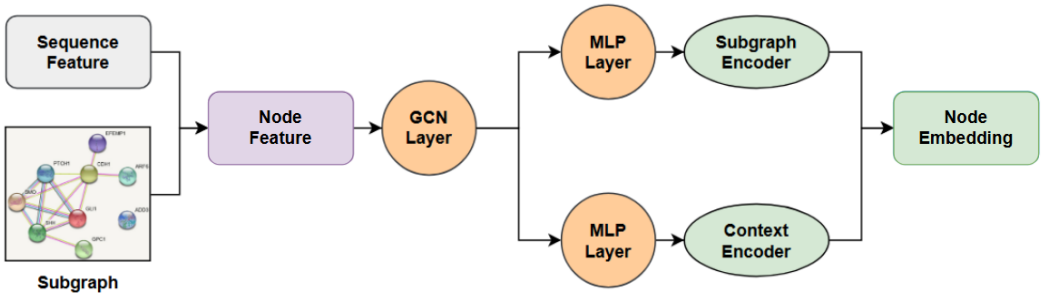

Figure 4.5: SCAGCN Model Architecture

4.3.1 GNN Module

The GNN module is the core module in the SCAGCN model. The first layer of this module is a GCN layer. This layer takes node features and edge indices as input and provides the updated node fetures as output. This layer updates each node of subgraphs by aggregating informations from its neighbours. The next layer is a 1-dimensional batch normalization layer. This layer normalizes node features and provides stable training. After this layer, a *ReLU* activation function is used to introduce non-linearity to the model. This helps the subgraphs to better learn complex node features. Also, a dropout layer is added after the activation. This is used to prevent overfitting. It applies regularization by randomly dropping fraction of nodes while training phase.

The last layer is either an Identity layer or an MLP layer based on if the input channels and output channels are equal or not. If equal, Identity layer is used which directly returns the previous layer's output. If not equal, a MLP layer is used to map node features to create a final representation. This MLP layer has 2 hidden layers with each having 1024 channels.

4.3.2 Subgraph Encoder

The Subgraph Encoder is a part of the encoders for SCAGCN model. This encoder uses multilayer perceptron (MLP) layer. The MLP layers have two hidden layers with the same direction as both input and output. They are used to process node features of the subgraphs. These layers focus on embedding node features from a subset of nodes of the subgraph. This encoder refines the feature presentation of the nodes in the subgraphs.

4.3.3 Context Encoder

The Context Encoder is a part of the encoders for SCAGCN model. This encoder uses multilayer perceptron (MLP) layer. The MLP layers have two hidden layers with the same direction as both input and output. They are used to process node contexts of the subgraphs. These layers focus on embedding node contexts related to its neighbour nodes of the subgraph. This encoder captures contextual relationships of the nodes in the subgraph.

4.3.4 Aggregation

The SCAGCN model follows aggregation method of the 1-hop egonet. The model first extracts the batches and maps nodes of the subgraphs. Then the node features are passed to the GNN module for updating node features for the subgraphs. Then some pivot nodes of the subgraphs are mapped accordingly for being selected as the initial occurances for its corresponding subgraph batch. After that, an embedding condition is imposed to the dual encoders in this model. This embedding condition enforces to choose the Context Encoder or not.

If Context Encoder is chosen based on the embedding condition, the encoded node features are passed through the Context Encoder for processing node contexts for the node embeddings of the subgraphs. Otherwise the previously achieved encoded node features remain. After the above embedding condition, the resulting node embeddings from the dual encoders are then aggregated through a summation for the model.

4.4 MidZPPI Model

The **MidZPPI** (Midnight Zone Protein Protein Interaction) prediction model is our proposed model for predicting multi-label protein-protein interactions in Midnight Zone of protein sequence alignments using a sequence based approach. Our model is divided into two blocks: subgraph block and classifier block. Firstly, MidZPPI pretrains protein sequences using both CBRVAE model and MLPAE model for extracting sequence features for the protein protein interaction prediction. Then, SCAGCN model of subgraph block contributes in extracting node features from the subgraphs. After combining both sequence features and subgraph features, the classifier block analyzes and produces prediction of the labels for the protein-protein interaction pairs. MidZPPI's architecture is shown in Figure 4.6.

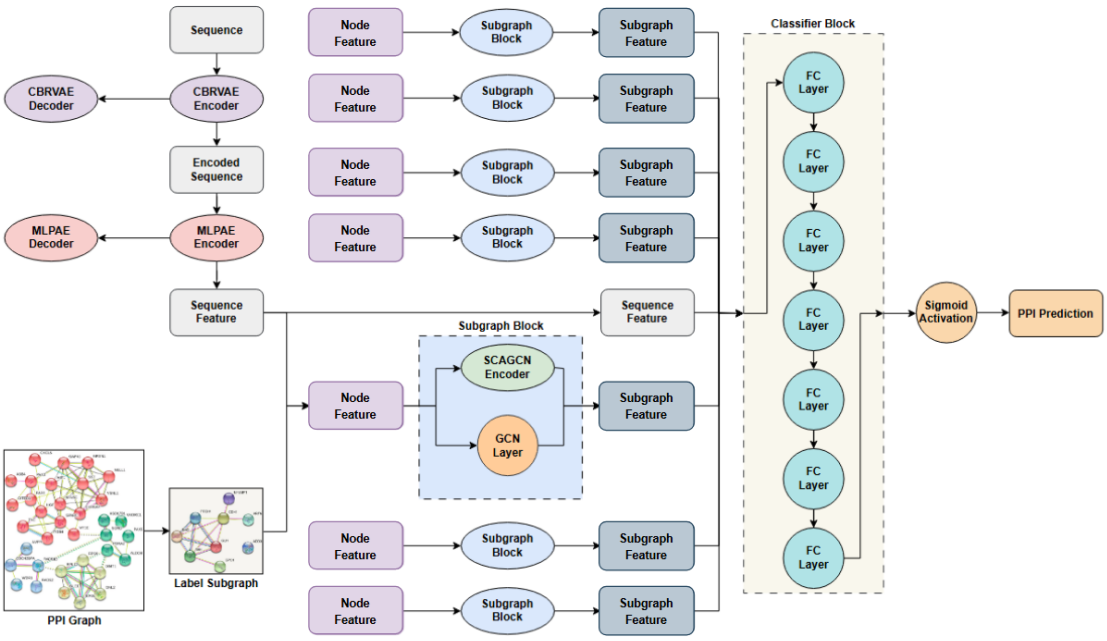

Figure 4.6: MidZPPI Model Architecture

4.4.1 Subgraph Block

The subgraph block is the first block of MidZPPI model. This block is used to process node features of subgraphs by introducing more complexity in the subgraphs and global graph. This block is separated into three layers: subgraph layer, global layer and final layer. the subgraph layer contains SCAGCN model in it. The global layer contains a GCN layer in its 1st layer. This layer performs GCN on the entire graph for producing a broader context for the nodes of the graph. This layer is applied to aggregate features over the entire graph. The final layer also contains multiple layers. The 1st layer is a 1-dimensional Batch normalization layer. This layer normalizes output features to stable training process.

4.4.2 Classifier Block

The classifier block firstly concatinates aggregated sequence features and subgraph features. After that, two interacting protein embeddings are extracted from it. We perform multiplication on both the protein embeddings to extract interaction feature. After that, the representation is sent to the classifier block for classification. This classifier block consists of seven fully connected layers. The number of input and output features for the fully connected layers in the classifier are based on the number of classes, which is seven, and hidden channels, which is 1024. These layers gradually reduce dimensionality of the features before classification. Lastly,

we perform sigmoid activation on the output provided by the classifier block to predict the PPI label.

4.4.3 Loss Function

For classifier block to predict labels, we adopt to using the ZLPR [23] loss function. This loss function is effective for multi-label categorical classification tasks. We use the CrossEntropy loss with Soft label to predict the label probabilities. The ZLPR [23] loss function formula:

$$L_{ZLPR} = \log\left(1 + \sum_{i \in \Omega_{pos}} e^{-s_i}\right) + \log\left(1 + \sum_{j \in \Omega_{neg}} e^{s_j}\right)$$

This loss function balances the influence of both positive and negative classes perfectly and is effective in capturing features for labels.

4.5 Other Models

4.5.1 GNN-PPI Model

GNN-PPI model utilizes Graph representation where all proteins are treated as nodes of the graph and all the protein-protein interactions (PPI) is considered as the edges for the graph [20]. Proteins are encoded in two ways, Protein-Independent Encoding (PIE) and Protein-Graph Encoding (PGE) [20]. Protein-Independent Encoding (PIE) extracts features from individual protein sequences using convolutional layers, recurrent layers and fully connected layers [20]. Protein-Graph Encoding (PGE) processes the PPI graph to capture relationships among neighbor nodes using GNN and aggregation of neighbor nodes by it enhances feature representation [20]. This model has utilized Random, BFS and DFS based graph dataset splitting.

4.5.2 DL-PPI Model

DL-PPI consists of GNN, Self Attention and FRN layers in it. In DL-PPI model, protein sequences are pre-processed using a pre-trained embedding model for en-

coding protein sequence information which is known as Inception. A graph neural network (GNN) is used in this model to learn relationships between proteins in the graph for broader context understanding [29]. A Self Attention Mechanism is used in this model to focus on important parts of protein features. This model uses Feature Relational Reasoning Network (FRN) to compute how the proteins are interacting based on relational reasoning [29]. The last fully connected layer of this model predicts the type of interacting pairs [29].

4.5.3 laruGL-PPI Model

LaruGL-PPI represents proteins and interactions as a graph where proteins are nodes and interactions are edges. This model introduces 3 types of graph encodings: Protein-Graph Encoding, PPI-Graph Encoding and Label-Graph Encoding. Protein-Graph Encoding represents each proteins by their amino acids. Graph Convolutional Networks (GCN) is used to extract useful features from this encoding [36]. PPI-Graph Encoding is built upon all the proteins being the nodes and the interactions being the edges of the graph which is constructed using all the PPI data. Graph Isomorphism Network (GIN) is used for this encoding to learn about the interactions [36]. Lastly, Label-Graph Encoding is constructed based on a subset of nodes and edges which belongs to a specific label. These label specific subsets represents subgraphs of the main PPI graph. Edge Based Subgraph Sampling is used to train graphs which reduces memory usage and improves efficiency. The GNNExplainer model of this model helps in prediction of the subgraph labels and identify most important residues for predicting interactions [36]. GNNExplainer model also provides interpretability of laruGL-PPI model.

Chapter 5

Result Analysis

5.1 Evaluation Metrics

In this section we will be discussing and analyzing the performance of our proposed **MidZPPI** model. We will be using our prepared **MHS62k** dataset as the benchmark dataset for performance evaluation of our model. We will be analyzing our model's performance based on Accuracy, Precision, Recall, F1-Score evaluation metrics. We will also be showing performance comparisons for other models which are GNN-PPI, DL-PPI and laruGL-PPI models with our proposed **MidZPPI** model based on the above mentioned evaluation metrics. We will also be showing visualizations of the comparisons based on the evaluation metrics for each train-test split methods which are Random Split, BFS Split, DFS Split. Alongside these, we will provide visualizations of model performance comparisons using Barplots and Confusion Matrices. Lastly, we will be providing comparative discussion on model architectures of all the models discussed for better understanding the robustness of the models.

5.1.1 Accuracy Metric

Accuracy is considered to be the baseline metric to understand how a model is performing. But accuracy can be misleading for imbalanced dataset. Accuracy will be calculated as the proportion of correctly predicted interctions to the total number of interaction predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Here, TP is true positives which are correctly predicted interactions, TN is true negatives which are correctly predicted non-interactions, FP is false positives which are incorrectly predicted interactions, FN is false negatives which are incorrectly predicted non-interactions.

5.1.2 Precision Metric

Precision is important for cases where false positives hold greater importance. Precision will be calculated as the proportion of predicted interactions that exists.

$$Precision = \frac{TP}{TP + FP}$$

5.1.3 Recall Metric

Recall is important in cases where a miss in detection could lead to serious consequences. Recall is calculated as the proportion of actual interactions which are correctly being predicted.

$$Recall = \frac{TP}{TP + FN}$$

5.1.4 F1-Score Metric

F1-score provides balance between both precision and recall by providing harmonic mean of both metrics. It is one of the most useful metric to evaluate imbalanced dataset.

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5.2 Performance Analysis

Our benchmark **MHS62k** dataset has is a multi-label imbalanced dataset. So, 'Micro Average F1-Score' is the most reliable metric to consider based on the state of our dataset. Accuracy is not chosen because its an imbalance dataset with multi-label targets. Accuracy will be dominated by the majority class labels, providing misleading information. Also, it will be unable to provide performances of minority classes. Also, Precision and Recall is not chosen also because of imbalanced nature of the dataset and multi-labeled targets which introduces biasness in these metrics. Either a high recall with low precision or a low recall with high precision is observed in this nature of datasets which leads to misinterpretation of target labels. For these reasons, we will only be evaluating better model based on the Micro Average F1-Score which is a fairly balanced and reliable evaluation metric for multi-label imbalanced datasets.

$$MicroAverageF1 - Score = 2 \times \frac{\sum TP}{\sum TP + \sum FP \sum FN}$$

5.2.1 Model Evaluation

We have used bfs, dfs and random partition techniques for evaluating GNN-PPI, DL-PPI, laruGL-PPI and MidZPPI models. This partitioning method is suitable for better evaluating graph based models. We have recorded the Micro Average F1 Scores of the models in Table 5.1 as we have chosen it as the best possible evaluation metric to provide accurate model performance indications.

Partition Methods	**GNN-PPI**	**DL-PPI**	**laruGL-PPI**	**MidZPPI**
BFS	0.6567	0.7215	0.8099	**0.8447**
DFS	0.6856	0.7832	0.8280	**0.8563**
Random	0.6966	0.7810	0.8543	**0.8976**

Table 5.1: Micro F1-Scores

From the table we can see that, our proposed **MidZPPI** model is leading in Micro Average F1 Score in BFS by 3.48%, in DFS by 2.83% and in Random by 4.33% from the second best model. The least compatible model for our study has been GNN-PPI model based on the above statistics.

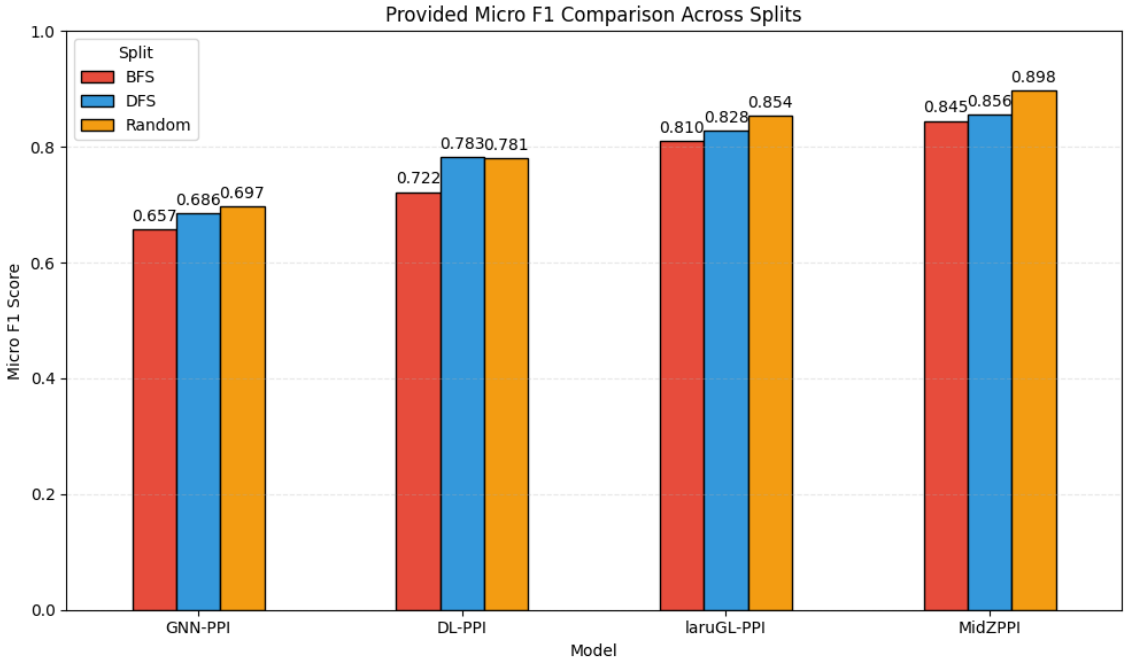

Figure 5.1: Micro F1-Score Comparisons

This barplot is shown in Figure 5.1 for better visualization for how the models are performing based on the Micro F1 Scores alongside which partition method have been the best/worst for them. It is clearly visible that all the models have struggled predicting PPI types in BFS partition and models tend to perform exceptionally good at Random split. We are also sharing partition wise comparison of all the metrics for all the models as follows. Micro Recalls and Micro Precision alongside the standard Micro F1 score would also be helpful in visualizing model's performance in each split in Figure 5.2, Figure 5.3 and Figure 5.4.

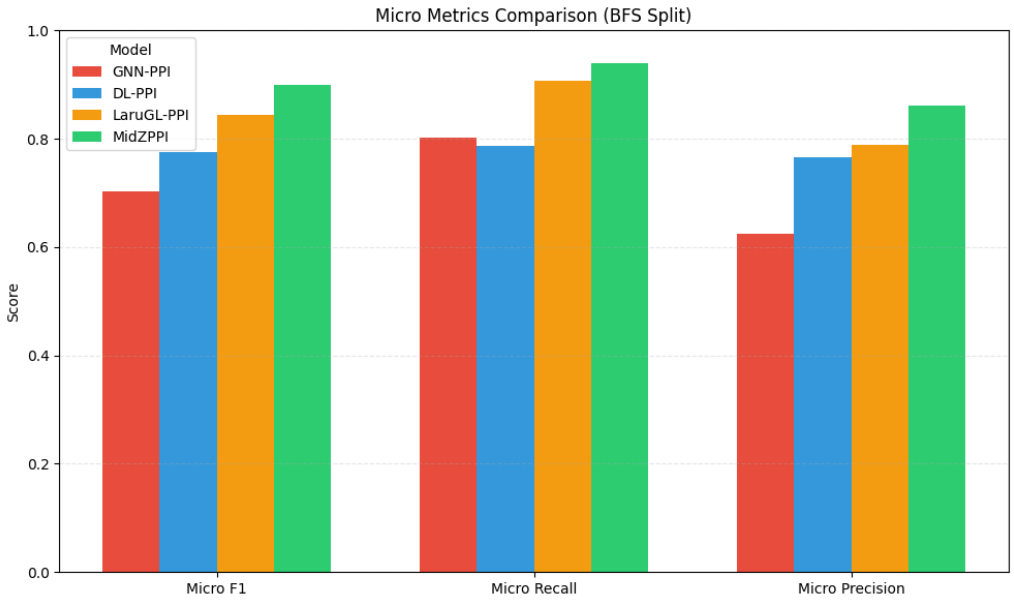

Figure 5.2: BFS Split Metrics Comparisons

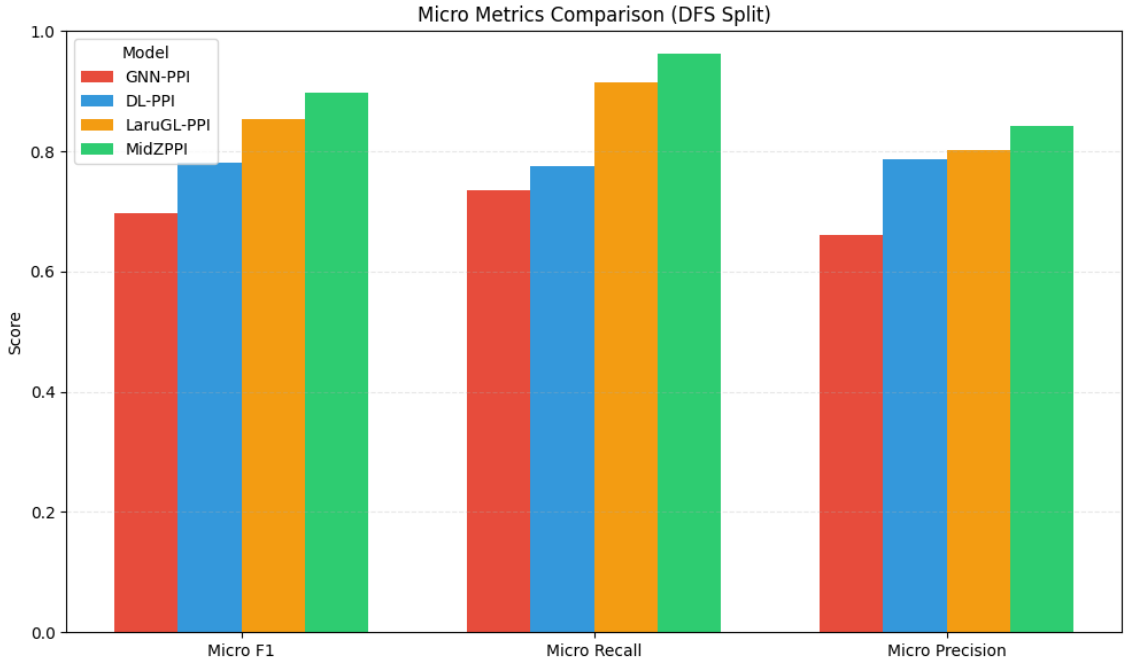

Figure 5.3: DFS Split Metrics Comparisons

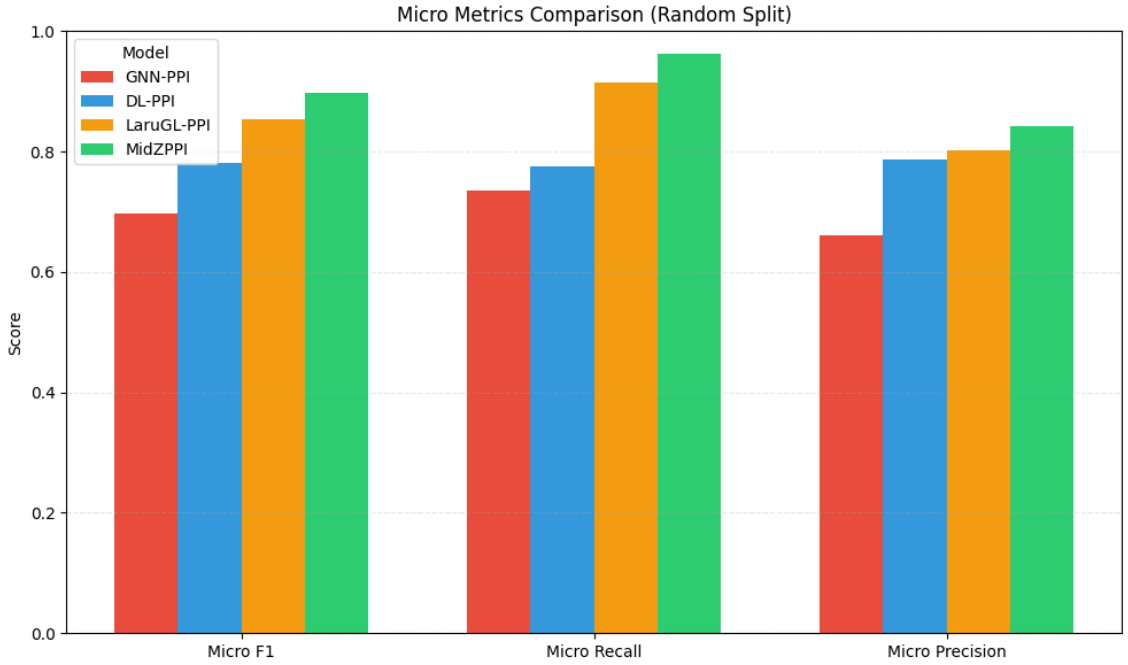

Figure 5.4: Random Split Metrics Comparisons

5.2.2 Confusion Matrices

As we are working with Multi-label dataset, there would be multiple confusion matrices. There are confusion matrices for each labels of 7 types. Having each

confusion matrices based on labels provides more clearer understanding on what are the label-wise performances for a model.

MidZPPI Model

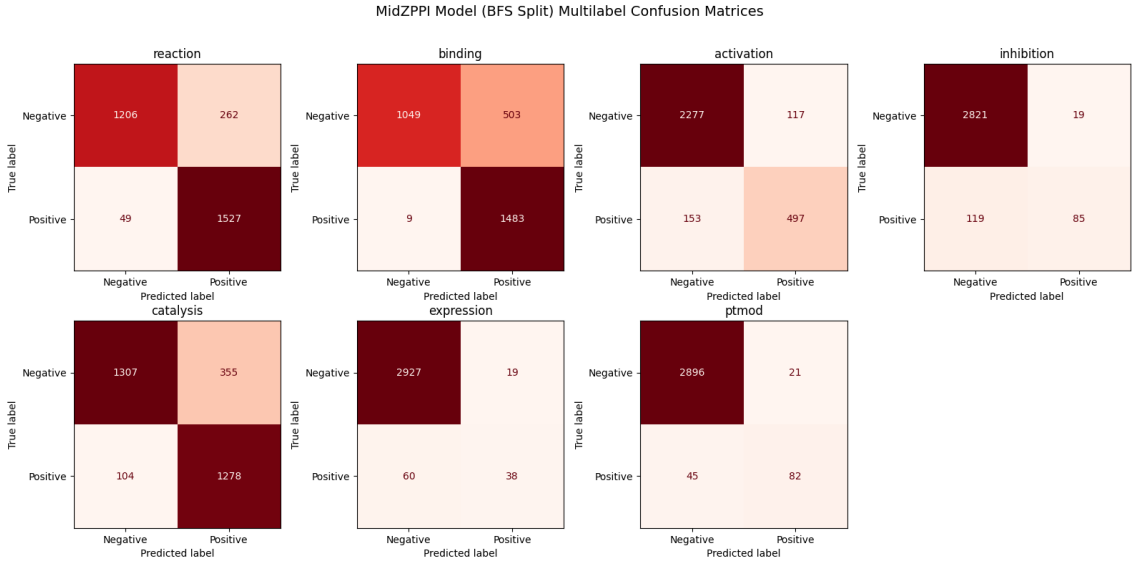

Figure 5.5: MidZPPI BFS Split Confusion Matrices

In Figure 5.5, these are the BFS Split confusion matrices of MidZPPI model. From the matrices we can see that, MidZPPI performs exceptionally well in predicting reaction labels in BFS split. But struggles majorly in predicting expression labels.

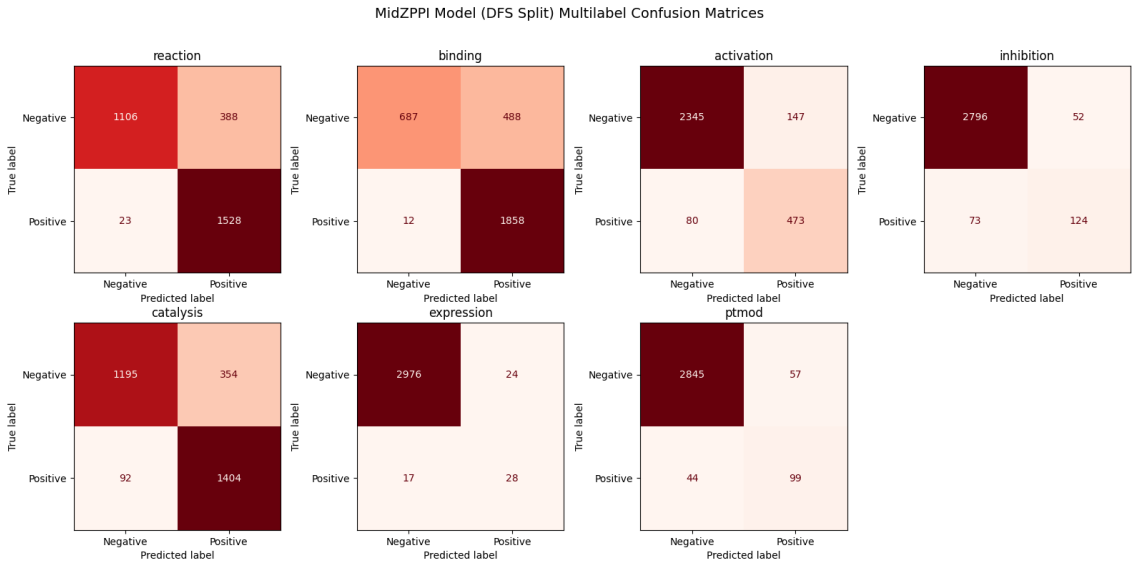

Figure 5.6: MidZPPI DFS Split Confusion Matrices

In Figure 5.6, these are the DFS Split confusion matrices of MidZPPI model. From the matrices we can see that, MidZPPI performs exceptionally well in predicting reaction labels in DFS split. But struggles majorly in predicting expression labels.

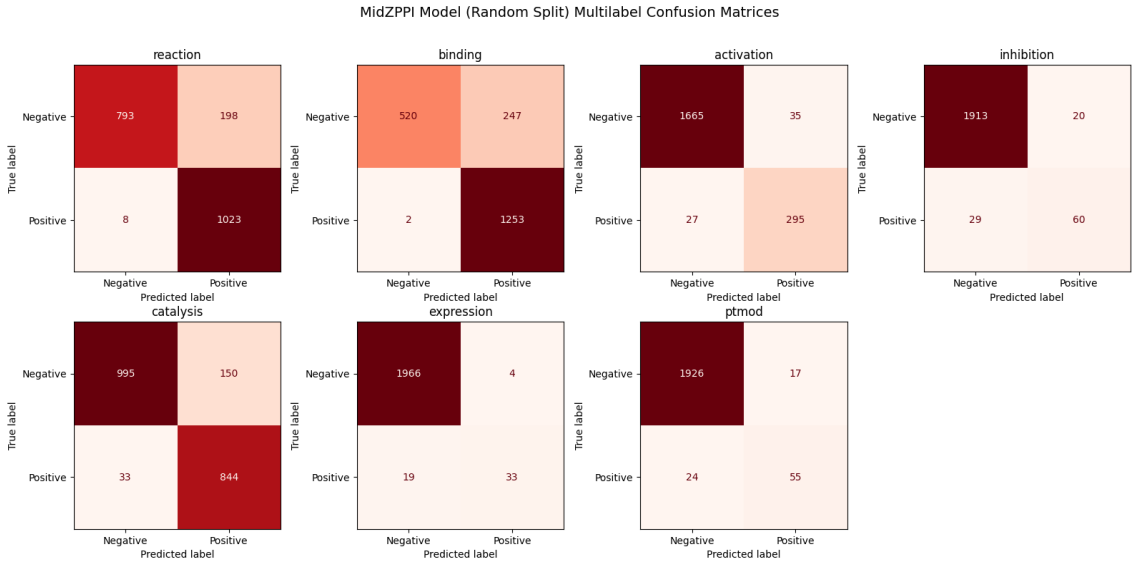

Figure 5.7: MidZPPI Random Split Confusion Matrices

In Figure 5.7, these are the Random Split confusion matrices of MidZPPI model. From the matrices we can see that, MidZPPI performs exceptionally well in predicting binding labels in Random split. But struggles majorly in predicting inhibition labels.

LaruGL-PPI Model

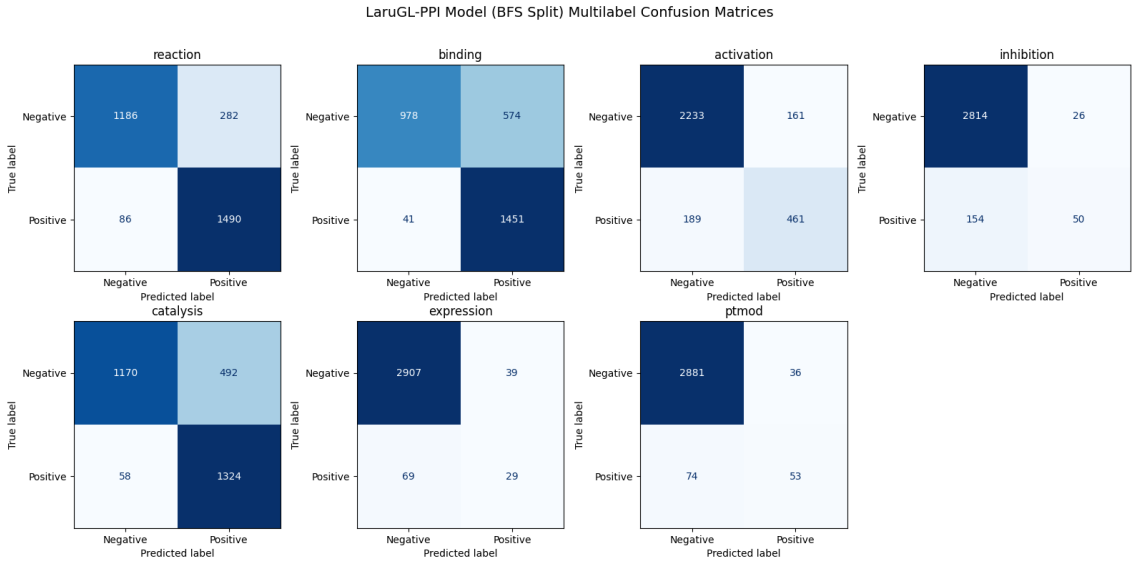

Figure 5.8: LaruGL-PPI BFS Split Confusion Matrices

In Figure 5.8, these are the BFS Split confusion matrices of LaruGL-PPI model. From the matrices we can see that, LaruGL-PPI performs exceptionally well in predicting catalysis labels in BFS split. But struggles majorly in predicting expression labels.

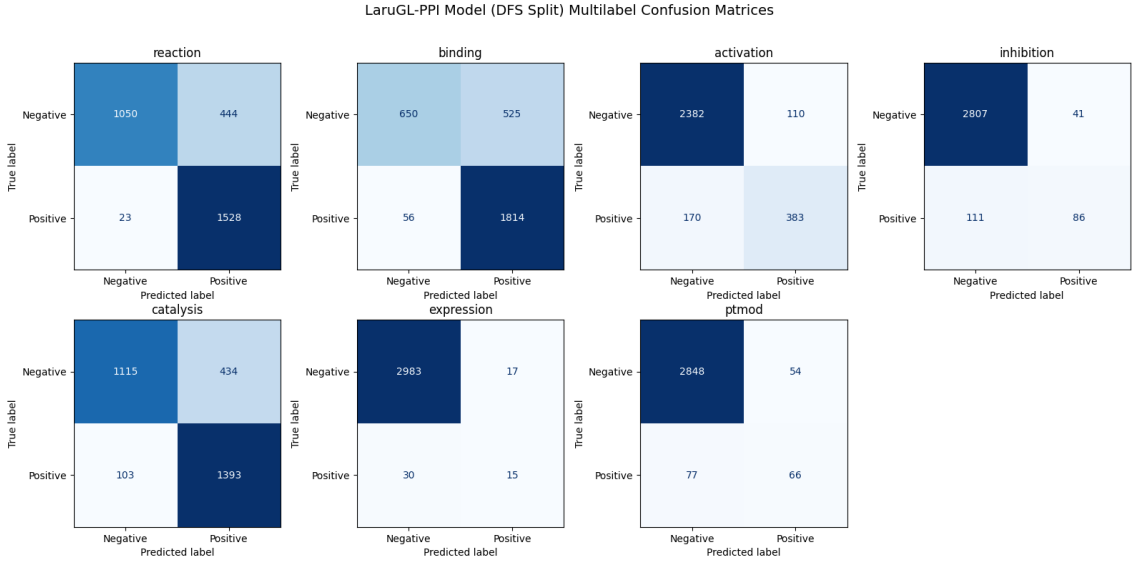

Figure 5.9: LaruGL-PPI DFS Split Confusion Matrices

In Figure 5.9, these are the DFS Split confusion matrices of LaruGL-PPI model. From the matrices we can see that, LaruGL-PPI performs exceptionally well in predicting reaction labels in DFS split. But struggles majorly in predicting expression labels.

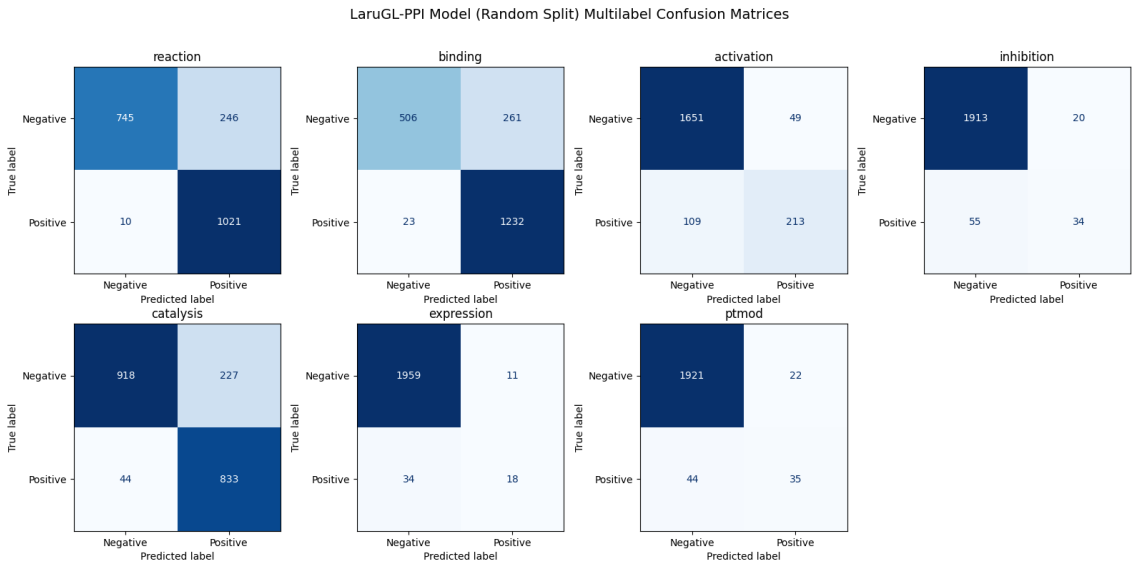

Figure 5.10: LaruGL-PPI Random Split Confusion Matrices

In Figure 5.10, these are the Random Split confusion matrices of LaruGL-PPI model.

From the matrices we can see that, LaruGL-PPI performs exceptionally well in predicting binding labels in Random split. But struggles majorly in predicting expression labels.

DL-PPI Model

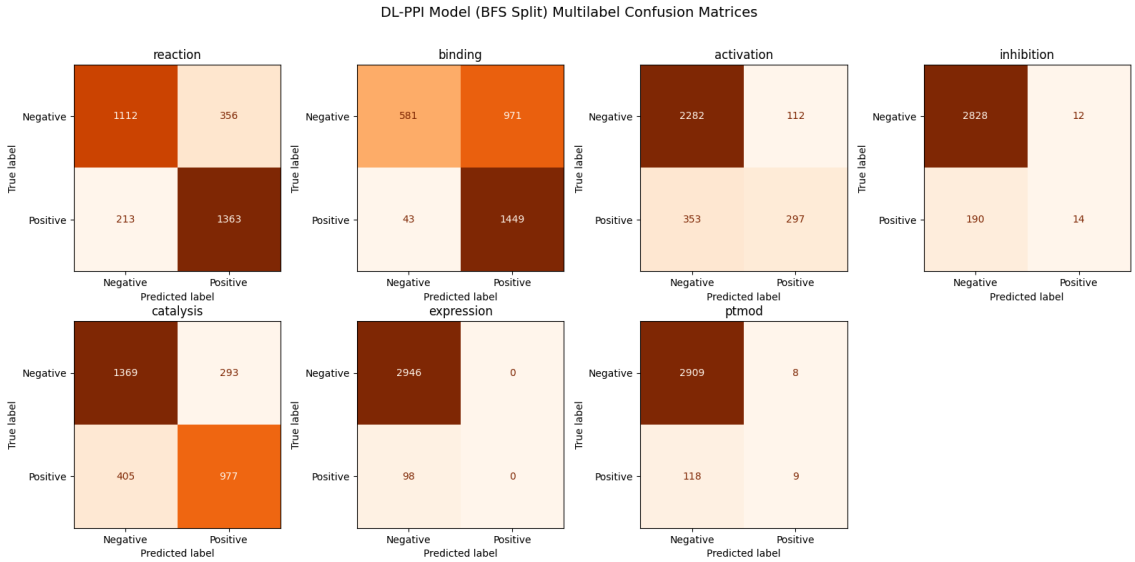

Figure 5.11: DL-PPI BFS Split Confusion Matrices

In Figure 5.11, these are the BFS Split confusion matrices of DL-PPI model. From the matrices we can see that, DL-PPI performs exceptionally well in predicting reaction labels in BFS split. But struggles majorly in predicting expression labels.

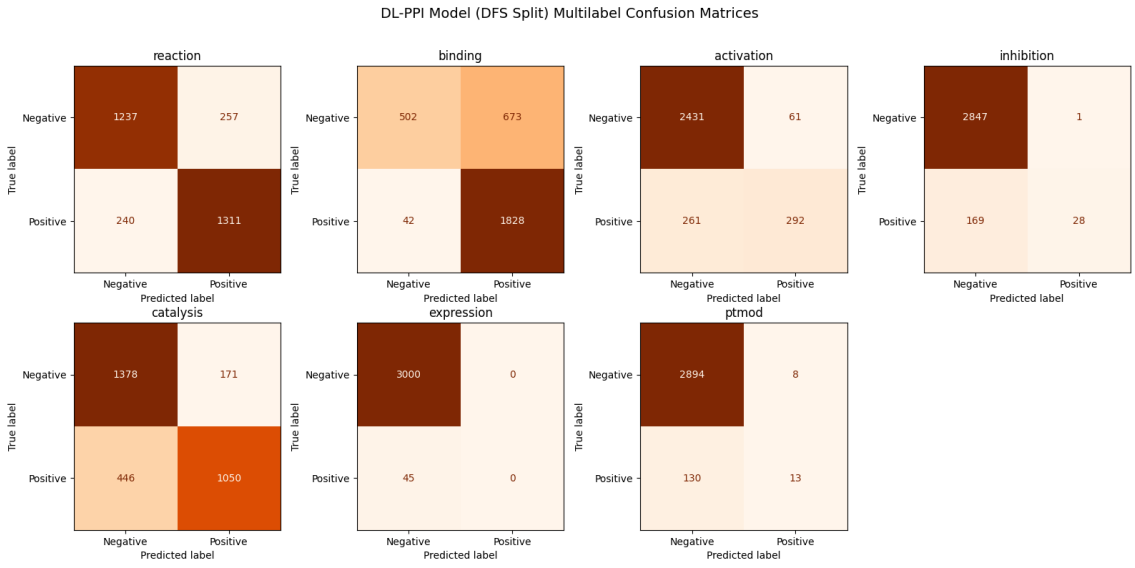

Figure 5.12: DL-PPI DFS Split Confusion Matrices

In Figure 5.12, these are the DFS Split confusion matrices of DL-PPI model. From the matrices we can see that, DL-PPI performs exceptionally well in predicting reaction labels in DFS split. But struggles majorly in predicting expression labels.

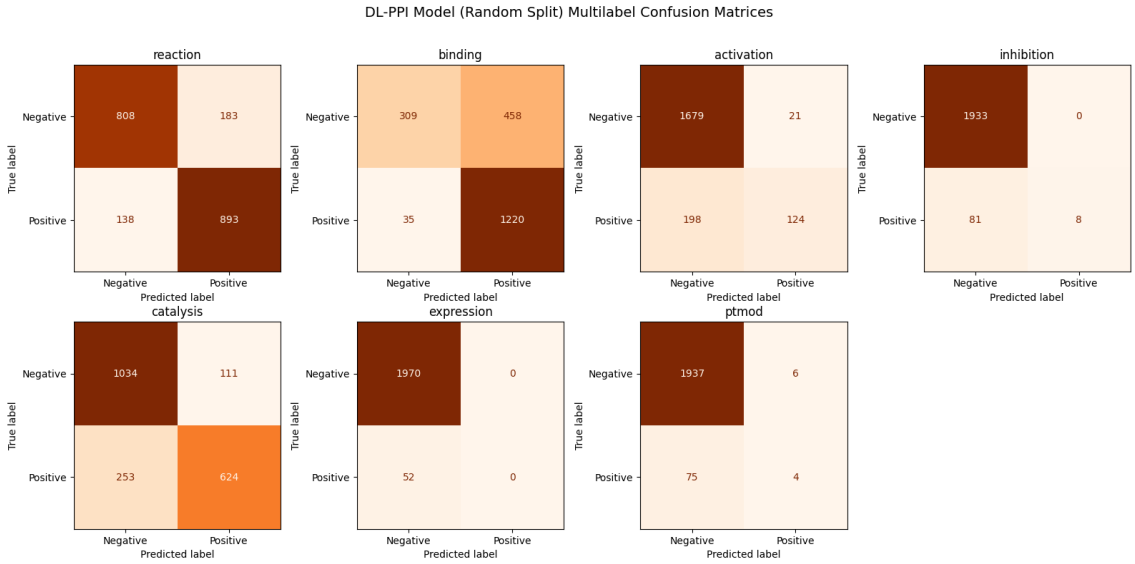

Figure 5.13: DL-PPI Random Split Confusion Matrices

In Figure 5.13, these are the Random Split confusion matrices of DL-PPI model. From the matrices we can see that, DL-PPI performs exceptionally well in predicting reaction labels in Random split. But struggles majorly in predicting expression labels.

GNN-PPI Model

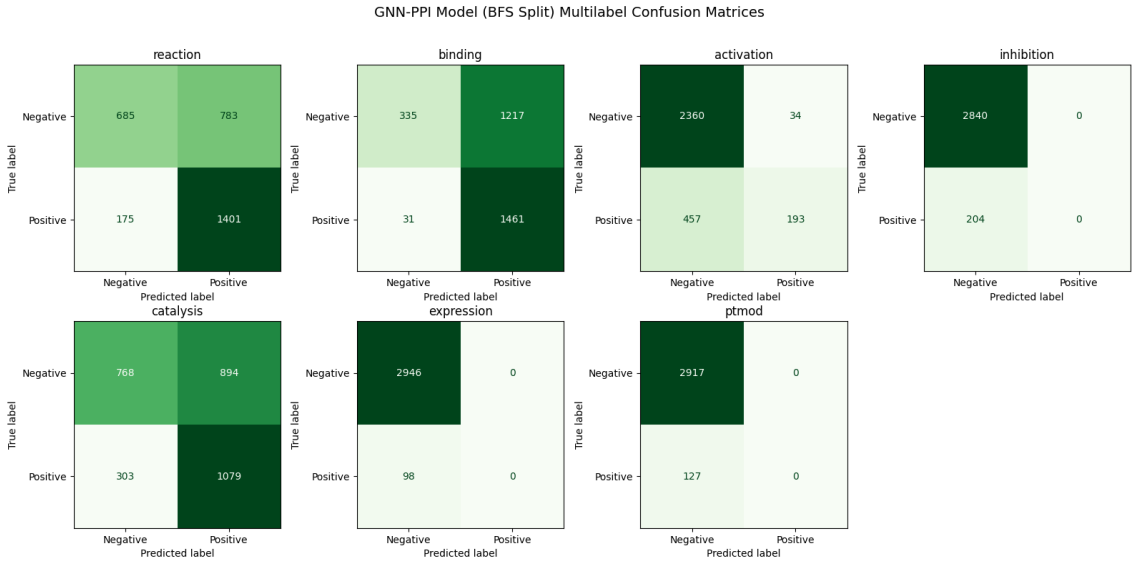

Figure 5.14: GNN-PPI BFS Split Confusion Matrices

In Figure 5.14, these are the BFS Split confusion matrices of GNN-PPI model. From the matrices we can see that, GNN-PPI performs exceptionally well in predicting reaction labels in BFS split. But struggles majorly in predicting expression, inhibition and ptmod labels.

Figure 5.15: GNN-PPI DFS Split Confusion Matrices

In Figure 5.15, these are the DFS Split confusion matrices of GNN-PPI model. From the matrices we can see that, GNN-PPI performs exceptionally well in predicting binding labels in DFS split. But struggles majorly in predicting expression, inhibition and ptmod labels.

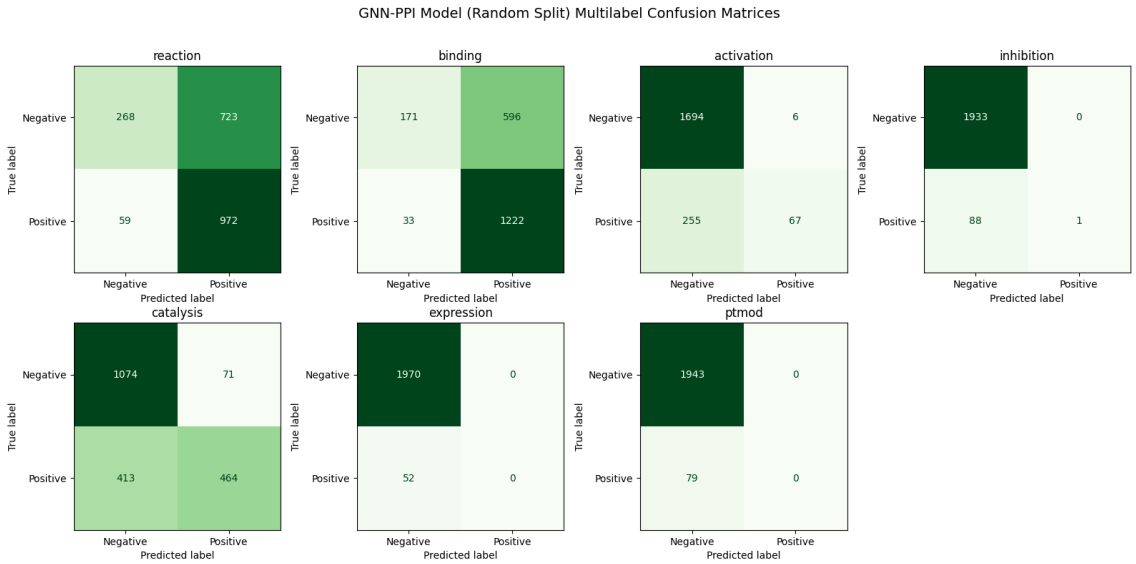

Figure 5.16: GNN-PPI Random Split Confusion Matrices

In Figure 5.16, these are the Random Split confusion matrices of GNN-PPI model.

From the matrices we can see that, GNN-PPI performs exceptionally well in predicting binding labels in Random split. But struggles majorly in predicting expression, inhibition and ptmod labels.

5.3 Comparative Analysis

Now we will be looking into all the models and see which model provides better features compared to others. We will be comparing model architectures, benefits it provides compared to other models to see in which standpoint our model stays and how better/worse our model is compared to others.

Features	**GNN-PPI**	**DL-PPI**	**laruGL-PPI**	**MidZPPI**
Focuses on Midnight Zone Proteins	✗	✗	✗	✓
Follows Sequence-based Approaches	✓	✓	✓	✓
Follows Embedding Sequences	✓	✓	✗	✓
Follows Subgraph-based Approaches	✗	✗	✓	✓
Focuses on Interpretability	✗	✗	✓	✗

Table 5.2: Comparative Study on Models

So, we can see from the Table 5.2 that our paper has an unique analysis on Midnight Zone of protein sequence alignments which other studies lack. Our study also uses a subgraph based approach alongside laruGL-PPI model which helped in achieving greater results regarding studies in Midnight Zone of protein sequence alignments. Also, laruGL-PPI focuses on interpretability rather than predicting on probability. This approach also helps in interpreate why certain interaction has that specific features. It could be a future work for our model as well.

Chapter 6

Conclusion

We have presented here MidZPPI, a sequence-based algorithm, which specifically aims at predicting multi-label protein-protein interactions within the Midnight Zone of protein sequences, where standard sequence alignment methods frequently have low similarity. We have been successful in our method of describing the interactions between proteins based on both in-depth interactions as co-occurring patterns and physicochemical parameters of amino acids as embeddings within a graphical framework to capture the interactions as nodes and edges in the graphical representation. Further improvement of the ability to extract meaningful features that are specific to the type of interaction was made possible by subgraph construction and traversal approaches like BFS and DFS.

These findings indicate that predicting multi-label PPIs based on the focus on Midnight Zone proteins is a better way to identify rare or previously unknown interactions and better predictive capabilities. In general, MidZPPI has been shown to offer a powerful protein interaction map in highly divergent protein sequences and has a lot of potential applications in protein function prediction, drug discovery, and systems biology.

6.1 Further Research Area

Despite the promising potential of sequence-based PPI prediction models, particularly when studying proteins in extreme environments such as the Midnight Zone, such issues as the difficulty of available data and the complexity of protein sequences remain. Nevertheless, progress in the field of deep learning and explainable AI is making them more reliable and understandable. Along the way, it is worth considering the biological accuracy and the interpretability of these models in order to

be able to apply them to the problem of protein functions and interactions in the challenging environments [34].

To overcome the challenges of midnight zone proteins, the future studies should be oriented at further extending the data on Midnight Zone proteins and using more sophisticated models, which will be able to reflect the peculiarities of these proteins. It might also be improved by hybrid methods that use sequence data with structural predictions using tools such as AlphaFold to enhance the accuracy of the PPI predictions of these extreme organisms [28].

Future research might also use 3D structural data of proteins in conjunction with sequence embeddings to enhance the accuracy of prediction and particularly those interactions that require protein folding or conformational changes. It may be useful to predict multi-label PPIs in other species with the help of the further extension of MidZPPI to uncover conserved patterns of interactions and evolution of Midnight Zone proteins.Biological relevance of the predicted interactions would be improved by the addition of contextual biological data including tissue specificity, cellular localization or post-translational modifications.The experimentally verifiable interactions between predicted interactions in the Midnight Zone should be verified, and would be useful to refine the computational model and scale up reliable PPI databases of low-similarity proteins.

Bibliography

[1] S. A. Benner, M. A. Cohen, and G. H. Gonnet, "Empirical and structural models for insertions and deletions in the divergent evolution of proteins," *Journal of molecular biology*, vol. 229, no. 4, pp. 1065–1082, 1993.

[2] S. F. Altschul et al., "Gapped blast and psi-blast: A new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[3] B. Rost, "Twilight zone of protein sequence alignments," *Protein Engineering*, vol. 12, no. 2, pp. 85–94, Feb. 1999, ISSN: 0269-2139. DOI: 10.1093/protein/12.2.85 eprint: https://academic.oup.com/peds/article-pdf/12/2/85/18542824/120085.pdf. [Online]. Available: https://doi.org/10.1093/protein/12.2.85

[4] M. Ashburner et al., "Gene ontology: Tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[5] J. Söding, A. Biegert, and A. N. Lupas, "The hhpred interactive server for protein homology detection and structure prediction," *Nucleic acids research*, vol. 33, no. suppl_2, W244–W248, 2005.

[6] G. J. Barton, "Quantification of the variation in percentage identity for protein sequence alignments," *BMC bioinformatics*, vol. 7, no. 1, p. 415, 2006.

[7] J. De Las Rivas and C. Fontanillo, "Protein–protein interaction networks: Unraveling the wiring of molecular machines within the cell," *Briefings in functional genomics*, vol. 11, no. 6, pp. 489–496, 2012.

[8] I. H. Moal and J. Fernández-Recio, "Skempi: A structural kinetic and energetic database of mutant protein interactions and its use in empirical models," *Bioinformatics*, vol. 28, no. 20, pp. 2600–2607, 2012.

[9] Q. C. Zhang et al., "Structure-based prediction of protein–protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[10] F. R. Sutandy, J. Qian, C.-S. Chen, and H. Zhu, "Overview of protein microarrays," *Current protocols in protein science*, vol. 72, no. 1, pp. 27–1, 2013.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[12] Z.-S. Wei, K. Han, J.-Y. Yang, H.-B. Shen, and D.-J. Yu, "Protein–protein interaction sites prediction by ensembling svm and sample-weighted random forests," *Neurocomputing*, vol. 193, pp. 201–212, 2016.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[14] T. C. Northey, A. Barešić, and A. C. Martin, "Intpred: A structure-based predictor of protein–protein interaction sites," *Bioinformatics*, vol. 34, no. 2, pp. 223–229, 2018.

[15] M. Chen et al., "Multifaceted protein–protein interaction prediction based on siamese residual rcnn," *Bioinformatics*, vol. 35, no. 14, pp. i305–i314, Jul. 2019, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz328 eprint: https://academic.oup.com/bioinformatics/article-pdf/35/14/i305/50721406/bioinformatics_35_14_i305_s4.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz328

[16] A. K. Thakur and L. Movileanu, "Real-time measurement of protein–protein interactions at single-molecule resolution using a biological nanopore," *Nature biotechnology*, vol. 37, no. 1, pp. 96–101, 2019.

[17] H. Lu et al., "Recent advances in the development of protein–protein interactions modulators: Mechanisms and clinical trials," *Signal transduction and targeted therapy*, vol. 5, no. 1, p. 213, 2020.

[18] E. J. Stollar and D. P. Smith, "Uncovering protein structure," *Essays in biochemistry*, vol. 64, no. 4, pp. 649–680, 2020.

[19] S. Das and S. Chakrabarti, "Classification and prediction of protein–protein interaction interface using machine learning algorithm," *Scientific reports*, vol. 11, no. 1, p. 1761, 2021.

[20] G. Lv, Z. Hu, Y. Bi, and S. Zhang, "Learning unknown from correlations: Graph neural network for inter-novel-protein interaction prediction," *arXiv preprint arXiv:2105.06709*, 2021.

[21] M. Heinzinger, M. Littmann, I. Sillitoe, N. Bordin, C. Orengo, and B. Rost, "Contrastive learning on protein embeddings enlightens midnight zone," *NAR Genomics and Bioinformatics*, vol. 4, no. 2, lqac043, Jun. 2022, ISSN: 2631-9268. DOI: 10.1093/nargab/lqac043 eprint: https://academic.oup.com/nargab/article-pdf/4/2/lqac043/44245898/lqac043.pdf. [Online]. Available: https://doi.org/10.1093/nargab/lqac043

[22] X. Li, P. Han, G. Wang, W. Chen, S. Wang, and T. Song, "Sdnn-ppi: Self-attention with deep neural network effect on protein-protein interaction prediction," *BMC genomics*, vol. 23, no. 1, p. 474, 2022.

[23] J. Su, M. Zhu, A. Murtadha, S. Pan, B. Wen, and Y. Liu, "Zlpr: A novel loss for multi-label classification," *arXiv preprint arXiv:2208.02955*, 2022.

[24] D. Szklarczyk et al., "The string database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic Acids Research*, vol. 51, no. D1, pp. D638–D646, Nov. 2022, ISSN: 0305-1048. DOI: 10.1093/nar/gkac1000 eprint: https://academic.oup.com/nar/article-pdf/51/D1/D638/48440966/gkac1000.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkac1000

[25] N. Zhang et al., "Ontoprotein: Protein pretraining with gene ontology embedding," *arXiv preprint arXiv:2201.11147*, 2022.

[26] N. Zhao, M. Zhuo, K. Tian, and X. Gong, "Protein–protein interaction and non-interaction predictions using gene sequence natural vector," *Communications Biology*, vol. 5, no. 1, p. 652, 2022.

[27] R. Dhanuka, J. P. Singh, and A. Tripathi, "A comprehensive survey of deep learning techniques in protein function prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 3, pp. 2291–2301, 2023.

[28] N. Kewalramani, A. Emili, and M. Crovella, "State-of-the-art computational methods to predict protein–protein interactions with high accuracy and coverage," *Proteomics*, vol. 23, no. 21-22, p. 2 200 292, 2023.

[29] J. Wu, B. Liu, J. Zhang, Z. Wang, and J. Li, "Dl-ppi: A method on prediction of sequenced protein–protein interaction based on deep learning," *BMC bioinformatics*, vol. 24, no. 1, p. 473, 2023.

[30] T. U. Consortium, "Uniprot: The universal protein knowledgebase in 2025," *Nucleic Acids Research*, vol. 53, no. D1, pp. D609–D617, Nov. 2024, ISSN: 1362-4962. DOI: 10.1093/nar/gkae1010 eprint: https://academic.oup.com/nar/article-pdf/53/D1/D609/60719276/gkae1010.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkae1010

[31] D. Kalifa, U. Singer, and K. Radinsky, "Goproteingnn: Leveraging protein knowledge graphs for protein representation learning," *arXiv preprint arXiv:2408.00057*, 2024.

[32] M. Xu et al., "Graph neural networks for protein-protein interactions–a short survey," *arXiv preprint arXiv:2404.10450*, 2024.

[33] Y.-H. Zhu, Z. Liu, Y. Liu, Z. Ji, and D.-J. Yu, "Uldna: Integrating unsupervised multi-source language models with lstm-attention network for high-accuracy protein–dna binding site prediction," *Briefings in Bioinformatics*, vol. 25, no. 2, bbae040, 2024.

[34] M. N. Asim, T. Asif, F. Hassan, and A. Dengel, "Protein sequence analysis landscape: A systematic review of task types, databases, datasets, word embeddings methods, and language models," *Database*, vol. 2025, baaf027, 2025.

[35] X. Zheng et al., "Pring: Rethinking protein-protein interaction prediction from pairs to graphs," *arXiv preprint arXiv:2507.05101*, 2025.

[36] Y. Zhou et al., "Effectiveness and efficiency: Label-aware hierarchical subgraph learning for protein–protein interaction," *Journal of Molecular Biology*, vol. 437, no. 6, p. 168 737, 2025.