PRIFYSGOL CYMRU
Y Drindod Dewi Sant
UNIVERSITY OF WALES
Trinity Saint David
SWANSEA - ABERTAWE

School of
Applied Computing

PROTEIN FUNCTION PREDICTION USING GRAPH CONVOLUTIONAL NETWORK

Yuanhao Chen

2304723

Supervisor: Dr. Seena Joseph

Project submitted as part of the requirements for the award of MSc: Software Engineer and Artificial Intelligence

September 2024

Declaration of Originality

I, ...(Yuanhao Chen)…… declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed :Yuanhao Chen .....

Date :08/28/2024 .....

FORM OF CONSENT

I Yuanhao Chen hereby consent that my Project, submitted in candidature for the Master Degree , if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed :Yuanhao Chen .....

Date : 08/28/2024.....

Copyright Acknowledgement

ABSTRACT/SYNPOSIS

This project advances protein function prediction by integrating protein language models (PLMs) and graph convolutional networks (GCNs), addressing the limitations of traditional methods that rely heavily on sequence similarity. The proposed model leverages diverse protein features, including sequences, protein-protein interaction (PPI) networks, and InterPro domains, to create a robust computational framework.

Utilizing the Evolutionary Scale Modeling (ESM-1b) PLM, high-dimensional feature embeddings are generated from protein sequences. These are integrated with PPI network data and InterPro domains through a two-layer GCN, enabling the model to capture complex interdependencies. The model's performance was evaluated using metrics such as Fmax and Area Under the Precision-Recall Curve (AUPR) across different Gene Ontology (GO) categories: Molecular Function (MFO), Biological Process (BPO), and Cellular Component (CCO).

The findings demonstrate that the model outperforms the most advanced techniques currently in use for BPO and CCO forecasts. However, MFO predictions require improvement, suggesting that future efforts should concentrate on more accurately identifying sequence-specific motifs.

The study problem and objectives are presented first in the report, which is structured to give a thorough overview. A review of the literature, the research technique, and a detailed analysis of the experimental data are then included. The study concludes with reflections, highlighting areas for future research and the broader implications for biomedical and biotechnological applications.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF CHART

ACKNOWLEDGEMENTS

I want to sincerely thank Dr. Seena Joseph from the University of Wales Trinity Saint David for all of her help and support over the last three months as I worked on my dissertation. Dr. Joseph has not only provided me with essential writing techniques but has also imparted a level of dedication and thoroughness that has greatly enriched my academic journey. I first encountered Dr. Joseph during my second semester courses, where her commitment and responsibility as a teacher were evident. It has been an honor to have her as my dissertation supervisor, and without her insightful feedback and unwavering support, I would not have been able to complete this dissertation so efficiently.

I also want to express my sincere gratitude to Dr. Huang from Wuhan University of Technology, whose knowledge and suggestions have greatly influenced the course of my work. The success of my project has been greatly attributed to Dr. Huang's direction in helping me develop creative ways and providing insightful analysis of my research emphasis.

Lastly, I wish to thank my fellow classmates who traveled to the UK with me. This has been our first experience studying and living abroad, and without their friendship and support, I cannot imagine how I would have navigated this challenging yet rewarding year. Their encouragement and companionship have made this journey possible and truly memorable.

CHAPTER 1 INTRODUCTION

Proteins are indispensable to cellular functions, playing critical roles in maintaining the acid-base balance, distributing water, transmitting genetic information, and transporting various vital substances within human organisms [1]. To systematically catalog diverse function of proteins, Gene Ontology (GO) provides structured vocabulary to classify protein functions into molecular function, biological process, and cellular component [2].

Protein function prediction helps in understanding life processes and disease mechanisms, thereby aiding in disease diagnosis and drug development. It also plays a key role in biotechnology applications, such as designing specific enzymes to enhance industrial and agricultural efficiency[3]. In recent years, the quantity of protein sequences stored in public databases has surged, enhancing our comprehension of protein diversity. And deep learning has shown promise in unearthing intricate patterns in high-dimensional data, making it ideal for tasks such as protein function classification[4]. Initial deep learning approaches for this purpose primarily utilized protein sequence data and positional information within Protein-Protein Interaction (PPI) networks[5]. Examples include DeepGO, which uses convolutional neural networks to extract sequence features, and its successor, DeepGOPlus [6], which enhances efficiency by combining a simplified neural network approach with nearest-neighbor algorithms. Another method, DeepAdd[7], interprets protein sequences through natural language processing techniques to generate feature representations. These protein function prediction methods still suffer from issues of insufficient accuracy and the inability to integrate all protein features comprehensively.

The evolution of Graph Neural Networks (GNNs) has introduced new methodologies for protein function prediction by effectively representing proteins'3D structures and PPI networks as graph-structured data[8]. For instance, DeepFri[9]

employs a self-supervised language model to derive residue features, which are then propagated using graph convolutional networks. Similarly, DeepGraphGO applies semi-supervised learning to integrate PPI and InterPro features for protein function prediction[10].

Protein Language Models (PLM) are computational tools that use machine learning to analyze and predict protein sequences, structures, and functions based on patterns learned from large datasets of protein data. ESM (Evolutionary Scale Modeling) was developed by Facebook AI Research, it uses Transformer architectures to generate representations of protein sequences[11].UniRep uses a recurrent neural network (RNN) approach to condense protein sequences into fixed-length vectors[12]. ProtTrans adapts Bidirectional Encoder Representations from Transformer (BERT) and T5 models from NLP to the protein sequencing field[13]. ProteinBERT applies the Masked Language Model (MLM) approach to predict amino acids in sequences[14]. Protein language models have demonstrated high accuracy in generating protein embeddings.

1.1 RESEARCH PROBLEM STATEMENT

The traditional methods for annotating protein functions remain costly and slow[15], leading to a significant annotation backlog as new proteins are discovered faster than they can be functionally characterized[16]. Protein language models like ESM-1b have demonstrated high accuracy in generating protein embeddings[11]. However, their potential has not been fully realized in combination with graph-based methods such as DeepGoPlus,DeepGo and DeepGraphGo [6]. Hence there is a critical need for reliable computational prediction models that can efficiently and accurately predict protein functions by integrating heterogeneous protein features.

1.2 RESEARCH AIM

The primary aim of this study is to develop a novel protein function prediction model by integrating data from protein sequences, protein-protein interaction (PPI) networks, and InterPro domains using PLM and Graph Convolutional Network(GCN), to generate accurate predictions of protein functions.

1.3 RESEARCH OBJECTIVES

The main aim of this study is achieved through the following objectives:

Obj1 - To generate embeddings from protein sequences using the pre-trained protein language model (ESM-1b) for Feature Extraction .

Obj2 - To integrate the embeddings from protein sequences and InterPro domains with adaptive feature weights into the PPI graph, and use GCNs to generate protein features.

Obj3 - To develop a classification model that combines the features weights and protein features vector generated by PLM,PPI and GCNs.

Obj4 - To evaluate and compare the performance of the developed model against existing state-of-the-art methods using well-known evaluation metrics.

1.4 SIGNIFICANCE/CONTRIBUTION OF THIS RESEARCH

This research marks a significant advancement in protein function prediction by pioneering the integration of protein language models (PLMs) with graph convolutional networks (GCNs). By harnessing a rich spectrum of protein features—

including sequence data, protein-protein interaction (PPI) networks, and InterPro domain information—this study develops a comprehensive and highly robust computational model. This innovative approach not only improves the accuracy and efficiency of protein function prediction but also tackles the current challenges in protein annotation, particularly the growing backlog of uncharacterized proteins. As a result, the proposed model offers a faster and more reliable method for the functional characterization of newly discovered proteins, driving forward progress in biomedical research and biotechnology, and enabling more informed and efficient applications in these fields.

1.5 STRUCTURE OF THE THESIS

Chapter one provides an introduction to the project, outlining the research problem, aims, objectives, and the significance of the study. It sets the foundation for understanding the necessity of advancing protein function prediction using innovative computational techniques.

Chapter two delves into a comprehensive review of the literature, covering traditional protein function prediction methods, recent advances with deep learning models, and the application of protein language models (PLMs) and graph convolutional networks (GCNs). This chapter highlights the current state of research, identifying gaps that this study aims to address.

Chapter three discusses the research methodology employed in this study. It explains the philosophical approach, research design, and the specific methodologies used, including the data collection methods, tools, and the detailed steps taken to develop and evaluate the proposed model.

Chapter four focuses on the design and implementation of the proposed protein function prediction model. This chapter elaborates on how the ESM-1b model and GCNs were integrated, providing technical details of the model architecture, training process, and the computational frameworks used.

Chapter five presents the testing and evaluation of the model. It includes a detailed analysis of the model's performance against existing state-of-the-art methods, using evaluation metrics such as Fmax and AUPR, and discusses the implications of the findings.

Chapter six concludes the project with a summary of the findings, conclusions drawn from the research, and recommendations for future work. This chapter reflects on the success of the project in meeting its objectives and suggests potential areas for further research to enhance protein function prediction.

CHAPTER 2 - REVIEW OF LITERATURE

The purpose of this chapter is to provide a comprehensive analysis of the existing body of literature related to protein function prediction using Convolutional Neural Network (CNN) and Graph Neural Networks (GNNs). This chapter identifies key trends, arguments, and gaps within the field, focusing on both traditional methods and modern deep learning approaches. The scope of the literature covered includes various subfields such as protein language models, traditional prediction methods, deep learning-based models, and the application of Gene Ontology (GO). The timeframe spans from foundational methods to the latest advancements in the field.

2.1 PROTEIN LANGUAGE MODELS

A protein language model is the transfer application of the language models in the field of biochemistry enabling tasks such as protein structure prediction, protein function prediction, and sequence generation[17]. It takes protein sequences as input and learns the underlying biochemical properties, secondary and tertiary structures, and functional patterns, A language model is a type of neural network that can predict the next character or word based on previous text, learning the statistical patterns of characters or words in a given language and generating new sequences that adhere to these patterns[18]. Language models encompass various architectures including recurrent, convolutional neural networks, and Transformer-based models, widely applied in natural language processing tasks.

One of the pioneering models in this area is Evolutionary Scale Modeling (ESM), developed by Facebook AI Research[11]. Evolutionary Scale Modeling (ESM) employs Transformer architectures to generate representations of protein

sequences that capture their evolutionary and functional nuances. These models have shown great promise in tasks such as predicting protein structure and function directly from sequence data, providing a deep understanding of protein dynamics without the need for traditional experimental methods.

Similarly, UniRep [12], developed by researchers at Harvard, utilizes a recurrent neural network (RNN) approach to condense protein sequences into fixed-length vectors. This model has been effectively used in predicting protein stability and fluorescence, showcasing its utility in both basic biological research and practical applications such as biotechnology.

On the other hand, ProtTrans extends the BERT and T5 models from Natural Language Processing (NLP) to the protein sequencing field, adapting these powerful Transformer-based models to tackle protein-related tasks such as structure prediction and function classification[13]. This adaptation underscores the versatility of NLP techniques in extracting meaningful patterns from biological data.

ProteinBERT takes a direct cue from its NLP counterpart, applying the Masked Language Model (MLM) approach to predict amino acids in sequences[14]. This methodology helps in understanding protein functions and interactions, thereby aiding in the annotation of unknown proteins and the exploration of genetic variations.

Moreover, DeepSequence utilizes variational autoencoders to study the effects of genetic mutations on protein functionality [19]. This model provides insights into how alterations in protein sequences can impact their biological function, which is crucial for understanding genetic disorders and guiding the engineering of novel proteins.

While not a traditional language model, AlphaFold by DeepMind has revolutionized structural biology by predicting protein structures with unprecedented accuracy[20]. AlphaFold's approach, which can be seen as an extension of language modeling principles to structural prediction, has been transformative, offering detailed protein structure predictions that can accelerate drug discovery and biological research.

The protein language model ESM which is used in this study is an open-source project introduced by Facebook Research [11]. It takes protein sequences as input and is trained as a high-capacity Transformer with hyperparameter optimization. After training, the model produces feature representations that contain implicit information about the protein's secondary and tertiary structures, functions, homology, and more. Moreover, these representations can be visualized through linear projection. Literature shows the evidence of several methods models of protein function prediction such as traditional models [21], [22], [23] ,CNN [6], [7], [9], [24], GNN [25], [26], [27] and GCN [28], [29], [30] models.

2.2 TRADITIONAL PROTEIN FUNCTION PREDICTION MODELS

Traditional models in protein function prediction rely heavily on sequence similarity and structural homology to infer function, leveraging well-established databases and algorithms to compare unknown proteins with characterized ones[31]. These models often use techniques such as sequence alignment and motif detection to identify functional similarities.

The Naive method is one of the benchmark methods used for comparing protein function predictions in CAFA [32]. Its principle relies on the hierarchical structure of Gene Ontology (GO), where lower-level GO terms propagate upwards, resulting in the aggregation of numerous functional annotations at higher-level GO terms. Under the assumption of annotating the same set of GO terms for all proteins, comparable prediction results can be obtained based on annotation frequencies.

The BLAST-KNN method is a K-Nearest Neighbors approach based on protein sequence similarity scores, leveraging the classical sequence alignment tool BLAST [21].

Within the domain of machine learning, logistic regression stands out as one of the extensively employed algorithms. In a study by You et al.[22], text data sourced

from the MEDLINE biomedical literature database underwent transformation into text features. Logistic regression was subsequently utilized for training, aiming to forecast the correlation between protein molecular function, biological process, and cellular component. DeepText2GO, significantly outperformed both text-based and sequence-based methods. Specifically, DeepText2GO achieved higher F-max scores (0.627 for MFO, 0.442 for BPO, and 0.694 for CCO), lower S-min scores (5.240 for MFO, 17.713 for BPO, and 4.531 for CCO), and higher AUPR scores (0.605 for MFO, 0.336 for BPO, and 0.729 for CCO) compared to other models. This demonstrates the model's superiority in leveraging deep semantic representations and integrating various data sources to enhance protein function prediction accuracy. Another innovation by Lee et al.[23] introduced a protein interaction network kernel logistic regression model. Leveraging the diffusion kernel, this model demonstrated superior prediction accuracy compared to a model based on Markov random field for protein function prediction.

2.3 DEEP LEARNING BASED PROTEIN FUNCTION PREDICTION MODELS

Deep learning has become an effective method for predicting the function of proteins by using its capacity to automatically identify and understand intricate patterns in vast amounts of biological data. Convolutional layers allow convolutional neural networks (CNNs) to automatically and adaptively learn spatial hierarchies of features. CNNs are a class of deep learning models that are mostly employed for processing grid-like data[33], such as photographs. This idea is extended to graph-structured data by Graph Neural Networks (GNNs), which capture dependencies between nodes in a graph[34]. Another variant of GNNs that generalize the convolution operation to graph data is Graph Convolutional Networks (GCNs), which allow for efficient learning of node representations by aggregating data from neighbors.

2.3.1 CNN based models in protein function prediction

Using protein sequences and known interactions, Kulmanov et al.[24] presented a new method for protein function prediction called DeepGO. This model combines two multilayer neural network-based representation learning algorithms to extract features useful for predicting protein functions. One method focuses on learning features from protein sequences, while the other learns protein representations based on their positions within the protein interaction network. The sequence features undergo processing through a one-dimensional convolutional layer and a subsequent max pooling operation. After that, characteristics from the protein interaction network are fused with the output of max pooling to create a fully connected layer that houses 1024 neurons. The ultimate classification is executed through a hierarchical neural network employing an S-shaped activation function. DeepGO's performance was evaluated using the standards set by the CAFA challenge, demonstrating significant improvements over baseline methods such as BLAST, particularly in predicting cellular locations. The model achieved higher F-max scores across the GO sub-ontologies: 0.395 for Biological Process (BP), 0.470 for Molecular Function (MF), and 0.633 for Cellular Component (CC). These results underscore DeepGO's ability to capture both explicit and implicit dependencies between GO terms, leading to more accurate protein function predictions.

Two years later, Kulmanov et al.[6] introduced the DeepGOplus model. This model employs a parameterless one-hot encoding, replacing the embedding layer, resulting in a substantial reduction in the number of parameters. The embedding layer, susceptible to memorizing training data, can potentially lead to overfitting. In contrast to DeepGO, where each convolutional layer shares the same filter, DeepGOplus configures different filters for periodic convolutional layers. Additionally, DeepGOplus utilizes a flat classification layer instead of a hierarchical classifier. This adaptation is necessary because DeepGOplus constructs a unified model encompassing over 5,000 classes for all three GO ontology terms. The constraints of memory and time complexity preclude the establishment of a hierarchical classifier for such a large number of classes.

Du et al.[7] enhanced the DeepGO model, introducing the DeepAdd model. In this iteration, protein sequences are treated akin to natural language, and the word2vec method is employed to define a feature set representing proteins. DeepAdd integrates a sequence similarity graph to learn features that leverage functional relationships across various similarity levels. In instances where the protein interaction network features for a target protein are absent, the sequence similarity features of the protein are deduced as supplementary features. These sequence similarity features comprise a compilation of scores indicating the sequence similarities calculated with respect to all sequences in the training set. The performance of DeepAdd was evaluated against various baseline models, including DeepGO, on datasets such as CAFA3 and SwissProt. The evaluation metrics included Fmax, AUC (Area Under the ROC Curve), and MCC (Mathews Correlation Coefficient). On the CAFA3 dataset, DeepAdd achieved Fmax scores of 0.345 for Biological Process (BP), 0.516 for Molecular Function (MF), and 0.547 for Cellular Component (CC); AUC scores of 0.896 for BP, 0.912 for MF, and 0.958 for CC; and MCC scores of 0.335 for BP, 0.585 for MF, and 0.511 for CC. On the SwissProt dataset, DeepAdd achieved Fmax scores of 0.393 for BP, 0.580 for MF, and 0.619 for CC; AUC scores of 0.907 for BP, 0.947 for MF, and 0.968 for CC; and MCC scores of 0.395 for BP, 0.606 for MF, and 0.592 for CC. These results demonstrate that DeepAdd consistently outperforms the baseline models across different GO sub-ontologies, particularly in scenarios where PPI data is missing or incomplete.

Renfrew et al.[9] introduced DeepFri, a model employing long short- term memory networks and graph convolutional networks for protein function prediction. The model takes both sequence and sequence- based features (structure predicted from the sequence) as input. The initial segment of the model constitutes a self-supervised language model structured with a recursive neural network incorporating long short-term memory. Pre-training is performed on the protein family database, utilizing it to extract residue features from sequences within the PDB database. The subsequent part of the model is a graph convolutional neural network that employs graph convolution to transmit residue-level features among neighboring residues, constructing a feature representation for the protein. In the final step, all the features

output from the graph convolutional layers are concatenated and fed into a fully connected layer for protein function. The outcomes of DeepFRI are noteworthy. When evaluated on experimentally annotated protein structures from the PDB, DeepFRI achieved an Fmax score of 0.657 for native structures, outperforming the sequence-only CNN-based method DeepGO, which had an Fmax score of 0.525. Furthermore, DeepFRI showed robustness in predicting functions of proteins with low sequence identity to the training set, achieving a median Fmax of 0.545 for proteins with ≤30% sequence identity, compared to 0.514 for FunFams and 0.491 for DeepGO .

2.3.2 GNN based models

Termed Graph Residual Neural Network (GRNN)[25], employs multi-relational graphs and utilizes learnable parameters to weigh the influence of different relations. This architecture combines local information from input data through parameterized linear transformations and non-linear functions, progressively extracting useful information. GRNN's residual layers allow for increased flexibility by capturing multiple types of diffusion, thus enhancing the learning capacity of the network. Through numerical tests on protein networks, the study proved the efficacy of GRNN, demonstrating notable performance advantages over state-of-the-art alternatives.

The outcomes of the GRNN framework are noteworthy. When evaluated on protein-to-protein interaction datasets, GRNN achieved a macro F1 score of 0.86 for the brain cells dataset with 440 labeled nodes, significantly outperforming the single-relational Graph Convolutional Network (GCN) which had a macro F1 score of 0.49. Similarly, for the circulation cells dataset, GRNN attained a macro F1 score of 0.77, while GCN scored 0.48. In the generic cells dataset, GRNN scored 0.70 compared to 0.49 for GCN. These results highlight GRNN's robustness and superior performance in predicting protein functions across multiple cell types, validating its potential as a powerful tool in bioinformatics.

By adding characteristics from the Evolutionary Scale Modeling (ESM) of proteins[26], which creates sequence embeddings using transformers trained on 250 million protein sequences, PANDA2 expands upon these developments. This

integration boosts PANDA2's prediction capability by enabling it to extract structural and sequence information from proteins. PANDA2 tied for first place in Biological Process Ontology (BPO) with a Fmax score of 0.3964 but a higher coverage rate, placed first in Cellular Component Ontology (CCO) with a Fmax score of 0.6374, and second in Molecular Function Ontology (MFO) with a Fmax score of 0.5849 when compared to top-performing methods in the CAFA3 challenge[32].These findings demonstrate the reliability and efficiency of PANDA2 in protein function prediction, which makes it an invaluable resource for bioinformatics research.

By applying a graph neural network (GNN) framework to anticipate the impact of mutations on protein stability, ProS-GNN (Protein Stability Graph Neural Network) displays notable gains. ProS-GNN uses GNNs to describe the complex interactions between atoms in protein structures. It does this by using message passing to capture the links between molecular structure and property and by integrating raw atom coordinates to provide spatial insights. This method builds upon previous work that used convolutional neural networks (CNNs) and other deep learning approaches for predicting protein stability changes, but it uniquely focuses on the structural data of proteins, improving the accuracy and efficiency of its predictions.

The ProS-GNN results are remarkable. ProS-GNN demonstrated excellent results in terms of bias reduction and data generalization when it was trained and evaluated on many datasets [27]. In particular, ProS-GNN obtained a Pearson correlation coefficient (r) of 0.62 and a root mean square error (RMSE) (σ) of 1.11 for direct mutations and r = 0.60 and σ = 1.12 for reverse mutations when tested on the S2648 dataset. ProS-GNN obtained r = 0.61, σ = 1.23 for direct mutations and r = 0.56, σ = 1.30 for reverse mutations on the Ssym dataset.

The outcomes of ProS-GNN are noteworthy. When trained and tested on various datasets[27], ProS-GNN achieved high performance in terms of data generalization and bias suppression. Specifically, when evaluated on the S2648 dataset, ProS-GNN achieved a Pearson correlation coefficient (r) of 0.62 and a root mean square error (RMSE) (σ) of 1.11 for direct mutations, and r = 0.60, σ = 1.12 for reverse mutations. On the Ssym dataset, ProS-GNN achieved r = 0.61, σ = 1.23 for direct

mutations, and r = 0.56, σ = 1.30 for reverse mutations. ProS-GNN further shown its robustness and efficiency by outperforming fifteen other algorithms in terms of prediction accuracy on the Ssym dataset. Furthermore, ProS-GNN obtained r = 0.48, σ = 1.27 for direct mutations and r = 0.43, σ = 1.19 for reverse mutations on the Myoglobin dataset. These findings show that ProS-GNN is a useful tool for bioinformatics research and may have therapeutic implications due to its capacity to learn feature representations efficiently and provide precise predictions of changes in protein stability following mutation.

2.3.3 GCN based model

The Graph Convolutional Network (GCN)-based hierarchical multi-label classification framework has demonstrated significant advancements[28]. This system uses sequence data and the hierarchical structure of Gene Ontology (GO) words to predict protein functions by combining GCNs with pre-trained language models. This technique improves on earlier research for sequence-based protein function prediction using convolutional neural networks (CNNs) and other deep learning techniques; however, it combines the hierarchical information of GO keywords in a novel way to increase prediction accuracy.Notable are the results of the hierarchical multi-label categorization framework based on GCN. When evaluated on the CAFA3 dataset, which includes molecular function ontology (MFO), biological process ontology (BPO), and cellular component ontology (CCO), the proposed method outperformed state-of-the-art approaches. Specifically, it achieved an Fmax score of 0.518 for MFO, 0.470 for BPO, and 0.637 for CCO, and an AUPR score of 0.476 for MFO, 0.368 for BPO, and 0.626 for CCO. These results highlight the method's robustness in handling large-scale hierarchical graphs, particularly in the BPO domain, where it significantly improved performance from an Fmax of 0.398 to 0.470 compared to the previous best model, TALE. This demonstrates the model's effectiveness in capturing the complex relationships within the hierarchical structure of GO terms, leading to more accurate protein function predictions

DeepGraphGO, introduced by Ronghui You et al.[29], is a sophisticated model that employs graph neural networks (GNNs) to predict protein functions across multiple species, utilizing both protein sequence and protein network information. The model integrates advanced graph-based methods with sequence analysis, improving upon traditional methods which often consider only one type of data. By leveraging multispecies data, DeepGraphGO enhances the training process and achieves better generalization on protein function prediction tasks. Extensive tests conducted on large-scale datasets using the CAFA (Critical Assessment of protein Function Annotation) settings proved DeepGraphGO's performance. State-of-the-art techniques including DeepGOPlus, GeneMANIA, deepNF, and clusDCA were surpassed by DeepGraphGO. The highest Fmax scores were obtained by DeepGraphGO for Molecular Function Ontology (MFO), Biological Process Ontology (BPO), and Cellular Component Ontology (CCO), with values of 0.623, 0.327, and 0.692, respectively. It also obtained AUPR ratings of 0.543 for MFO, 0.194 for BPO, and 0.695 for CCO, indicating its greater performance and resilience in predicting protein function across several GO domains.

The unsupervised protein embeddings approach leverages pre-trained deep sequence models in an unsupervised setting to extract complex feature representations that are subsequently applied to the supervised task of protein molecular function prediction[30]. This method expands on earlier work with convolutional neural networks (CNNs) and other deep learning architectures, which frequently use direct sequence input or hand-crafted features. By utilizing unsupervised learning, the method captures more intricate relationships within protein sequences, leading to improved predictive performance. The outcomes of this approach are noteworthy. Evaluations on the CAFA3 benchmark demonstrated that the unsupervised protein embeddings method achieved a competitive Fmax score of 0.55, placing it among the top-performing models. Further testing on the PDB dataset yielded an Fmax score of 0.52, an Smin score of 0.48, and a ROCAUC of 0.84. These results indicate the model's robustness and effectiveness in capturing functional information from protein sequences, significantly outperforming traditional methods that rely solely on supervised learning and hand-crafted features.

This success underscores the potential of unsupervised pre-training on large-scale protein sequence data as a powerful tool in protein function prediction.

2.4 GENE ONTOLOGY(GO)

Protein functions are accurately described hierarchically by the Gene Ontology (GO). In the cells of organisms, proteins are involved in membrane transport, signal transduction, recognition, regulation, and catalysis of processes. These protein functions are directly related to their three-dimensional structures and indirectly depend on their DNA sequences[35]. The GO knowledge base mainly consists of two components: GO terms, which provide the logical structure and relationships of biological processes. The relationships between different GO terms mainly include is_a and part_of [36], which can be represented by a directed acyclic graph as shown in Figure 1; and GO annotations, which provide annotations for the GO terms, describing their functions. In GO annotations, GO can be divided into three main categories based on different levels of interdependence among protein functions: Molecular functions refer to activities at the molecular level, such as catalysis. These are usually predicted computationally to determine homologs; Biological processes describe broader functions that are assembled from molecular functions, such as specific metabolic pathways; Cellular components describe the locations within the cell where proteins perform their functions, such as the nucleus or cytoplasm[35].

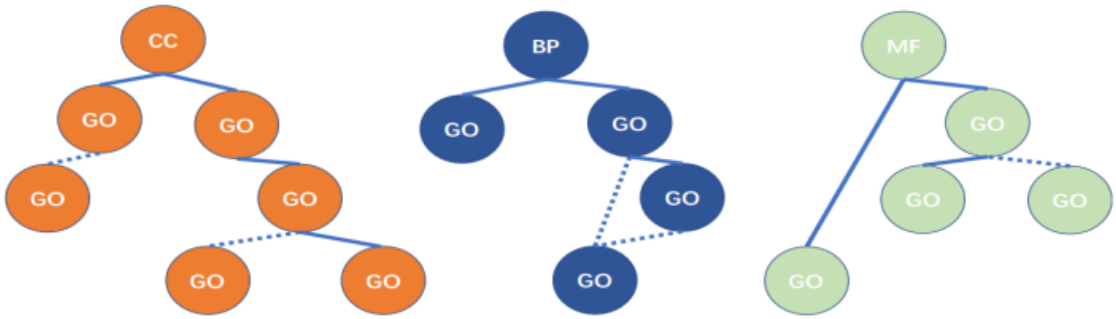

Figure 1:Example of a hierarchical structure for GO

Proteome function prediction is a difficult multi-classification problem since proteins might have many activities in the domains of biological processes, cellular components, and molecular functions[37]. Because deep learning is so good at extracting meaning from high-dimensional data, it may be applied to a wide range

of classification issues, including the prediction of protein function. Additionally, deep learning requires minimal manual processes, thus it can easily utilize the increasing availability of computational resources and data.

2.5 GCN ARCHITECTURE

A potent paradigm for learning from graph-structured data—which is common in fields including social networks, biological networks, and recommendation systems—has emerged: graph neural networks, or GNNs[38]. A typical GNN operates by representing data as nodes and edges, where nodes signify entities (e.g., proteins, users) and edges represent their relationships (e.g., interactions, friendships). Each node is characterized by its features, and during the learning process, nodes aggregate information from their neighbors to update their own representations. This process, known as message passing, involves iterating through the graph to refine node states until a meaningful embedding is obtained for each node. Then, different tasks like node classification, connection prediction, or graph classification can be performed using these embeddings.

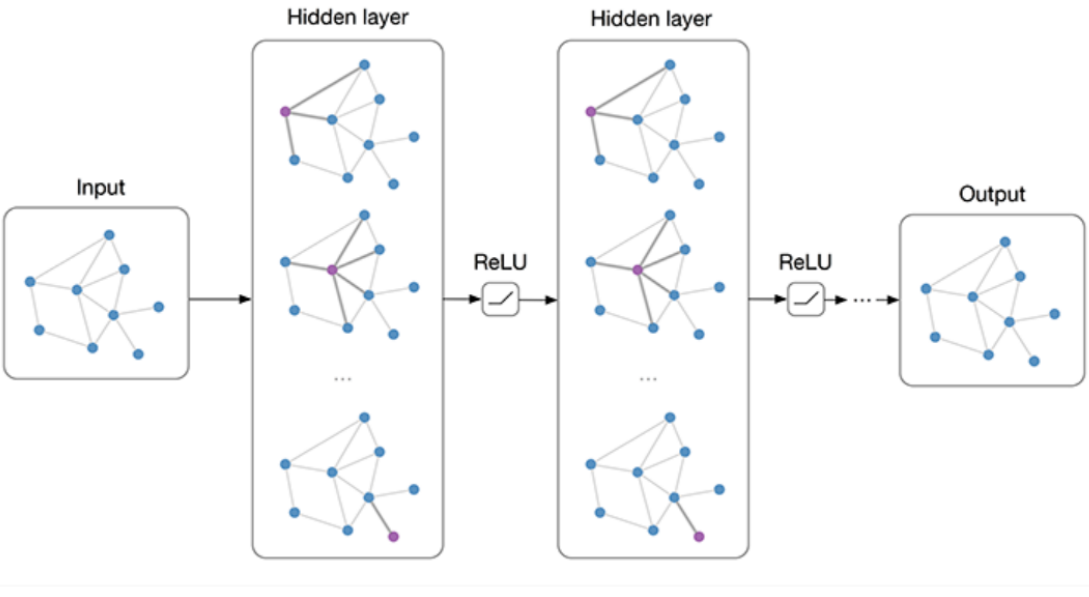

Figure 2:Multi-layer Graph Convolutional Network (GCN)[39]

Graph Convolutional Networks (GCNs) represent a significant evolution in the field of GNNs[39], introducing a specific type of convolutional operation adapted for graphs. Unlike traditional GNNs that utilize generic message passing, GCNs apply convolutional operations to graph data, which allows them to efficiently aggregate and transform node features from their local neighborhoods. The pioneering work by Kipf and Welling (2017)[40] simplified spectral convolutions to make GCNs more computationally efficient, thus enabling the application of deep learning techniques to larger graph datasets. By normalizing the adjacency matrix and employing layer-wise propagation rules, GCNs effectively capture the local structure of graphs, leading to improved performance in tasks like semi-supervised node classification.

The development of advanced GCN architectures, such as Graph Attention Networks (GAT) and Graph Isomorphism Networks (GIN)[41], [42], further enhanced the expressiveness and applicability of GNNs. GATs introduced attention mechanisms to dynamically weigh the importance of neighboring nodes during feature aggregation, thereby allowing the model to focus on more relevant connections. GINs, on the other hand, improved the discriminative power of GNNs by closely mimicking the Weisfeiler-Lehman graph isomorphism test[43], ensuring better differentiation between non-isomorphic graphs. These advancements have broadened the scope of GNN applications, making them indispensable in fields like bioinformatics, where they are used for protein function prediction and other complex biological tasks, demonstrating their ability to learn from both sequence and structural information effectively.

2.6 LITERATURE SUMMARY

The literature review highlights significant advancements in protein function prediction, transitioning from traditional sequence and homology-based methods to modern deep learning models that incorporate a variety of data sources and advanced computational techniques. Traditional methods, while effective in specific contexts, often struggle with proteins that lack homologous sequences or well-characterized motifs. Large-scale biological data may now be automatically mined for complex patterns through the use of Convolutional Neural Networks (CNNs) and

Graph Neural Networks (GNNs), which are the key components of deep learning. Important discoveries highlight how well different data types—like sequence, structural, and interaction data—integrate and how deep learning models perform better at capturing the complex interactions seen in biological systems. However, gaps remain in efficiently handling large-scale data, integrating multi-source heterogeneous data, and improving model interpretability and robustness. This study's proposed method aims to address these gaps by leveraging advanced GNN architectures and pre-trained protein language models, promising to set new standards in protein function prediction.

2.7 CHAPTER SUMMARY

This chapter has provided a detailed review of the literature on protein function prediction, emphasizing the evolution of methods from traditional sequence and homology-based approaches to state-of-the-art deep learning-based models. It has identified key contributions and existing gaps within the field, highlighting the challenges and opportunities in integrating multi-source data and improving prediction accuracy. This comprehensive analysis sets the stage for the proposed study, which seeks to enhance protein function prediction by employing advanced machine learning techniques, including GNNs and pre-trained protein language models. The following chapter presents the methodology, experimental design, performance evaluation metrics and outlining the steps taken to achieve the aim of the study.

CHAPTER 3 - RESEARCH METHODOLOGY

This chapter outlines the research methodology employed in this study, detailing the research methods, research model, data analysis, evaluation metrics, and research materials. The methodology is designed to provide a systematic approach to achieve the research objectives and ensure the reliability and validity of the results.

3.1 RESEARCH METHOD

The research philosophy adopted for this study is positivism. Positivism involves the use of rigorous and systematic approaches to investigate phenomena[44], relying on observable and measurable facts. This approach is consistent with the goal of the study, which is to create a unique protein function prediction model using quantifiable information from InterPro domains, protein-protein interaction networks, and protein sequences. Adopting a positivist approach ensures objectivity and reliability by focusing on measurable phenomena[24], [29].

The research approach utilized is deductive. Deductive research begins with a theoretical framework, followed by the collection and analysis of data to test hypotheses derived from that theory[45]. In this study, existing theories and models related to protein function prediction, such as those using protein language models and graph convolutional networks, are used as the foundation. The hypotheses are then tested through the development and evaluation of the proposed model. This approach facilitates the testing of existing theories and models in the context of protein function prediction.

Quantitative methodology was selected for this study. Numerical data is gathered and analyzed using quantitative methods in order to identify trends, evaluate theories, and forecast outcomes [45], [46]. This method works well for assessing the protein function prediction model's performance using statistical measures including Fmax, accuracy, precision, and recall [47], [48], [49]. Quantitative methods provide robust statistical analysis to evaluate model performance[46], [50].

An experimental research strategy is being used. In experimental research, variables are changed to see how they affect other variables. [51]. In this article, experiments are carried out using machine learning techniques and data integration from various sources to evaluate the performance of the protein function prediction model. An experimental strategy allows for controlled testing of model variables and their effects.

This research has a cross-sectional time horizon. Studies that examine cross-sectional data do so at a particular point in time[52]. This method works well for assessing how well the protein function prediction model is performing right now with the datasets that are accessible. A cross-sectional time horizon enables a snapshot evaluation of the model's current performance.

The research employed many approaches and procedures, such as data collecting, model creation, and performance evaluation. Data is collected from UniProt[53], STRING[54], and SwissProt databases[55]. The PyTorch and DGL frameworks are utilized in the development of the model, and a Linux server with 32GB of RAM and an NVIDIA 1080TI GPU with 12GB of VRAM is used for the tests. Metrics for performance evaluation, including Fmax, recall, accuracy, and precision, are utilized to evaluate the model [48]. These techniques and procedures ensure comprehensive data collection, model development, and rigorous performance evaluation.

This structured approach, guided by the research onion framework, ensures a thorough and methodical investigation into protein function prediction, leveraging advanced computational techniques and high-quality datasets to achieve reliable and accurate results.

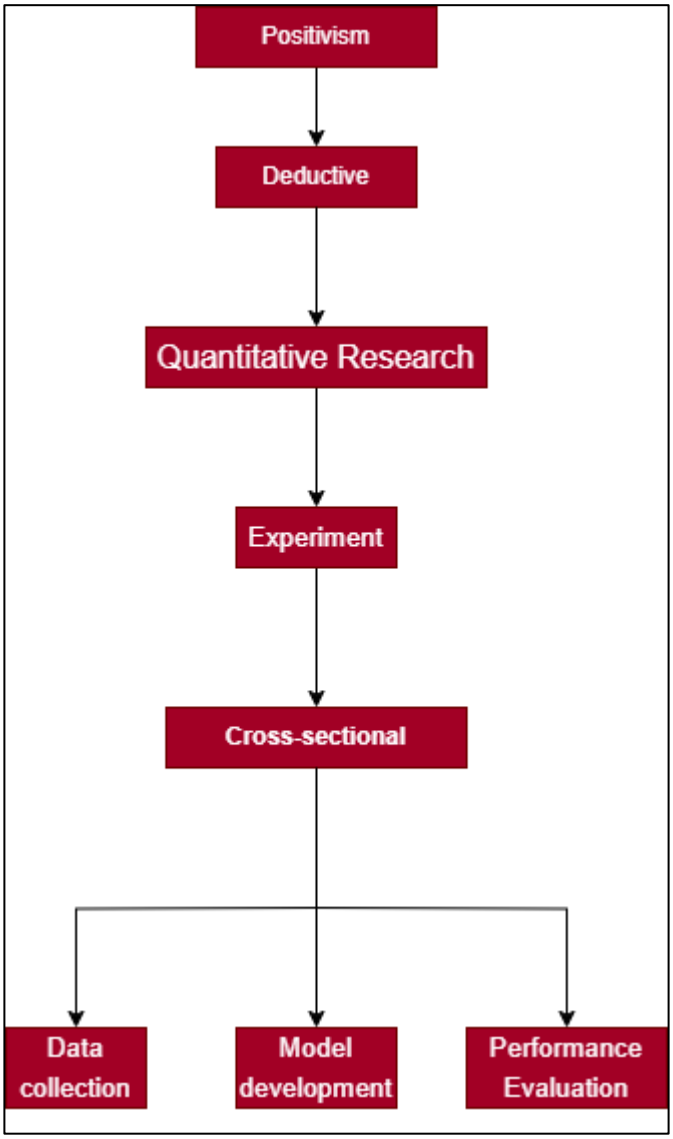

Figure 3:Research DESIGN

3.2 PROPOSED RESEARCH MODEL

This research methodology first employs the ESM-1b (Evolutionary Scale Modeling-1b) protein language model to extract protein features. Subsequently, it tunes the parameters of the DeepGraphGO graph neural network to integrate multi-source protein feature information and predict the GO scores of proteins.

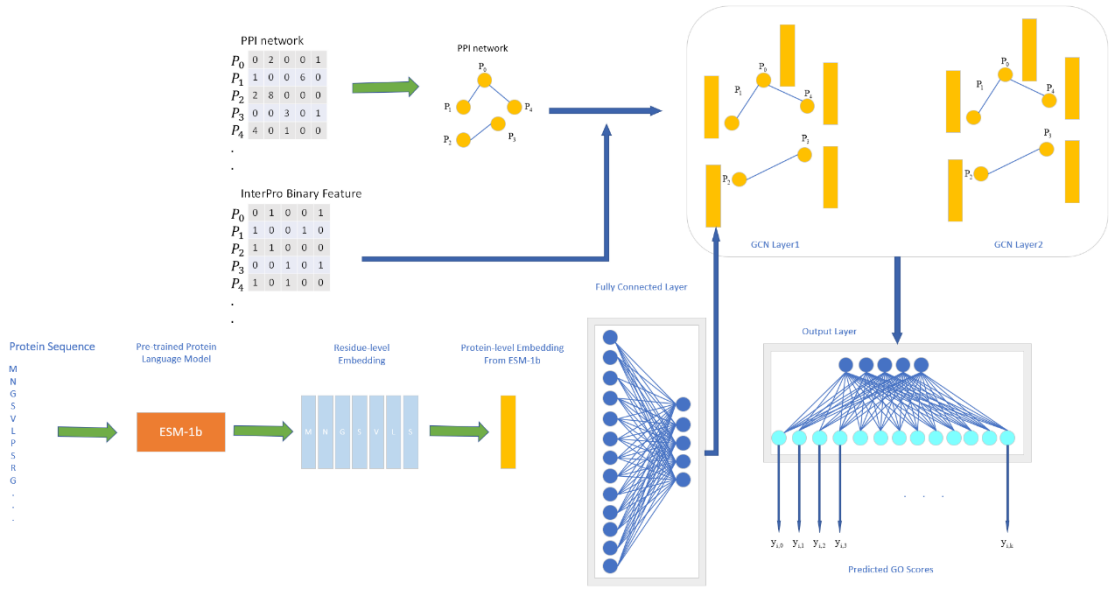

Figure 4:Proposed Model

3.2.1 ESM-1b

The ESM-1b (Evolutionary Scale Modeling-1b) is a deep Transformer-based language model designed to process protein sequences[11]. Its architecture leverages the Transformer model, which has shown exceptional performance in natural language processing tasks[14], [56]. The ESM-1b model is trained on an extensive dataset of 250 million protein sequences, totaling 86 billion amino acids, to capture the evolutionary diversity of proteins.

Multiple layers of feed-forward neural networks and self-attention processes make up the Transformer architecture employed in ESM-1b[11]. The Transformer's layers analyze input sequences through self-attention, enabling the model to assess the relative significance of various amino acids in a sequence. Modeling long-range dependencies and interactions within protein sequences requires the use of this mechanism[57]. These weighted inputs are then subjected to non-linear transformations by the feed-forward networks, which help the model extract intricate patterns and characteristics from the data.

During training, the ESM-1b model uses a masked language modeling objective. In this configuration, the model is trained to anticipate the masked positions based on

the surrounding context, with a portion of the amino acids in each input sequence being randomly masked. By using this method, the model is compelled to acquire meaningful representations of protein sequences that capture the secondary structures, biological characteristics, and evolutionary links included in the data.

The ESM-1b model encodes biochemical properties of amino acids into its representations. These properties include hydrophobicity, polarity, and molecular weight. Visualization techniques such as t-SNE (t-distributed stochastic neighbor embedding) reveal distinct clustering of amino acids based on these biochemical properties[58], indicating that the model effectively captures and utilizes this information. The representations learned by ESM-1b contain rich information about the secondary and tertiary structures of proteins. Linear projections from the model's hidden layers can predict secondary structure elements such as alpha-helices and beta-sheets with high accuracy. Additionally, the model excels in predicting long-range residue-residue contacts, which are essential for determining the three-dimensional conformation of proteins. ESM-1b's ability to detect remote homologs—proteins with similar structures but low sequence identity—surpasses traditional methods. By evaluating the similarity of vector representations in the model's learned space, ESM-1b can identify structurally related proteins even when sequence similarity is minimal. This capability is particularly useful for annotating proteins of unknown function and understanding evolutionary relationships. ESM-1b consistently outperforms baseline models such as LSTMs (Long Short-Term Memory networks) and n-gram models. For instance, in tasks like secondary structure prediction and contact prediction, ESM-1b achieves higher accuracy and precision. The model's performance improves further when trained on diverse and high-capacity datasets, underscoring the importance of data diversity and model scale. The features learned by ESM-1b generalize well across various downstream tasks, including mutational effect prediction and protein engineering[9], [34]. Fine-tuning the model on specific datasets for these tasks yields state-of-the-art results, demonstrating the versatility and robustness of the learned representations.

In summary, the ESM-1b model, with its deep Transformer architecture and extensive training on diverse protein sequences, provides powerful and accurate

representations of protein sequences. These representations capture a wide range of biological information, enabling superior performance in structural prediction, homology detection, and other critical bioinformatics tasks. The success of ESM-1b highlights the potential of large-scale unsupervised learning in advancing our understanding of protein biology.

3.2.2 DeepGraphGO

The research medel employs a model closely aligned with DeepGraphGO, as illustrated in Figure 3.2. The input node features consist of 1280-dimensional protein features trained using the ESM-1b protein language model. The input adjacency matrix represents the top 100 interaction strengths for each node within the protein-protein interaction network. The model processes these inputs through a fully connected layer, followed by two GCN layers and an output layer to generate the predicted GO scores.

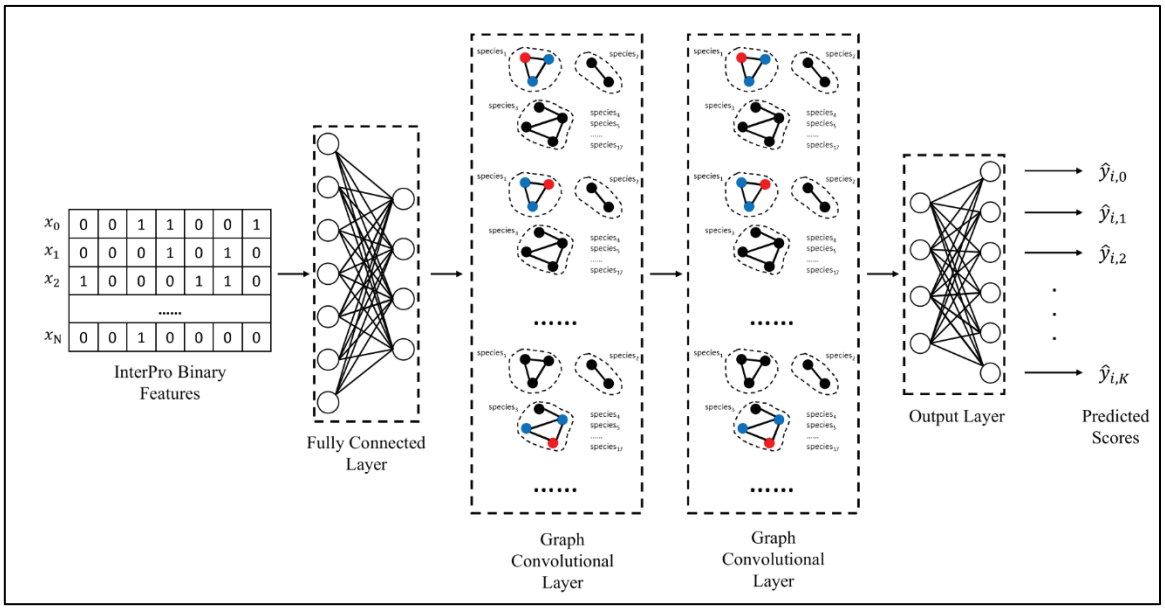

Figure 5:DeepGraphGO (Source : You et al.,[29])

Graph convolutional networks have been proven capable of extracting features from the natural representation of data for one or multiple graphs. The concept of the graph convolutional layers in DeepGraphGO and proposed model is that graph convolutional networks are an appropriate method for extracting features from

protein interaction networks, considering the graph-based structure represented by the protein interaction network. Based on the work of Kipf and Welling [34], the model proposed in this paper updates the representation vectors $\mathrm{H}^{(\mathrm{l})}$in $\mathbb{R}^{\mathrm{N}\times\mathrm{d}}$ of the graph convolutional network layer at the l-th level and the residual connections as follows:

$$\mathbf{H}^{(l)} = f\left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}\mathbf{H}^{(l-1)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)}\right) + \mathbf{H}^{(l-1)} \qquad (3.1)$$

Where A represents the adjacency matrix, and I represents the identity matrix of dimension N, and $\tilde{A} = A + I.\widetilde{D}$is the degree matrix of $\tilde{A}.\mathbf{W}^{(l)}$ in $\mathbb{R}^{\mathrm{d}\times\mathrm{d}}$ and $\mathbf{b}^{(l)}$ in $\mathbb{R}^{\mathrm{d}}$ are the weights and biases, respectively. M consecutive graph convolutional layers can capture high-order node information of order M.

The output layer primarily transforms the matrix obtained from information propagation and mathematical calculations of the previous model to change its dimensionality to match the number of GO categories[29], [58]. The resulting vector for each item represents the predicted score for each GO category. The anticipated score for every GO category is shown in the resulting vector for every item. An activation function and a fully linked layer are used to implement the output layer. The fully connected layer gets its name because it uses all local features to function as a classifier throughout the entire model. By connecting every node in this layer to every other layer's node, the previously extracted features are integrated. The feature matrix and the weights of the fully connected layer must match because the dimensionality of the weight matrix in the fully connected layer stays constant, requiring the input dimensions of the feature matrix from the preceding layer to be the same.

Activation functions are a key part of neural network design. Their role is to make the model more flexible. Each deep learning neural network model must choose activation functions based on particular conditions because different hidden layer activation functions will produce different learning outcomes for the network model on the training dataset; the output layer activation functions will determine the kinds of predictions the model can make. The sigmoid activation function, sometimes

referred to as the logistic function, is the activation function that is employed in this article. Algorithms for logistic regression classification employ the same function. Any real value between 0 and 1 can be used as the input and output of this function. The output value approaches 1 in proportion to the input value, and it approaches 0 in proportion to the input value. For the ith protein and the jth GO category, the predicted score $\hat{y}_{ij}$ is obtained through the following output layer:

$$\hat{y}_{ij} = \sigma\left(\mathbf{w}_j^{(o)} h_i + \mathbf{b}_j^{(o)}\right) \quad (3.2)$$

Where $\mathbf{w}_j^{(o)} \in \mathbb{R}^{\mathrm{d}}$ and $\mathbf{b}_j^{(o)} \in \mathbb{R}$, are the weights and deviations of the functions predicting the j-th GO class, respectively, and σ is the activation function[29].

The difference between the computation algorithm's intended output and current output determines the value of the loss function. One way to evaluate the impact of a data modeling algorithm is to use the loss function. It can be separated into two categories: regression (continuous values) and classification (discrete values). This article uses a binary cross-entropy function as its loss function. The likelihood of each prediction is compared to the actual class output—which can be either 0 or 1—using binary cross-entropy. The difference between the predicted value and the expected value will then be used to compute the probability score, which indicates how near or how far the actual value is from the predicted value. The negative average of the logarithm of the corrected predicted probabilities is, in essence, the binary cross-entropy. Its expression is as follows:

$$J = -\frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log\left(\hat{y}_{ij}\right) + \left(1 - y_{ij}\right)\log\left(1 - \hat{y}_{ij}\right) \quad (3.3)$$

Where, K, means the number of GO classes, y_{ij} is the true value and $\hat{y}_{ij}$ is the predicted value.

3.3 DATA ANALYSIS AND EVALUATION METRICS

This study's data analysis uses a rigorous methodology that incorporates quantitative techniques to assess the prediction model's performance in detail. [59].

Through the use of quantitative analysis, the research is able to obtain a more nuanced knowledge of the behavior of the model and identify possible areas for improvement by delving deeper into the underlying patterns and correlations within the data. Furthermore, quantitative analysis provides a strong statistical assessment that makes it possible to precisely analyze the model's accuracy, recall, precision, and other important metrics.[47], [60]. By combining the advantages of qualitative insights with the thoroughness of quantitative evaluation, this dual approach guarantees a comprehensive review that fully validates the prediction model's efficacy and reliability.

3.3.1 Fmax

In the field of bioinformatics, Fmax is a crucial evaluation metric that is especially useful for jobs involving multi-label classification issues, such as protein function prediction. It stands for the highest F-measure, which is determined across several thresholds[47], [61]. It is a harmonic mean of precision and recall. Because it takes into account both precision (the accuracy of positive predictions) and recall (the capacity to locate all pertinent instances), the Fmax metric offers a fair assessment of a model's performance.

A protein may fall into more than one functional category in multi-label classification tasks like protein function prediction, which increases the complexity of evaluating prediction models. The model's performance may not be fully captured by conventional metrics like accuracy, particularly when dealing with imbalanced datasets. To tackle this, Fmax offers a solitary metric that strikes a compromise between recall and precision, guaranteeing that false positives and false negatives are considered equally [62], [63].

By integrating accuracy and recall, Fmax provides a fair assessment of prediction models. This is important in situations where both false positives and false negatives might have serious consequences, like in biological research. Fmax, which is calculated across a range of thresholds, gives researchers the option to choose the best threshold for their particular application. This is especially helpful when working with datasets that exhibit variable levels of class imbalance. Fmax, which combines

precision and recall, offers a more thorough assessment of a model's performance than accuracy alone. This makes it crucial in multi-label classification, as the existence of several classes can impede the evaluation process.

The calculation of Fmax involves several steps. Firstly, precision and recall are computed at various thresholds. Precision (P) and recall (R) are defined as:

$$\text{Precision}=\frac{TP}{TP+FP}=\frac{TP}{\text{all detections}} \tag{3.4}$$

$$\text{Recall}=\frac{TP}{TP+FP}=\frac{TP}{\text{all ground truth}} \tag{3.5}$$

whereas genuine positives that were not projected as such are called false negatives (FN), false positives (FP) are called erroneously predicted positive samples, and true positives (TP) are called properly predicted positive samples [47].

Fmax refers to the F-measure value of the protein center calculated at all predicted thresholds. First, the mean precision and recall were calculated using the following formula:

$$\begin{aligned} pr_i(t) &= \frac{\sum_f I(f\in P_i(t)\wedge f\in T_i)}{\sum_f I(f\in P_i(t))} \\ rc_i(t) &= \frac{\sum_f I(f\in P_i(t)\wedge f\in T_i)}{\sum_f I(f\in T_i)} \\ AvgPr(t) &= \frac{1}{m(t)}\cdot\sum_{i=1}^{m(t)} p\, r_i(t) \\ AvgRc(t) &= \frac{1}{n}\cdot\sum_{i=1}^{n} r_i\,(t) \end{aligned} \tag{3.6}$$

And f represents a GO class; T_i represents the correctly annotated set; $P_i(t)$ represents the annotated set of predicted proteins i at the threshold t; $m(t)$ represents the number of predicted proteins with more than one function; And n represents the number of all proteins; I represents a recognition function. The prediction correctly returns 1, otherwise it returns 0. Then, Fmax was calculated at a threshold t ∈[0,1] with an update pace of 0.01. If the predicted score of a GO class is greater than t, the protein is considered predicted to have this function:

$$F_{\max} = \max_{t} \left\{ \frac{2 \cdot AvgPr(t) \cdot AvgRc(t)}{AvgPr(t) + AvgRc(t)} \right\} \quad (3.7)$$

In practice, Fmax is particularly useful in evaluating models for protein function prediction. Given the complexity and multi-label nature of protein functions, Fmax helps in assessing how well the model can predict multiple functions simultaneously. By using this metric, researchers can fine-tune their models to achieve a better balance between precision and recall, ultimately leading to more accurate and reliable predictions[64].

In summary, Fmax is an essential metric for evaluating multi-label classification models, providing a balanced and comprehensive measure of performance. Its calculation through the optimization of the F1 score across various thresholds ensures that both precision and recall are adequately considered, making it a preferred choice for complex prediction tasks in bioinformatics.

3.3.2 Area Under the Precision-Recall Curve

A crucial assessment statistic in machine learning is Area Under the Precision-Recall Curve (AUPR), especially when dealing with classification issues with unbalanced data [48]. The Precision-Recall (PR) curve, which plots recall—the percentage of genuine positive outcomes among all actual positive instances—against precision—the percentage of true positive results among all positive results predicted by the model—is the source of AUPR. This curve is obtained at different threshold settings.

When there is an imbalance in the dataset—that is, when there are much fewer positive cases than negative cases—AUPR becomes especially useful. In these situations, conventional criteria such as accuracy may be deceptive due to their potential dominance by the majority class [47]. AUPR provides a more informative picture by focusing on the performance of the model with respect to the minority class. AUPR is highly sensitive to the imbalance between classes, making it an ideal metric for applications where the positive class is rare. This sensitivity helps in evaluating the model's ability to correctly identify the minority class without being

overwhelmed by the majority class. By considering both precision and recall, AUPR gives a comprehensive measure of a model's ability to retrieve all relevant instances (recall) while minimizing false positives (precision). This is crucial in applications like medical diagnosis, fraud detection, and information retrieval, where both precision and recall are important[48]. AUPR integrates performance across all possible thresholds, providing a single scalar value that summarizes the model's performance. This makes it easier to compare different models or to assess the performance of a single model without worrying about the choice of threshold.

The Precision-Recall curve is integrated to determine the AUPR. Plotting precision and recall values at different threshold levels results in this curve. The definitions of precision (P) and recall (R) are given in equation 3.1, where true positives (TP) are positive samples that were successfully predicted, false positives (FP) are positive samples that were mistakenly forecasted, and false negatives (FN) are positive samples that were not predicted in the first place.

To create the PR curve, these precision and recall values are calculated at multiple threshold levels. The AUPR is then the area under this curve, which can be computed using numerical integration methods such as the trapezoidal rule. The mathematical expression for AUPR can be represented as:

$$\text{AUPR} = \int_0^1 P(R)dR \tag{3.8}$$

This integral can be approximated by summing the areas of the trapezoids formed between successive points on the PR curve:

$$\text{AUPR} = \sum_{i=1}^{n-1} (R_{i+1} - R_i) \times \frac{P_i + P_{i+1}}{2} \tag{3.9}$$

where P_i and R_i are the precision and recall at the i-th threshold, and n is the number of thresholds[64].

AUPR is particularly useful in fields such as bioinformatics, medical diagnosis, and information retrieval, where identifying the positive class accurately is more critical than predicting the majority class correctly. For instance, in protein function

prediction, accurately identifying the correct function (positive class) among a large number of non-functions (negative class) is crucial. AUPR provides a meaningful evaluation metric that reflects the model’s capability to perform well under these conditions.

In summary, AUPR is an essential metric for evaluating classification models, especially in imbalanced datasets. It provides a balanced measure of a model’s precision and recall across all thresholds, making it a preferred choice for assessing performance in scenarios where accurately predicting the minority class is critical. Its calculation through the integration of the Precision-Recall curve ensures a comprehensive evaluation of the model's ability to distinguish between positive and negative classes.

3.4 RESEARCH MATERIALS

This section provides an overview of the materials used in this research, including the datasets, software, and hardware. The materials are selected to ensure the robustness, reliability, and efficiency of the research process, aligning with the research objectives.

3.4.1 Research Data

The bioinformatics community has initiated competitions such as the Critical Assessment of Functional Annotation (CAFA) challenge to address performance evaluation issues in automatic protein function prediction[32]. CAFA provides guidelines for constructing datasets for protein function prediction problems and criteria for evaluating prediction results. This article's protein data uses the same dataset as DeepGraphGO[29], following CAFA's principles and using the same 17

reference species as CAFA4. The protein sequence data is sourced from UniProt [53], totaling about 18,000 entries. The protein interaction network data comes from the eleventh edition of the STRING database [54], covering approximately 24 million proteins. The GO terms data is sourced from SwissProt [55], extracting all experimental annotation data, categories including: 'IDA', 'IPI', 'EXP', 'IGI', 'IMP', 'IEP', 'IC', or 'TA', all of which are combined to form an annotation dataset. Ultimately, a fasta file containing about 18,000 protein sequence data entries, a filtered protein interaction network matrix (where positions in the matrix are non-zero if there is an interaction between two proteins), and a text file containing the GO terms associated with the respective proteins will be obtained.For future dataset statistics, see Table 1.

Table 4. 1:Datasets from Deepgraphgo[29]

Datasets	MFO	BPO	CCO
Train	*35092*	54276	48093
Valid	490	1579	923
Test	426	925	1224
Total	36008	56780	50240

UniProt: The Universal Protein Resource UniProt (https://www.uniprot.org/help/downloads) is a comprehensive repository of protein sequence and functional information, offering extensive coverage of protein diversity. It provides detailed annotations for proteins, including information on their functions, structures, and roles in biological processes.

STRING Database: Protein-protein interactions that are known or expected can be found in the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (https://string-db.org). The vast and trustworthy dataset is ensured by the fact that these interactions are derived from multiple sources, such as public text collections, computer prediction techniques, and experimental data.

SwissProt: A curated protein sequence database that provides a high level of annotation, including information on protein function, domain structure, and post-translational modifications. SwissProt (https://www.ebi.ac.uk/GOA) focuses on experimentally validated data, making it a gold standard for functional annotation.

3.4.2 Data Collection Methods and Tools:

The chosen datasets are highly suitable for this research due to their comprehensive coverage[36], [53], [54], [55], high-quality interaction data, and reliable functional annotations. UniProt provides an extensive array of protein sequences, ensuring broad representation across various species and functions. This diversity is crucial for capturing the wide range of biological activities that proteins can perform. The STRING database is renowned for its detailed and high-confidence interaction data, which is essential for constructing accurate protein-protein interaction networks. These networks are a fundamental component of the research, as they help to elucidate the complex interactions that dictate protein function. Additionally, SwissProt offers meticulously curated and experimentally validated annotations, providing a solid and reliable foundation for training and evaluating the prediction model. The combination of these datasets ensures that the research is grounded in high-quality, diverse, and trustworthy data, which is vital for developing a robust and accurate protein function prediction model.

The research adheres to ethical standards by utilizing publicly available datasets, ensuring compliance with data usage policies and avoiding issues related to data privacy and consent[53], [54], [55]. Proper attribution is given to all data sources, acknowledging the original contributors and maintaining academic integrity. Furthermore, the datasets used do not contain any personal or sensitive information, thereby minimizing ethical concerns related to data handling. This ethical approach

not only ensures the integrity of the research process but also aligns with best practices for using publicly accessible data in scientific research. By addressing these ethical considerations, the research maintains transparency and respect for the data providers and the broader scientific community.

3.4.3 Software and hardware

The code for this study was written using the PyTorch and DGL (Deep Graph Library) frameworks due to their robust capabilities and flexibility. PyTorch is renowned for its dynamic computation graph, which simplifies debugging and allows for easy modifications during model development. This is particularly beneficial in research settings where models often need to be iteratively refined. Additionally, PyTorch's extensive support for neural network components, including pre-built modules, loss functions, and optimizers, streamlines the creation of complex models such as those used in protein function prediction.

DGL complements PyTorch by providing specialized tools for handling graph-based data, essential for implementing Graph Neural Networks (GNNs). Protein-protein interaction networks can be naturally represented as graphs, and DGL's optimized graph operations ensure efficient processing of these structures. Together, PyTorch and DGL offer a high-performance, scalable solution that leverages GPU acceleration for handling large datasets, a critical requirement for this study. Moreover, the strong community support and rich ecosystems of both frameworks facilitate collaboration and the integration of existing research, enhancing the overall productivity and impact of the study.

The training and experiments for the models in this study were conducted on a Linux server equipped with 32GB of RAM and an NVIDIA 1080TI GPU with 12GB of VRAM.

3.5 CHAPTER SUMMARY

In this chapter, the research methodology for developing a novel protein function prediction model was thoroughly outlined. The chapter began by detailing the research philosophy of positivism and the deductive approach employed. This structured methodology ensures objectivity and reliability, focusing on quantifiable data from protein sequences, protein-protein interaction networks, and InterPro domains.

The chapter proceeded to describe the selection and justification of research materials, including high-quality datasets from UniProt, STRING, and SwissProt. These datasets provide comprehensive coverage and reliable annotations crucial for accurate model training and evaluation. The choice of PyTorch and DGL frameworks was justified by their robust capabilities in handling neural networks and graph-based data, respectively. The computational experiments were conducted on a powerful Linux server equipped with 32GB of RAM and an NVIDIA 1080TI GPU, ensuring efficient processing.

This chapter lays the groundwork for the upcoming design and implementation of the prediction model in Chapter 4, providing a strong foundation of meticulous approach and superior materials. By combining graph convolutional networks with protein language models, the area of protein function prediction should become more accurate and efficient, filling in some of its current shortcomings. This process will be expanded upon in the next chapter, which will include specifics on the creation, application, examination, and assessment of the suggested model.

CHAPTER 4 – EXPERIMENT AND RESULT ANALYSIS

This chapter explores the experimental design, the outcomes of using the suggested model, and a thorough examination of these outcomes. The experimental setup, results and analysis, discussion, and summary make up the chapter's four sections. This framework guarantees a thorough comprehension of the study's performance, methods, and consequences.

4.1 EXPERIMENTAL SETUP

In this chapter the details of the experiments conducted to obtain the study results are provided. The experimental setup integrates multiple high-quality datasets and leverages advanced computational frameworks to predict protein functions using a novel model that combines protein language models (PLMs) and Graph Convolutional Networks (GCNs).

And the datasets used in this experiment underwent the following preprocessing steps.

- Protein Sequences: Data is extracted from UniProt in FASTA format[53]. Automated scripts are used to download sequences, ensuring that the dataset is comprehensive and up-to-date.
- PPI Network Data: Interactions from the STRING database are filtered to retain only the top 100 interactions for each protein based on interaction strength. This filtering ensures that the data includes the most biologically relevant interactions[54].
- GO Terms: Gene Ontology annotations are sourced from SwissProt[55], focusing on experimental evidence codes such as 'IDA' (Inferred from Direct Assay), 'IMP' (Inferred from Mutant Phenotype), and others. This ensures that the functional annotations used are reliable and validated.

The model was trained using a two-layer Graph Convolutional Network (GCN).Table 2 shows the parameters setting for the model training.

Table 4. 2:Parameters setting

Input size	Hidden size	Drop out rate	Epochs	Batch size	Optimizer	Learning rate	Activation Function
1280	512	0.5	20	8	Adam	1e-3	Sigmoid

The choice of the Adam optimizer was due to its adaptive learning rate capabilities[65], which efficiently handle sparse gradients and noisy problems. The Sigmoid function was selected to suit the multi-binary classification requirement[66], ensuring that the model could effectively distinguish between the presence and absence of multiple features.

4.2 RESULTS AND ANALYSIS

The evaluation of the model's performance was based on Fmax and AUPR[47], [48], [49], the most common metrics in the protein function prediction area[29]. The scores for these metrics, corresponding to the predictions made by the model in this study for the Molecular Function Ontology (MFO), Biological Process Ontology (BPO), and Cellular Component Ontology (CCO), are presented in a tabular format. A comparison was conducted with BLAST-KNN—,LR-InterPro and Net-KNN proposed by R. You et al.[67], DeepGO[24], DeepGOPlus[6], and

DeepGraphGO[29]. The data ranked first in individual performance is highlighted in bold.

Table 4. 3:Performance comparison of Proposed Model

Method	Fmax			AUPR		
	MFO	**BPO**	**CCO**	**MFO**	**BPO**	**CCO**
BLAST-KNN[67]	0.590	0.274	0.650	0.455	0.113	0.570
LR-InterPro[67]	0.617	0.278	0.661	0.530	0.133	0.672
Net-KNN[67]	0.426	0.305	0.667	0.276	0.157	0.641
DeepGO[24]	0.434	0.248	0.632	0.306	0.101	0.573
DeepGOPlus[6]	0.593	0.290	0.672	0.398	0.108	0.595
DeepGraphGO[29]	**0.623**	0.290	0.672	**0.543**	0.194	0.695
ProposedModel	0.531	**0.336**	**0.686**	0.456	**0.212**	**0.705**

Based on the performance metrics provided in Table 4.2 and Figure 4.1, DeepGraphGO outperforms other models in Molecular Function Ontology (MFO) prediction, achieving the highest Fmax (0.623) and AUPR (0.543) scores. This superior performance can be attributed to its ability to effectively leverage graph convolutional networks (GCNs) to integrate both sequence information and protein interaction data. The multiple layers of GCNs in DeepGraphGO allow for capturing complex, high-order relationships within protein networks, which are particularly critical for accurately predicting molecular functions[29]. Additionally, the inclusion of InterProScan features that identify specific protein domains and motifs enhances the model's ability to capture the fine-grained biochemical properties necessary for MFO predictions. These capabilities make DeepGraphGO particularly adept at handling the intricate dependencies and specificities involved in molecular function prediction, leading to its strong performance in this area.

Chart 1:Performance

In the BPO category, the proposed model achieves the highest Fmax score (0.686) and AUPR score (0.705), outperforming all other models. This strong performance can be attributed to the model's ability to effectively integrate diverse features from protein sequences, PPI networks**[54]**, and InterPro domains. The combination of these features allows the model to capture the complex interactions and pathways that are essential for predicting biological processes. The use of ESM-1b for extracting high-dimensional sequence features, coupled with the graph convolutional network (GCN) layers, enables the model to learn and generalize the

intricate dependencies between proteins within a biological process. This multi-source integration is crucial for BPO predictions[15], where understanding the functional interactions between proteins is key to accurately identifying the underlying biological processes.

In the CCO category, the proposed model again demonstrates superior performance with the highest Fmax score (0.705) and AUPR score (0.705). The model's success in CCO predictions can be attributed to its effective utilization of PPI network data, which is critical for understanding the spatial distribution and localization of proteins within the cell. By leveraging the GCN layers, the model is able to capture the complex network context that influences cellular localization. The integration of protein sequence features with network-based information allows the model to accurately predict the cellular components where proteins are likely to function[68]. This approach is particularly effective for CCO predictions, as it combines structural information with network interactions, providing a comprehensive understanding of protein localization within cellular structures.

4.3 DISCUSSION

The results of this study, as shown in Tables 4.1 and 4.2, reveal that the proposed model's performance varied significantly across different ontologies. This section explores potential reasons behind these performance differences, supported by relevant literature, and provides a detailed analysis of the model's strengths and weaknesses.

- Model Adaptability

  Traditional models like BLAST-KNN and LR-InterPro have long relied on sequence homology for predicting protein functions. These models utilize sequence alignment and similarity scoring techniques, which are particularly effective for tasks where sequence information alone is a strong predictor of function, such as in Molecular Function Ontology (MFO) predictions. For example, studies by Pearson et al.[68] and Altschul et al. [69]emphasize that

sequence-based approaches often excel in identifying proteins with similar functions when the sequences are sufficiently homologous .

In contrast, the proposed model integrates diverse features from multiple sources, including protein sequences, PPI networks, and InterPro domains. While this comprehensive integration offers a richer context for function prediction, it may also introduce complexity and noise, potentially making it less effective for MFO predictions, where simple sequence similarity might suffice. The added complexity could obscure specific sequence-based signals that are critical for accurate MFO predictions, as suggested by Leung et al.[70], who highlighted that over-complicating models with multi-source data can sometimes dilute the effectiveness of straightforward predictions .

- Data Utilization

  The proposed model leverages ESM-1b, a pre-trained protein language model, for feature extraction, capturing intricate details of protein sequences. ESM-1b has been shown to provide a powerful representation of protein sequences, capturing not just the sequence information but also structural and functional properties that are important for understanding protein behavior (Rives et al.[11]) . This comprehensive feature representation likely contributes to the model's strong performance in Biological Process Ontology (BPO) and Cellular Component Ontology (CCO) predictions, where understanding complex interactions and structural contexts is crucial.

  For BPO predictions, the integration of PPI network data enhances the model's ability to capture interactions and pathways involving multiple proteins, which are key to understanding biological processes. Similarly, for CCO, the model's effectiveness in capturing cellular localization likely benefits from the detailed network context provided by the PPI data. This aligns with the findings of Barabási and Oltvai[71], who noted that biological networks are integral to understanding cellular functions and localizations .

- Integration of Multi-Source Features

The integration of multi-source protein features, including sequence data, interaction networks, and domain information, likely played a significant role in the proposed model's superior performance in BPO and CCO. By combining these diverse data types, the model can develop a holistic view of protein function, which is particularly beneficial for complex tasks involving multiple interacting components. This approach is supported by previous research, such as that by Zhang et al. [39], which demonstrated that multi-source integration can significantly improve the accuracy of complex biological predictions by providing a more comprehensive context .

However, this approach may not be as effective for MFO predictions, which often depend more directly on specific sequence motifs or active sites. The added complexity from integrating multiple data sources might dilute the impact of these critical sequence-specific features, potentially leading to less accurate predictions. This potential downside is echoed in the work of Almagro Armenteros et al.[72], who suggested that while multi-source integration can enhance performance, it must be carefully managed to avoid introducing noise .

- Graph Convolutional Networks

Graph Convolutional Networks (GCNs) are particularly effective at capturing relationships in graph-structured data, such as PPI networks. In the proposed model, GCN layers effectively aggregate information from neighboring nodes in the PPI network, allowing the model to learn about the broader network context of each protein. This capability is especially useful for BPO and CCO predictions, where the network context can provide critical insights into biological processes and cellular localization. Kipf and Welling [34] demonstrated that GCNs are highly effective for tasks involving network data, supporting the findings of this study .

However, the effectiveness of GCNs might be less pronounced in MFO, where functional relationships may not be as graph-dependent and might rely more on localized sequence features. This limitation is consistent with the observations of Shervashidze et al.[43], who noted that while GCNs are powerful for graph-

structured data, they may not always capture more localized, sequence-based information as effectively as other models .

- Challenges in MFO Prediction

  Molecular Function Ontology (MFO) predictions often involve predicting specific biochemical activities of proteins, such as enzymatic functions, binding affinities, or catalytic roles. These functions are typically determined by specific amino acid residues and motifs within the protein sequence. The proposed model's reliance on broader network and domain features might not capture these specific sequence motifs as effectively as models focused solely on sequence alignment and homology. As discussed by Jones et al.[73], sequence-specific models are particularly well-suited for identifying these types of functions .

  Additionally, the complexity of integrating diverse features could introduce noise that impacts the model's ability to make accurate MFO predictions. This challenge is highlighted in the work of Zhou et al. [32], who pointed out that while multi-source integration can provide a more complete picture, it also risks introducing irrelevant information that can degrade performance for tasks requiring highly specific feature recognition .

- Potential Improvements

  To enhance the proposed model's performance in MFO predictions, future work could focus on refining the feature integration process to minimize noise and improve the model's ability to capture specific sequence motifs. Techniques such as feature selection and weighting could be optimized to prioritize the most relevant features for MFO tasks. Additionally, incorporating more sophisticated techniques for feature extraction and selection, such as attention mechanisms or hierarchical models, could help improve the model's adaptability to different ontologies. Vaswani et al.[57] demonstrated that attention mechanisms can significantly enhance model performance by allowing the model to focus on the most important parts of the input data, suggesting a promising direction for future work .

Overall, while the proposed model shows strong performance in BPO and CCO predictions, there are clear areas for improvement, particularly in MFO predictions. By addressing these challenges and incorporating advanced techniques, the model's effectiveness and applicability could be further enhanced, contributing to more accurate and comprehensive protein function predictions in future research.

4.4 SUMMARY

This chapter has outlined the experimental setup, results, and analysis of the proposed model for protein function prediction. The proposed model demonstrated superior performance in BPO and CCO, achieving the highest Fmax and AUPR scores, but underperformed in MFO predictions compared to other models. The analysis highlights the strengths of integrating multi-source features and the adaptability of GCNs for specific ontologies while pointing out areas needing improvement for better MFO predictions. This detailed discussion underscores the importance of continued research into optimizing computational prediction models and suggests specific areas for enhancement in future work , such as refining feature integration and improving the capture of specific sequence motifs.

CHAPTER 5 – SUMMARY CONCLUSION AND RECOMMENDATIONS

5.1 SUMMARY

The primary aim of this study was to develop a novel protein function prediction model by integrating data from protein sequences, protein-protein interaction (PPI) networks, and InterPro domains using a Protein Language Model (PLM) and Graph Convolutional Network (GCN) to generate accurate predictions of protein functions. The main objectives of this study were achieved through the following steps:

The first objective was to generate embeddings from protein sequences using the pre-trained protein language model (ESM-1b) for feature extraction. The approach involved utilizing ESM-1b to process and extract high-dimensional feature representations from protein sequences. This method successfully generated detailed embeddings that capture the complex biochemical properties of proteins, forming the foundational features necessary for accurate function prediction.

The second objective focused on integrating the embeddings from protein sequences and InterPro domains with adaptive feature weights into the PPI graph and using GCNs to generate protein features. By combining the extracted embeddings with PPI network data and InterPro domain features, and applying GCNs for feature integration, a comprehensive feature set was created. This integration significantly enhanced prediction accuracy by leveraging multiple sources of protein information.

The third objective was to develop a classification model that combines the feature weights and protein feature vectors generated by PLM, PPI, and GCNs. A GCN-based model was designed and implemented to process the integrated feature vectors and predict protein functions. This robust model demonstrated its capability to predict protein functions across different Gene Ontology (GO) categories effectively.

The fourth objective involved evaluating and comparing the performance of the developed model against existing state-of-the-art methods using well-known evaluation metrics. Extensive testing and evaluation were conducted using metrics

such as Fmax and AUPR. The proposed model demonstrated superior performance in Biological Process Ontology (BPO) and Cellular Component Ontology (CCO) predictions, validating the effectiveness of the proposed integration and methodology.

5.2 CONCLUSION

The proposed model utilized an integrated approach combining ESM-1b embeddings, PPI networks, and GCNs to predict protein functions. The methodology involved extracting detailed protein features using ESM-1b, which captures high-dimensional representations of protein sequences, followed by integrating these features through Graph Convolutional Networks (GCNs). This integration process enabled the model to leverage multiple sources of information, such as sequence data, protein-protein interactions, and domain-specific features, to enhance the overall prediction accuracy.

The outcomes of this integrated approach were promising, with the proposed model achieving an Fmax of 0.531 for MFO, 0.336 for BPO, and 0.686 for CCO, along with AUPR scores of 0.456 for MFO, 0.212 for BPO, and 0.705 for CCO. The proposed model excelled in predicting Biological Process Ontology (BPO) and Cellular Component Ontology (CCO), achieving the highest Fmax scores of 0.686 for CCO and 0.336 for BPO, and the highest AUPR scores of 0.705 for CCO and 0.212 for BPO, among the compared models. This indicates that the model's ability to incorporate diverse protein features and contextual information from PPI networks effectively captures the complex relationships and dependencies inherent in biological processes and cellular components. However, the model underperformed in Molecular Function Ontology (MFO) predictions. This underperformance highlights areas for potential improvement, suggesting that the model may need further refinement to better capture specific sequence motifs and biochemical properties relevant to molecular functions.

5.3 LIMITATION AND RECOMMENDATION

The proposed model exhibited variability in performance across different GO categories, notably underperforming in MFO predictions. Integrating multi-source features introduced complexity and potential noise, which may have affected the model's accuracy for specific tasks. Additionally, the model required significant computational resources for training and evaluation, which could limit its scalability and accessibility. Moreover, the model was implemented using an older version of the DGL framework (0.4.3post2), and the training was performed on an Nvidia 1080TI GPU, which is not the latest hardware. This resulted in slower training speeds and necessitated a smaller batch size.

Future research should focus on refining the feature integration process to reduce noise and enhance the model's ability to capture specific sequence motifs, particularly for MFO predictions. This could involve optimizing the feature selection and weighting mechanisms to ensure that the most relevant features are prioritized. Incorporating sophisticated techniques such as attention mechanisms or hierarchical models could improve the model's adaptability and performance across all GO categories. Attention mechanisms, for example, could help the model focus on the most critical parts of the input data, while hierarchical models could better capture the multi-level relationships between different protein features.

Developing more efficient training algorithms and exploring distributed computing approaches can help scale the model for larger datasets and broader applications. Implementing parallel processing and leveraging cloud-based platforms could enhance the model's scalability and make it more accessible for researchers with limited computational resources. Utilizing more diverse and extensive datasets could further improve the model's generalizability and robustness, providing a more comprehensive tool for protein function prediction. This includes incorporating additional sources of protein interaction data, exploring different types of protein annotations, and expanding the model's training on a wider variety of protein families and species.

Additionally, updating the implementation to use the latest version of the DGL framework and employing more advanced GPU hardware could significantly

improve training efficiency and model performance. This would allow for larger batch sizes and faster training times, potentially leading to better optimization and more accurate predictions.

By addressing these limitations and pursuing the proposed future work, the model's performance and applicability can be significantly enhanced, contributing to advancements in bioinformatics and protein function prediction.

CHAPTER6 – REFLECTION

In managing the work, I adopted a methodical approach by breaking down the project into smaller tasks with specific deadlines given by the supervisor. This allowed me to monitor progress effectively and ensure that each stage of the project was completed on time. Regular reviews and adjustments to the schedule ensured that I met all deadlines without compromising the quality of the work. The use of project management tools alongside AI assistance was key to maintaining this balance. I extensively used AI tools, particularly ChatGPT, which played a crucial role in enhancing the quality of my writing and refining complex ideas. ChatGPT assisted in organizing thoughts, structuring content, and improving the clarity and coherence of the research paper. This tool was invaluable in providing quick insights, generating suggestions for improvement, and ensuring that the writing was both technically accurate and accessible. Through this process, I not only improved my technical writing skills but also learned how to leverage AI tools to enhance productivity. This experience has taught me valuable lessons in time management, the importance of iterative refinement, and how to efficiently integrate AI tools into complex tasks to achieve the best outcomes.

Additionally, this project provided a significant opportunity to develop and enhance my technical skills, particularly in machine learning and bioinformatics. Learning to implement and fine-tune complex models like Graph Convolutional Networks (GCNs) and integrating them with pre-trained models such as ESM-1b was both challenging and rewarding. Conducting experiments on a Linux server equipped with an NVIDIA 1080TI GPU presented challenges in processing speed and memory limitations, necessitating smaller batch sizes and more efficient data processing techniques, which deepened my understanding of hardware-software interactions in high-performance computing environments.

REFERENCES

[1] R. F. Weaver, *Molecular Biology*. McGraw-Hill Education, 2011. [Online]. Available: https://books.google.co.uk/books?id=LawRtAEACAAJ

[2] M. Ashburner *et al.*, 'Gene Ontology: tool for the unification of biology', *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.

[3] R. Bonetta and G. Valentino, 'Machine learning techniques for protein function prediction', *Proteins Struct. Funct. Bioinforma.*, vol. 88, no. 3, pp. 397–413, 2020.

[4] M. Li, W. Shi, F. Zhang, M. Zeng, and Y. Li, 'A Deep Learning Framework for Predicting Protein Functions With Co-Occurrence of GO Terms', *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 20, no. 2, pp. 833–842, 2023, doi: 10.1109/TCBB.2022.3170719.

[5] P. Sun *et al.*, 'Protein Function Prediction Using Function Associations in Protein–Protein Interaction Network', *IEEE Access*, vol. 6, pp. 30892–30902, 2018, doi: 10.1109/ACCESS.2018.2806478.

[6] M. Kulmanov and R. Hoehndorf, 'DeepGOPlus: improved protein function prediction from sequence', *Bioinformatics*, vol. 36, no. 2, pp. 422–429, 2020.

[7] Z. Du, Y. He, J. Li, and V. N. Uversky, 'Deepadd: protein function prediction from k-mer embedding and additional features', *Comput. Biol. Chem.*, vol. 89, p. 107379, 2020.

[8] V. Gligorijević *et al.*, 'Structure-based protein function prediction using graph convolutional networks', *Nat. Commun.*, vol. 12, no. 1, p. 3168, May 2021, doi: 10.1038/s41467-021-23303-9.

[9] V. Gligorijević *et al.*, 'Structure-based protein function prediction using graph convolutional networks', *Nat. Commun.*, vol. 12, no. 1, p. 3168, 2021.

[10] R. You, S. Yao, H. Mamitsuka, and S. Zhu, ‘DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction’, *Bioinformatics*, vol. 37, no. Supplement_1, pp. i262–i271, Jul. 2021, doi: 10.1093/bioinformatics/btab270.

[11] A. Rives *et al.*, ‘Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences’, *Proc. Natl. Acad. Sci.*, vol. 118, no. 15, p. e2016239118, 2021.

[12] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, ‘Unified rational protein engineering with sequence-based deep representation learning’, *Nat. Methods*, vol. 16, no. 12, pp. 1315–1322, 2019.

[13] A. Elnaggar *et al.*, ‘Prottrans: Toward understanding the language of life through self-supervised learning’, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7112–7127, 2021.

[14] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, ‘ProteinBERT: a universal deep-learning model of protein sequence and function’, *Bioinformatics*, vol. 38, no. 8, pp. 2102–2110, 2022.

[15] C. L. Mills, P. J. Beuning, and M. J. Ondrechen, ‘Biochemical functional predictions for protein structures of unknown or uncertain function’, *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 182–191, 2015.

[16] K. K. Yang, Z. Wu, and F. H. Arnold, ‘Machine-learning-guided directed evolution for protein engineering’, *Nat. Methods*, vol. 16, no. 8, pp. 687–694, 2019.

[17] T. Bepler and B. Berger, ‘Learning the protein language: Evolution, structure, and function’, *Cell Syst.*, vol. 12, no. 6, pp. 654–669, 2021.

[18] A. Madani *et al.*, ‘Large language models generate functional protein sequences across diverse families’, *Nat. Biotechnol.*, vol. 41, no. 8, pp. 1099–1106, 2023.

[19] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, ‘Deep generative models of genetic variation capture the effects of mutations’, *Nat. Methods*, vol. 15, no. 10, pp. 816–822, 2018.

[20] J. Jumper *et al.*, ‘Highly accurate protein structure prediction with AlphaFold’, *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[21] S. J. Giri, P. Dutta, P. Halani, and S. Saha, ‘MultiPredGO: deep multi-modal protein function prediction by amalgamating protein structure, sequence, and interaction information’, *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1832–1838, 2020.

[22] R. You, X. Huang, and S. Zhu, ‘DeepText2GO: improving large-scale protein function prediction with deep semantic text representation’, *Methods*, vol. 145, pp. 82–90, 2018.

[23] H. Lee, Z. Tu, M. Deng, F. Sun, and T. Chen, ‘Diffusion kernel-based logistic regression models for protein function prediction’, *Omics J. Integr. Biol.*, vol. 10, no. 1, pp. 40–55, 2006.

[24] M. Kulmanov, M. A. Khan, and R. Hoehndorf, ‘DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier’, *Bioinformatics*, vol. 34, no. 4, pp. 660–668, 2018.

[25] V. N. Ioannidis, A. G. Marques, and G. B. Giannakis, ‘Graph neural networks for predicting protein functions’, in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, IEEE, 2019, pp. 221–225.

[26] C. Zhao, T. Liu, and Z. Wang, ‘PANDA2: protein function prediction using graph neural networks’, *NAR Genomics Bioinforma.*, vol. 4, no. 1, p. lqac004, 2022.

[27] S. Wang, H. Tang, P. Shan, Z. Wu, and L. Zuo, ‘ProS-GNN: predicting effects of mutations on protein stability using graph neural networks’, *Comput. Biol. Chem.*, vol. 107, p. 107952, 2023.

[28] K. Choi, Y. Lee, C. Kim, and M. Yoon, ‘An effective GCN-based hierarchical multi-label classification for protein function prediction’, *ArXiv Prepr. ArXiv211202810*, 2021.

[29] R. You, S. Yao, H. Mamitsuka, and S. Zhu, ‘DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction’, *Bioinformatics*, vol. 37, no. Supplement_1, pp. i262–i271, 2021.

[30] A. Villegas-Morcillo, S. Makrodimitris, R. C. van Ham, A. M. Gomez, V. Sanchez, and M. J. Reinders, ‘Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function’, *Bioinformatics*, vol. 37, no. 2, pp. 162–170, 2021.

[31] J. Chen, Z. Gu, L. Lai, and J. Pei, ‘In silico protein function prediction: the rise of machine learning-based approaches’, *Med. Rev.*, vol. 3, no. 6, pp. 487–510, 2023.

[32] N. Zhou *et al.*, ‘The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens’, *Genome Biol.*, vol. 20, pp. 1–23, 2019.

[33] L. Alzubaidi *et al.*, ‘Review of deep learning: concepts, CNN architectures, challenges, applications, future directions’, *J. Big Data*, vol. 8, pp. 1–74, 2021.

[34] T. N. Kipf and M. Welling, ‘Semi-supervised classification with graph convolutional networks’, *ArXiv Prepr. ArXiv160902907*, 2016.

[35] M. Ashburner *et al.*, ‘Gene ontology: tool for the unification of biology’, *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, 2000.

[36] G. O. Consortium, ‘The Gene Ontology (GO) database and informatics resource’, *Nucleic Acids Res.*, vol. 32, no. suppl_1, pp. D258–D261, 2004.

[37] G. O. Consortium, ‘The gene ontology resource: 20 years and still GOing strong’, *Nucleic Acids Res.*, vol. 47, no. D1, pp. D330–D338, 2019.

[38] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, ‘The graph neural network model’, *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, 2008.

[39] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, ‘Graph convolutional networks: a comprehensive review’, *Comput. Soc. Netw.*, vol. 6, no. 1, pp. 1–23, 2019.

[40] T. N. Kipf and M. Welling, ‘Semi-supervised classification with graph convolutional networks’, *ArXiv Prepr. ArXiv160902907*, 2016.

[41] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, ‘Graph attention networks’, *ArXiv Prepr. ArXiv171010903*, 2017.

[42] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, ‘How powerful are graph neural networks?’, *ArXiv Prepr. ArXiv181000826*, 2018.

[43] N. Shervashidze, P. Schweitzer, E. J. Van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, ‘Weisfeiler-lehman graph kernels.’, *J. Mach. Learn. Res.*, vol. 12, no. 9, 2011.

[44] Y. S. Park, L. Konge, and A. R. Artino Jr, ‘The positivism paradigm of research’, *Acad. Med.*, vol. 95, no. 5, pp. 690–694, 2020.

[45] D. R. Thomas, ‘A general inductive approach for analyzing qualitative evaluation data’, *Am. J. Eval.*, vol. 27, no. 2, pp. 237–246, 2006.

[46] J. W. Creswell and C. N. Poth, *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications, 2016.

[47] D. M. Powers, ‘Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation’, *ArXiv Prepr. ArXiv201016061*, 2020.

[48] T. Saito and M. Rehmsmeier, ‘The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets’, *PloS One*, vol. 10, no. 3, p. e0118432, 2015.

[49] J. Davis and M. Goadrich, ‘The relationship between Precision-Recall and ROC curves’, in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[50] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[51] T. D. Cook and D. T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Figures, 2007.

[52] M. S. Setia, ‘Methodology series module 3: Cross-sectional studies’, *Indian J. Dermatol.*, vol. 61, no. 3, pp. 261–264, 2016.

[53] ‘UniProt: the universal protein knowledgebase in 2023’, *Nucleic Acids Res.*, vol. 51, no. D1, pp. D523–D531, 2023.

[54] 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 Res.*, vol. 51, no. D1, pp. D638–D646, 2023.

[55] R. P. Huntley *et al.*, ‘The GOA database: gene ontology annotation updates for 2015’, *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1057–D1063, 2015.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘Bert: Pre-training of deep bidirectional transformers for language understanding’, *ArXiv Prepr. ArXiv181004805*, 2018.

[57] A. Vaswani *et al.*, ‘Attention is all you need’, *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[58] K. He, X. Zhang, S. Ren, and J. Sun, ‘Deep residual learning for image recognition’, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[59] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[60] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An introduction to statistical learning: With applications in python*. Springer Nature, 2023.

[61] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, 'Optimal thresholding of classifiers to maximize F1 measure', in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, Springer, 2014, pp. 225–239.

[62] G. Tsoumakas and I. Katakis, 'Multi-label classification: An overview international journal of data warehousing and mining', *Label Powerset Algorithm Call. PT3*, vol. 3, no. 3, 2006.

[63] M.-L. Zhang and Z.-H. Zhou, 'A review on multi-label learning algorithms', *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2013.

[64] R. Sharan, I. Ulitsky, and R. Shamir, 'Network-based prediction of protein function', *Mol. Syst. Biol.*, vol. 3, no. 1, p. 88, 2007.

[65] D. P. Kingma and J. Ba, 'Adam: A method for stochastic optimization', *ArXiv Prepr. ArXiv14126980*, 2014.

[66] H. Pratiwi *et al.*, 'Sigmoid activation function in selecting the best model of artificial neural networks', in *Journal of Physics: Conference Series*, IOP Publishing, 2020, p. 012010.

[67] R. You *et al.*, 'NetGO: improving large-scale protein function prediction with massive network information', *Nucleic Acids Res.*, vol. 47, no. W1, pp. W379–W387, 2019.

[68] W. R. Pearson, 'An introduction to sequence similarity ("homology") searching', *Curr. Protoc. Bioinforma.*, vol. 42, no. 1, pp. 3–1, 2013.

[69] S. F. Altschul, 'A protein alignment scoring system sensitive at all evolutionary distances', *J. Mol. Evol.*, vol. 36, pp. 290–300, 1993.

[70] H. W. Leung and J. Bovy, ‘Deep learning of multi-element abundances from high-resolution spectroscopic data’, *Mon. Not. R. Astron. Soc.*, vol. 483, no. 3, pp. 3255–3277, 2019.

[71] A.-L. Barabasi and Z. N. Oltvai, ‘Network biology: understanding the cell’s functional organization’, *Nat. Rev. Genet.*, vol. 5, no. 2, pp. 101–113, 2004.

[72] V. I. Jurtz *et al.*, ‘An introduction to deep learning on biological sequence data: examples and solutions’, *Bioinformatics*, vol. 33, no. 22, pp. 3685–3690, 2017.

[73] S. Lise and D. Jones, ‘Sequence patterns associated with disordered regions in proteins’, *PROTEINS Struct. Funct. Bioinforma.*, vol. 58, no. 1, pp. 144–150, 2005.

PROJECT MANAGEMENT

Effective project management was critical to the successful completion of this research project. To ensure that the project stayed on track, I employed a detailed plan that was reflected in two Gantt charts: the Original Gantt Chart and the Actual Gantt Chart showed in Figure 6 and Figure 7.

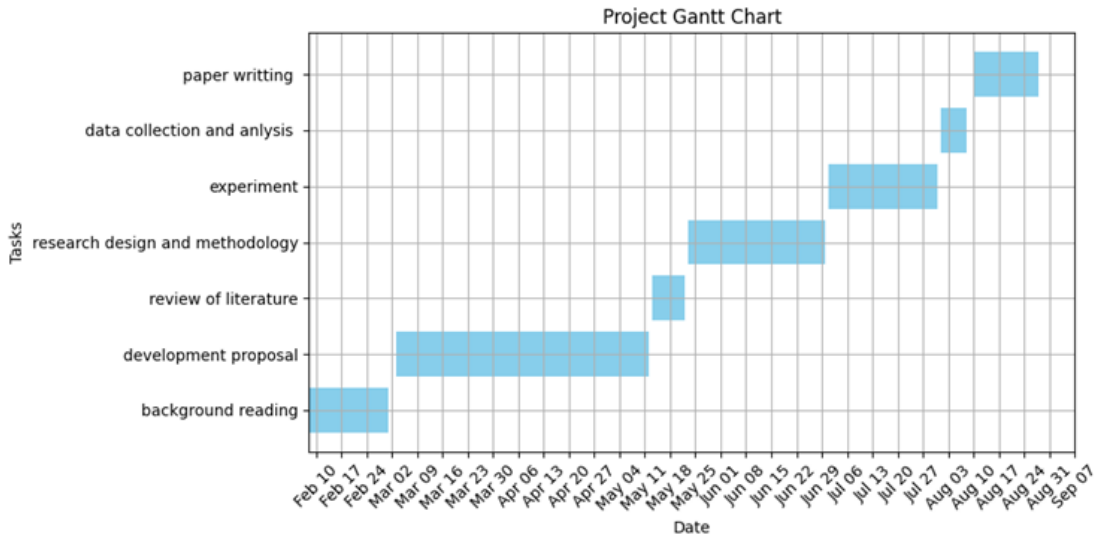

Figure 6:Original Gantt Chart

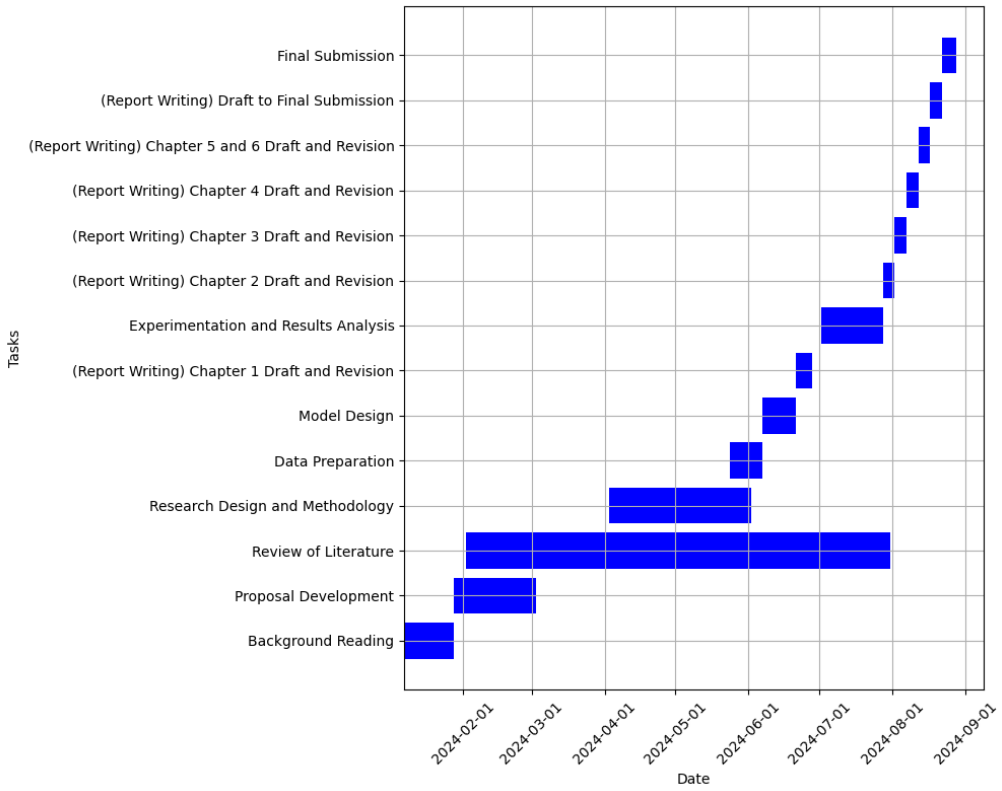

Figure 7：Actual Gantt Chart

The Original Gantt Chart laid out the initial timeline for the project, starting with the early stages of research, data collection, and literature review. It provided a structured framework for the project, with specific milestones set for each phase, including the development of the model, experimentation, analysis, and the writing of the thesis. As the project progressed, some adjustments were made, which are reflected in the Actual Gantt Chart. While the project generally followed the planned timeline, a few phases required more time than initially anticipated, particularly in the areas of model development and data analysis. These adjustments were necessary to address unexpected challenges, such as fine-tuning the model and interpreting complex data sets.

Overall, the project management strategy, as depicted in the Gantt charts, played a crucial role in organizing the workflow, meeting deadlines, and accommodating unforeseen challenges. The flexibility to adjust the schedule as needed ensured that the project objectives were met without compromising the quality of the research.

ETHICS FORM

ETHICS FORM – STEM MSc STUDENTS ONLY

APPLICATION FOR ETHICAL APPROVAL

In order for research to result in benefit and minimise risk of harm, it must be conducted ethically.

The University follows the OECD Frascati manual definition of **research activity**: "creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications". As such this covers activities undertaken by members of staff, postgraduate research students, and both taught postgraduate and undergraduate students working on dissertations/projects.

The individual undertaking the research activity is known as the "principal researcher".

This form must be completed and approved prior to undertaking any research activity.

SECTION A: About You (Principal Researcher)

1	Full Name:	Yuanhao Chen
2	Student Number:	2304723
3	Email address:	2304723@student.uwtsd.ac.uk
4	Programme of Study:	MSc Computing portfolio
5	Director of Studies/Supervisor:	Seena Joseph

SECTION B: Internal and External Ethical Guidance Materials

	Please list the core ethical guidance documents that have been referred to during the completion of this form (including any discipline-specific codes of research ethics, location-specific dodes of research ethics, and also any specific ethical guidance relating to the proposed methodology). Please tick to confirm that your research proposal adheres to these codes and guidelines. You may add rows to this table if needed.	
1	**UWTSD Research Ethics & Integrity Code of Practice**	☒
2	**UWTSD Research Data Management Policy**	☐
3		☐

SECTION C: Details of Research Activity

1	Indicative title:	Protein Function Prediction using Graph Neural Networks		
2	Proposed start date:	10th Feb	Proposed end date:	21st Aug
	Introduction to the Research (maximum 300 words in each section) **Ensure that you write for a <u>Non-Specialist Audience</u> when outlining your response to the three points below:** • *Purpose of Research Activity* • *Proposed Research Question* • *Aims of Research Activity* • *Objectives of Research Activity* Demonstrate, briefly, how **<u>Existing Research</u>** has informed the proposed activity and explain			

	• *What the research activity will add to the body of knowledge* • *How it addresses an area of importance.*
3	**Purpose of Research Activity** The primary goal of this research is to enhance the methods for predicting the functions of proteins using advanced machine learning techniques, specifically Graph Neural Networks (GNNs). Proteins are vital to numerous biological processes in organisms, and understanding their functions can lead to significant advancements in medical and biological sciences. However, as the number of discovered proteins increases rapidly, traditional experimental methods to determine protein functions are becoming inadequate due to their time-consuming and costly nature. This research seeks to address this gap by employing a novel computational approach. (this box should expand as you type)
4	**Research Question** This study aims to investigate whether integrating diverse data sources, such as protein-protein interaction networks and domain information, with cutting-edge graph neural network architectures can improve the accuracy and efficiency of protein function prediction compared to existing methods. (this box should expand as you type)
5	**Aims of Research Activity** The research aims to develop a more accurate and computationally efficient model for predicting protein functions. By integrating various types of biological data and utilizing the latest advancements in machine learning, this study seeks to overcome the limitations of current prediction methods and provide deeper insights into protein functions that are crucial for scientific advancements. (this box should expand as you type)
6	**Objectives of Research Activity** To implement a novel GNN architecture that leverages multiple data types for a comprehensive analysis. To improve the prediction accuracy of protein functions by utilizing sequence-derived features alongside structural and interaction data. To reduce computational costs through optimized data processing techniques suitable for large-scale biological data. To validate the effectiveness of the proposed model against established benchmarks, thereby setting new standards for protein function prediction in the field. (this box should expand as you type)
	Proposed data collection methods (maximum 600 words) Provide a brief summary of all the methods that **may** be used in the research activity to collect data, making it clear what specific techniques may be used. If methods other than those listed in this section are deemed appropriate later, additional ethical approval for those methods will be needed. You do not need to justify the methods here, but should instead describe how you intend to collect the data necessary for you to complete your project.
7	This should describe how you intend to collect data. It should not include a discussion of the theoretical basis for your data collection methods. Please note, that if you intend to collect any audio/video recordings of interviews with participants then these will be are classified as Personal Data under GDPR/DPA2018. If you intend to use these then note this in section H. Download the datasets from open source websites and open source github. (this box should expand as you type)

SECTION D: Scope of Research Activity

	Will the research activity include:	YES	NO
1	Use of a questionnaire or similar research instrument?	☐	☒

2	Use of interviews?	☐	☒
3	Use of focus groups?	☐	☒
4	Use of participant diaries?	☐	☒
5	Use of video or audio recording?	☐	☒
6	Use of computer-generated log files?	☐	☒
7	Participant observation with their knowledge?	☐	☒
8	Participant observation without their knowledge?	☐	☒
9	Access to personal or confidential information without the participants' specific consent?	☐	☒
10	Administration of any questions, test stimuli, presentation that may be experienced as physically, mentally or emotionally harmful / offensive?	☐	☒
11	Performance of any acts which may cause embarrassment or affect self-esteem?	☐	☒
12	Investigation of participants involved in illegal activities?	☐	☒
13	Use of procedures that involve deception?	☐	☒
14	Administration of any substance, agent or placebo?	☐	☒
15	Working with live vertebrate animals?	☐	☒
16	Procedures that may have a negative impact on the environment?	☐	☒
17	Other primary data collection methods. Please indicate the type of data collection method(s) below.		
	Details of any other primary data collection method: (this box should expand as you type)	☐	☒

If you have ticked NO to every question then the research activity is (ethically) low risk and you may skip section E and continue to section F.

If YES to any question, then no research activity should be undertaken until full ethical approval has been obtained.

SECTION E: Intended Participants

	Who are the intended participants:	**YES**	**NO**
1	Students or staff at the University?	☐	☐
2	Adults (over the age of 18 and competent to give consent)?	☐	☐
3	Vulnerable adults?	☐	☐
4	Children and Young People under the age of 18? (Consent from Parent, Carer or Guardian will be required)	☐	☐
5	Prisoners?	☐	☐
6	Young offenders?	☐	☐
7	Those who could be considered to have a particularly dependent relationship with the investigator or a gatekeeper?	☐	☐
8	People engaged in illegal activities?	☐	☐
9	Others. Please indicate the participants below, and specifically any group who may be unable to give consent.	☐	☐

	Details of any other participant groups: Complete this only if your participants cannot give consent. This includes animals (this box should expand as you type)		

	Participant numbers and source Provide an estimate of the expected number of participants. How will you identify participants and how will they be recruited?	
10	How many participants are expected?	Ballpark figures are fine, but make sure that you explain how you will identify and contact your participants. *(this box should expand as you type)*
11	Who will the participants be?	*(this box should expand as you type)*
12	How will you identify the participants?	*(this box should expand as you type)*

	Information for participants:	**YES**	**NO**	**N/A**
13	Will you describe the main research procedures to participants in advance, so that they are informed about what to expect?	☐	☐	☐
14	Will you tell participants that their participation is voluntary?	☐	☐	☐
15	Will you obtain written consent for participation?	☐	☐	☐
16	Will you explain to participants that refusal to participate in the research will not affect their treatment or education (if relevant)?	☐	☐	☐
17	If the research is observational, will you ask participants for their consent to being observed?	☐	☐	☐
18	Will you tell participants that they may withdraw from the research at any time and for any reason?	☐	☐	☐
19	With questionnaires, will you give participants the option of omitting questions they do not want to answer?	☐	☐	☐
20	Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs?	☐	☐	☐
21	Will you debrief participants at the end of their participation, in a way appropriate to the type of research undertaken?	☐	☐	☐
22	If NO to any of above questions, please give an explanation			
	You should be able to tick YES for all of these questions. If not, then explain why not in this box. *(this box should expand as you type)*			

	Information for participants:	**YES**	**NO**	**N/A**
24	Will participants be paid?	☐	☐	☐
25	Is specialist electrical or other equipment to be used with participants?	☐	☐	☐
26	Are there any financial or other interests to the investigator or University arising from this study?	☐	☐	☐
27	Will the research activity involve deliberately misleading participants in any way, or the partial or full concealment of the specific study aims?	☐	☐	☐

28	If YES to any question, please provide full details
	You should be able to tick NO for most of these questions. For any cases that you have ticked YES then provide details in this box. If you are using cameras/voice recorders to record interviews then please state that in this box. *(this box should expand as you type)*

SECTION F: Anticipated Risks

	Outline any anticipated risks that may adversely affect any of the participants, the researchers and/or the University, and the steps that will be taken to address them.	
1	**Risks to participants** For example: sector-specific health & safety, emotional distress, financial disclosure, physical harm, transfer of personal data, sensitive organisational information. If you have identified in section D that there are no participants then enter N/A and go skip to question 3.	
	Risk to participants: There are always risks. Do not write N/A unless you have no participants. N/A *(this box should expand as you type)*	*How you will mitigate the risk to participants:* N/A *(this box should expand as you type)*
2	If research activity may include sensitive, embarrassing or upsetting topics (e.g. sexual activity, drug use) or issues likely to disclose information requiring further action (e.g. criminal activity), give details of the procedures to deal with these issues, including any support/advice (e.g. helpline numbers) to be offered to participants. Note that where applicable, consent procedures should make it clear that if something potentially or actually illegal is discovered in the course of a project, it may need to be disclosed to the proper authorities	
	N/A *(this box should expand as you type)*	
3	**Risks to the investigator** For example: personal health & safety, physical harm, emotional distress, risk of accusation of harm/impropriety, conflict of interest	
	Risk to the investigator: There are always risks. Do not write NA. Emotional Distress: While not common in computational research, long hours, frustration over unexpected results, or challenges in problem-solving could lead to stress or burnout. *(this box should expand as you type)*	*How you will mitigate the risk to the investigator:* Manage workloads and maintain a balanced approach to research activities. *(this box should expand as you type)*
4	**University/institutional risks** For example: adverse publicity, financial loss, data protection	
	Risk to the University: There are always risks. Do not write NA. Failure to Meet Expectations: If the research does not meet the scientific community's expectations or if the results are misinterpreted by the public, it might lead to adverse publicity questioning the institution's research capabilities. *(this box should expand as you type)*	*How you will mitigate the risk to the University:* Transparent Communication: Maintain open and transparent communication channels with the public and research community to manage expectations and promptly correct any misinformation. *(this box should expand as you type)*
5	**Environmental risks** For example: accidental spillage of pollutants, damage to local ecosystems	
	Risk to the environment: You may write NA if there are no research-related environmental risks. Driving to the university does not count as a risk.	*How you will mitigate the risk to environment:* NA *(this box should expand as you type)*

	NA *(this box should expand as you type)*	

SECTION G: Feedback, Consent and Confidentiality

If you have identified in section D that there are no participants then enter skip this section and continue to section H.

1	**Feedback** What de-briefing and feedback will be provided to participants, how will this be done and when?
	You don't need to email your participants with your final report. A good alternative is to set up an email address that they will be able to contact for further details or results. *(this box should expand as you type)*
2	**Informed consent** Describe the arrangements to inform potential participants, before providing consent, of what is involved in participating. Describe the arrangements for participants to provide full consent before data collection begins. If gaining consent in this way is inappropriate, explain how consent will be obtained and recorded in accordance with prevailing data protection legislation.
	If you are using a paper questionnaire then you should have the participants sign an appropriate consent form. These forms will count as personal data and should be noted as such in section J. If you are using an online questionnaire, then you should have a screen before the questions start that acts as a consent form, informing participants that by clicking on the NEXT button they are providing consent. *(this box should expand as you type)*
3	**Confidentiality / Anonymity** Set out how anonymity of participants and confidentiality will be ensured in any outputs. If anonymity is not being offered, explain why this is the case.
	Do not collect names unless you really need them. Do not name participants or organisations in any research publications (including the thesis) without their explicit permission. *(this box should expand as you type)*

SECTION H: Data Protection and Storage

	Does the research activity involve personal data (as defined by the General Data Protection Regulation 2016 "GDPR" and the Data Protection Act 2018 "DPA")?	**YES**	**NO**
1	***"Personal data"*** *means any information relating to an identified or identifiable natural person ('data subject'). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. Any video or audio recordings of participants is considered to be personal data.*	☐	×
	If YES, provide a description of the data and explain why this data needs to be collected:		
2	This includes audio/video data of participants, but can also include IP addresses and usernames. Names, addresses and emails also count, as do consent forms. *(this box should expand as you type)*		
	Does it involve special category data (as defined by the GDPR)?	**YES**	**NO**
3	***"Special category data"*** *means sensitive personal data consisting of information as to the data subjects' –* *(a) racial or ethnic origin,* *(b) political opinions,* *(c) religious beliefs or other beliefs of a similar nature,*	☐	×

	(d) membership of a trade union (within the meaning of the Trade Union and Labour Relations (Consolidation) Act 1992), *(e) physical or mental health or condition,* *(f) sexual life,* *(g) genetics,* *(h) biometric data (as used for ID purposes),*		
	If YES, provide a description of the special category data and explain why this data needs to be collected:		
4	What counts as 'sensitive' will differ between cultures. Any information on behaviour that is not in accordance with cultural norms would count as sensitive personal data. *(this box should expand as you type)*		

	Will data from the research activity (collected data, drafts of the thesis, or materials for publication) be stored in any of the following ways?	**YES**	**NO**
5	Manual files (i.e. in paper form)?	☐	☒
6	University computers?	☐	☒
7	Private company computers?	☐	☒
8	Home or other personal computers?	☒	☐
9	Laptop computers/ CDs/ Portable disk-drives/ memory sticks?	☒	☐
10	"Cloud" storage or websites?	☐	☒
11	Other – specify:	☐	☒
12	For all stored data, explain the measures in place to ensure the security of the data collected, data confidentiality, including details of backup procedures, password protection, encryption, anonymisation and pseudonymisation:		
	If possible, save your data on computers that are secure and regularly backed up. Many cloud services only provide GDPR-compliant storage for business customers. An example of suitable text is given below. *All data will be kept in password protected cloud storage on the University Office 365 system which will not be shared. Audio/visual data will be transcribed and would be shown to participants to check accuracy of reporting. Any USB sticks used to store or transfer data will be password protected. All participants will be given a unique identifier to ensure confidentiality and this list will be kept securely in the password protected folder.* All data will be stored on a secure server with long-term backup, and users will need a password to access it. *(this box should expand as you type)*		

	Data Protection		
	Will the research activity involve any of the following activities:	**YES**	**NO**
13	Electronic transfer of data in any form?	☒	☐
14	Sharing of data with others at the University outside of the immediate research team?	☒	☐
15	Sharing of data with other organisations?	☒	☐
16	Export of data outside the UK or importing of data from outside the UK?	☒	☐
17	Use of personal addresses, postcodes, faxes, emails or telephone numbers?	☒	☐

18	Publication of data that might allow identification of individuals?	☐	☒
19	If YES to any question, please provide full details, explaining how this will be conducted in accordance with the GDPR and Data Protection Act (2018) (and/or any international equivalent):		
	This includes data such as drafts of your thesis as well as experimental or survey data. An example of suitable text is given below. *All data will be encrypted and kept in password protected cloud storage on the University Office 365 system which will not be shared. Any USB sticks used to store or transfer data will be password protected. All data transfers will be encrypted and password protected. All participants will be given a unique identifier to ensure confidentiality and this list will be kept securely in the password protected folder. The data will be stored until the completion of the project and then deleted. In accordance with the DPA2018, participants will have the right to ask to see what data is held relating to them, and this data will be deleted immediately if the participant requests this, in which case the data will not be used in the project.* All data is downloaded from open-source websites and will be uploaded to open-source websites, where it can be shared by all developers. *(this box should expand as you type)*		
20	List all who will have access to the data generated by the research activity:		
	Normally the principal researcher, possibly also the supervisor and, if the project has an industrial partner, a representative of that partner. Possibly also external examiner or second marker? All developers who want to access this data *(this box should expand as you type)*		
21	List who will have control of, and act as custodian(s) for, data generated by the research activity:		
	Usually the principal researcher. None *(this box should expand as you type)*		
22	Give details of data storage arrangements, including security measures in place to protect the data, where data will be stored, how long for, and in what form.		
	All data will be encrypted and kept in password protected cloud storage on the University Office 365 system which will not be shared. Any USB sticks used to store or transfer data will be password protected, and will be reformatted at the end of the project in order to destroy the data. The data will be stored until the completion of the project and then deleted. The open-source data is stored on servers used by open-source websites, accessible for everyone to download and use. The data I downloaded is stored on a local server. *(this box should expand as you type)*		
22	Confirm that you have read the UWTSD guidance on data management (see https://www.uwtsd.ac.uk/library/research-data-management/)	☒	
23	Confirm that you are aware that you need to keep all data until after your research has completed or the end of your funding	☒	

SECTION I: Declaration

	The information which I have provided is correct and complete to the best of my knowledge. I have attempted to identify any risks and issues related to the research activity and acknowledge my obligations and the rights of the participants. In submitting this application I hereby confirm that I undertake to ensure that the above named research activity will meet the University's Research Ethics and Integrity Code of Practice which is published on the website: https://www.uwtsd.ac.uk/research/research-ethics/		
1	**Signature of applicant:**	Yuanhao Chen	**Date:2024/5/9**
2	Director of Studies/Supervisor:	Seena Joseph	**Date:2024/5/9**

3	Signature:	Yuanhao Chen	

FOR INTERNAL USE ONLY:

	Ethical approval given		
1	**Signature of assessor:**		**Date:**
2	Name:		
3	Role:		

LOGBOOK

Date	Daily Activities	Thought Trails	Things to Do
2024-02-01	● Started background reading on protein function prediction using Graph Neural Networks (GNNs).	● Understanding the complexity of protein interactions and the role of GNNs in capturing these relationships.	● Identify key literature sources focusing on GNN applications in bioinformatics.
2024-02-03	● Continued exploring studies on integrating protein-protein interaction (PPI) networks with GNN models.	● Noted the importance of high-quality PPI data and its influence on GNN performance in function prediction.	● Search for datasets providing reliable PPI information.
2024-02-05	● Began Proposal Development. Initial discussions about the feasibility of using GNNs for protein function prediction.	● Considering the research aim, objectives, and feasibility of integrating multiple data types into a GNN framework.	● Draft the research proposal outline. Develop a preliminary timeline for the project phases.
2024-02-07	● Further refined the proposal, focusing on objectives and expected outcomes	● Exploring different GNN architectures, particularly those	● Expand the proposal to include a detailed literature review

	related to GNN-based models.	suitable for biological data integration.	section.
2024-02-10	● Continued literature review focusing on GNN-based models in protein function prediction.	● Identified gaps in current GNN applications, such as the need for models that better handle heterogeneous data.	● Collect more papers on deep learning applications in bioinformatics, specifically related to GNNs.
2024-02-12	● Started outlining potential modifications to existing GNN models to better suit protein function prediction tasks.	● The integration of PPI networks and sequence data might require custom GNN layers or modified architectures.	● Draft a section of the proposal on the proposed GNN modifications.
2024-02-15	● Detailed proposal drafted, focusing on using GNNs for protein function prediction.	● Exploring the feasibility of a GNN model that can process and integrate multiple biological data types effectively.	● Prepare a presentation for the proposal presentation. Plan for further deep dives into GNN model studies.
2024-03-01	● Completed background reading and final adjustments to the proposal.	● Finalizing the scope and methodology, ensuring the proposal aligns with the project's goals and resources.	● Submit the proposal and start preparing for literature review consolidation.

2024-03-10	● Deepened literature review with a focus on multi-source data integration techniques within GNN models.	● Emphasized the importance of effectively combining sequence data with PPI networks in function prediction.	● Finalize literature review draft and plan for the experimental setup.
2024-05-11	● Started research design and methodology, focusing on data collection and preprocessing for GNN training.	● Designed initial experiments and selected datasets for model training. Deciding on best data sources for PPI and sequences.	● Set up the experiment environment and organize data collection. Begin writing the research methodology chapter.
2024-06-01	● Completed data preparation and began initial GNN model training.	● Observed that the quality of preprocessed data will significantly impact the GNN's performance.	● Run initial training cycles to test GNN performance. Monitor and log training progress carefully.
2024-06-15	● Analyzed early experimental results and discussed model adjustments.	● Early results show promising accuracy but highlight the need for more robust handling of heterogeneous data.	● Continue experiments and explore regularization techniques to

			reduce overfitting.
2024-07-01	● Completed additional experiments and started integrating findings into the report.	● Results are now more consistent, and the model's accuracy is improving. Next focus on optimizing GNN performance.	● Begin drafting the experimentation section in the report. Prepare for the next phase of writing.
2024-07-09	● Completed Chapter 1 revisions and started outlining Chapter 2, focusing on the literature review.	● Strengthening the connection between reviewed studies and experimental findings is crucial for coherence.	● Draft Chapter 2, ensuring it ties literature review closely to the proposed GNN model.
2024-08-01	● Drafted Chapters 4 and 5, incorporating experiment results and discussions related to GNN performance.	● Early results indicate that the proposed GNN model outperforms existing techniques in accuracy, especially in integrating heterogeneous data.	● Prepare visuals and tables for the report to clearly illustrate the results.
2024-08-28	● Final report submission.	● Reflecting on the entire project process and outcomes. Preparing for future research directions based on this work.	● Finalize any loose ends and prepare for the viva. Start planning for

			future research directions.
2024-02-01	● Started background reading on protein function prediction using Graph Neural Networks (GNNs).	● Understanding the complexity of protein interactions and the role of GNNs in capturing these relationships.	● Identify key literature sources focusing on GNN applications in bioinformatics.
2024-02-03	● Continued exploring studies on integrating protein-protein interaction (PPI) networks with GNN models.	● Noted the importance of high-quality PPI data and its influence on GNN performance in function prediction.	● Search for datasets providing reliable PPI information.
2024-02-05	● Began Proposal Development. Initial discussions about the feasibility of using GNNs for protein function prediction.	● Considering the research aim, objectives, and feasibility of integrating multiple data types into a GNN framework.	● Draft the research proposal outline. Develop a preliminary timeline for the project phases.
2024-02-07	● Further refined the proposal, focusing on objectives and expected outcomes	● Exploring different GNN architectures, particularly those	● Expand the proposal to include a detailed

	related to GNN-based models.	suitable for biological data integration.	literature review section.
2024-02-10	● Continued literature review focusing on GNN-based models in protein function prediction.	● Identified gaps in current GNN applications, such as the need for models that better handle heterogeneous data.	● Collect more papers on deep learning applications in bioinformatics, specifically related to GNNs.
2024-02-12	● Started outlining potential modifications to existing GNN models to better suit protein function prediction tasks.	● The integration of PPI networks and sequence data might require custom GNN layers or modified architectures.	● Draft a section of the proposal on the proposed GNN modifications.
2024-02-15	● Detailed proposal drafted, focusing on using GNNs for protein function prediction.	● Exploring the feasibility of a GNN model that can process and integrate multiple biological data types effectively.	● Prepare a presentation for the proposal presentation. Plan for further deep dives into GNN model studies.
2024-03-01	● Completed background reading and final adjustments to the proposal.	● Finalizing the scope and methodology, ensuring the proposal	● Submit the proposal and start preparing for literature review consolidation.

		aligns with the project's goals and resources.	
2024-03-10	● Deepened literature review with a focus on multi-source data integration techniques within GNN models.	● Emphasized the importance of effectively combining sequence data with PPI networks in function prediction.	● Finalize literature review draft and plan for the experimental setup.
2024-05-11	● Started research design and methodology, focusing on data collection and preprocessing for GNN training.	● Designed initial experiments and selected datasets for model training. Deciding on best data sources for PPI and sequences.	● Set up the experiment environment and organize data collection. Begin writing the research methodology chapter.
2024-06-01	● Completed data preparation and began initial GNN model training.	● Observed that the quality of preprocessed data will significantly impact the GNN's performance.	● Run initial training cycles to test GNN performance. Monitor and log training progress carefully.

2024-06-15	• Analyzed early experimental results and discussed model adjustments.	• Early results show promising accuracy but highlight the need for more robust handling of heterogeneous data.	• Continue experiments and explore regularization techniques to reduce overfitting.
2024-07-01	• Completed additional experiments and started integrating findings into the report.	• Results are now more consistent, and the model's accuracy is improving. Next focus on optimizing GNN performance.	• Begin drafting the experimentation section in the report. Prepare for the next phase of writing.
2024-07-09	• Completed Chapter 1 revisions and started outlining Chapter 2, focusing on the literature review.	• Strengthening the connection between reviewed studies and experimental findings is crucial for coherence.	• Draft Chapter 2, ensuring it ties literature review closely to the proposed GNN model.
2024-08-01	• Drafted Chapters 4 and 5, incorporating experiment results and discussions related to GNN performance.	• Early results indicate that the proposed GNN model outperforms existing techniques in accuracy, especially in	• Prepare visuals and tables for the report to clearly illustrate the results.

		integrating heterogeneous data.	
2024-08-28	● Final report submission.	● Reflecting on the entire project process and outcomes. Preparing for future research directions based on this work.	● Finalize any loose ends and prepare for the viva. Start planning for future research directions.

GLOSSARY

GCN (Graph Convolutional Network): A particular kind of neural network intended for direct manipulation of graph-structured data. Protein-protein interaction networks and other diverse biological data sources are included in the protein function prediction process through the usage of GCNs.

GNN (Graph Neural Network): A type of neural network that processes data structured as graphs. GNNs are used to model the relationships between entities, such as proteins within a protein-protein interaction network.

GO (Gene Ontology): A bioinformatics project that attempts to harmonize the way different species represent the characteristics of genes and proteins. Three primary ontologies are used to classify GO terms: Molecular Function, Biological Process, and Cellular Component.

BPO (Biological Process Ontology): A category within the Gene Ontology framework that describes the biological processes, or sets of molecular events, in which proteins are involved.

MFO (Molecular Function Ontology): A Gene Ontology category that describes the molecular activities, such as catalytic or binding actions, performed by individual proteins.

CCO (Cellular Component Ontology): A Gene Ontology category that refers to the locations within the cell where proteins carry out their functions, such as the nucleus or cytoplasm.

ESM-1b (Evolutionary Scale Modeling-1b): A protein language model based on deep learning was created to parse protein sequences and extract functional and evolutionary characteristics that are essential for predicting the function of the protein.

DeepGraphGO: A model that combines information from interaction networks and protein sequences to predict the activities of proteins using graph neural networks.

PPI (Protein-Protein Interaction): The actual physical bonds formed by biochemical processes and/or electrostatic forces between two or more protein molecules. PPI networks are crucial for comprehending the intricate relationships that control biological processes.

InterPro: A database that combines various protein signature datasets to offer functional analysis of proteins through family classification and domain and significant site prediction.

Fmax: A bioinformatics metric that assesses the maximal F-measure over several thresholds, balancing recall and precision. In multi-label classification tasks such as protein function prediction, it is very helpful.

AUPR (Area Under the Precision-Recall Curve): A performance metric that is especially useful in situations where the datasets are unbalanced for assessing the quality of forecasts. It sheds light on how recall and precision are traded off at various threshold values.

Sigmoid Function: An activation function used in neural networks that outputs values between 0 and 1, often used in binary classification tasks.

UniProt: The Universal Protein Resource, a comprehensive database that provides a curated protein sequence and functional information.

STRING Database: A database including information on known and anticipated interactions between proteins, obtained from a variety of sources such as computational forecasts and experimental evidence.