Protein-Protein Interactions Prediction, Drug Repurposing and Molecular Docking for emerging viral threats

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Engineering
in
Computer Science and Engineering

Submitted by

Debarati Paul
Regn. No. - 163558 of 2022 - 2023
Exam Roll No. - M4CSE24004
Class Roll No. - 002210502007
Session - 2022 - 2024

Under the supervision of

Prof. Subhadip Basu

Department of Computer Science and Engineering
Jadavpur University
188, Raja S.C. Mallick Rd,
Kolkata - 700032, West Bengal, India

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University, Kolkata - 700032

CERTIFICATE OF RECOMMENDATION

This is to certify that the work embodied in this thesis entitled **"Protein-Protein Interactions Prediction, Drug Repurposing and Molecular Docking for emerging viral threats"** has been satisfactorily completed by **Debarati Paul** (Registration Number 163558 of 2022 - 2023; Class Roll No. 002210502007; Examination Roll No. M4CSE24004). It is a bonafide piece of work carried out under my supervision and guidance at Jadavpur University, Kolkata, for partial fulfillment of the requirements for the awarding of the **Master of Engineering** in **Computer Science and Engineering** degree of the Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, during the academic year 2022 - 2024.

Prof. Subhadip Basu,
Department of Computer Science and Engineering,
Jadavpur University.
(Supervisor)

Forwarded By:

Prof. Nandini Mukhopadhyay,
HOD, Department of Computer Science and Engineering,
Jadavpur University.

Prof. Dipak Laha,
DEAN, Faculty of Engineering and Technology,
Jadavpur University.

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University, Kolkata - 700032

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled **"Protein-Protein Interactions Prediction, Drug Repurposing and Molecular Docking for emerging viral threats"** is a bonafide record of work carried out by **Debarati Paul** (Registration Number 163558 of 2022 - 2023; Class Roll No. 002210502007; Examination Roll No. M4CSE24004) in partial fulfillment of the requirements for the award of the degree of **Master of Engineering** in **Computer Science and Engineering** in the Department of Computer Science and Engineering, during the period of September 2022 to May 2024. It is understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose of which it has been submitted.

_____
Signature of Examiner 1
Date:

_____
Signature of Examiner 2
Date:

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University, Kolkata - 700032

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that the thesis entitled **"Protein-Protein Interactions Prediction, Drug Repurposing and Molecular Docking for emerging viral threats"** contains a literature survey and original research work by the undersigned candidate, as a part of her degree of **Master of Engineering** in **Computer Science and Engineering, Jadavpur University**. All the information have been obtained and presented in accordance with academic rules and ethical conduct.
I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Debarati Paul

Examination Roll No.: M4CSE24004

Registration No.: 163558 of 2022 - 2023

Thesis Title: Protein-Protein Interactions Prediction, Drug Repurposing and Molecular Docking for emerging viral threats

Signature of the Candidate:

. . . . . .

Dedicated to my parents.
For their endless love, support and encouragement

. . . . . .

Acknowledgements

The success and outcome of this thesis required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I owe my deep gratitude to my mentor and thesis guide **Prof. Subhadip Basu**, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, who took a keen interest in my thesis work and guided me all along, till its completion. I express my heartfelt thanks for his dedicated involvement, constant suggestions, and editorial enhancements, which served as enduring sources of motivation and inspiration throughout the process.

I would like to thank **Dr. Mita Nasipuri**, former co-ordinator and **Prof. Mahantapas Kundu**, present co-ordinator of the Centre for Microprocessor Application for Training and Educational Research (CMATER) lab, Department of Computer Science and Engineering, Jadavpur University for permitting me to conduct this study in the renowned CMATER lab and rendering valuable suggestions.

I would also like to thank **Prof. Nandini Mukhopadhyay**, Head of the Department, Department of Computer Science and Engineering, Jadavpur University, for providing all the departmental facilities and her valuable comments.

My special thanks to **Dr. Anup Kr. Halder**, **Dr. Sovan Saha**, and **Dr. Tapabrata Chakraborti** for their hard work and substantial contribution to uplift the studies presented in this thesis. I am also grateful to my other co-authors Mr. Aanzil Akram Halsana, Mr. Tapas Chakraborty, and Ms. Rupali Patua for their support and encouragement.

I would express my sincere gratitude to all my Professors at Jadavpur University. I would also like to thank my classmates and labmates Mr. Sujan Sarkar, Dr. Neelotpal Chakraborty, Dr. Nirmal Das, Ms. Rakhi Tewari, Mr. Amartya Ray, Mr. Ankit Das, Mr. Akash Halder, Ms. Tirtha Das, and Ms. Godhuli Das for their continuous support and the enjoyable experiences shared over the past two years. Learning with them has been a great experience, and I appreciate their presence in this wonderful journey.

Most importantly, none of this could have happened without the love and support of my family, particularly my father, **Mr. Basudev Paul** and my mother **Mrs. Monidipa Paul**, who have been instrumental in supporting and encouraging me throughout this journey.

Debarati Paul

Examination Roll No.: M4CSE24004

Registration No.: 163558 of 2022 - 2023

Department of Computer Science and Engineering,

Jadavpur University.

Abstract

PPIs are fundamental to the functioning of biological systems, regulating a myriad of cellular processes. Therefore, accurate prediction of PPIs is crucial for disease diagnosis and the development of novel or repurposed therapeutic interventions. Computational techniques streamline the forecast of PPIs and prove to be more cost-effective in terms of resources and time than experimental approaches. In this thesis, initially, a CNN-based novel methodology EnMAS has been proposed to predict PPIs by transforming amino acid sequences of interacting proteins into 2D images using a colour encoding scheme derived from evolutionary data such as PAM and BLOSUM matrices. Additionally, EnMAS has been applied to human-SARS-COV-2 PPIN to predict interactions between human hosts and the COVID-19 virus. This is followed by another GNN-based study GSPPI where a topological feature of PPIN called graphlet (up to 5 nodes) has been leveraged to predict PPIs. Subsequently, in the next study, this graphlet feature and various amino acid composition-based features have been integrated to predict PPIs. This ensemble characteristic has demonstrated excellent prediction performance for PPI on the HIPPIE dataset. Owing to this performance, an illustrative case study has been conducted to predict the PHIs on the interacting protein network of MPXV and Humans. Mpox, a zoonotic illness triggered by MPXV presents a substantial risk to public health since it may be transmitted from person to person and there are no particular treatments available. Investigating how viral proteins interact with host proteins provides valuable insights into the dynamics of PHIs which is essential for designing effective vaccines. Utilizing the human protein targets for MPXV identified accurately, a drug repurposing study is undertaken to discover potential candidate drugs capable of curing Mpox. Additionally, validation has been conducted through subsequent docking studies on these promising drugs for MPXV using the prominent MPXV crystal structures currently available. The results of our study emphasize the effectiveness of using EnMAS and ensemble feature-based PPI prediction to understand the molecular processes involved in viral infection and to aid in the development of drugs for emerging viral threats such as Mpox and COVID-19. Though we have shown the efficacy of our method on Mpox as a case study, the approach is generalizable to a broad range of new emerging viral diseases with potential epidemic or pandemic threats, where the proposed method can reduce the risks and cost of developing new compounds and can offer a viable answer to the time constraints of traditional drug development.

Contents

Abstract **i**

Table of Contents **iii**

List of Publications **vii**

List of Figures **ix**

List of Tables **xi**

List of Abbreviations **xiii**

1 Introduction **1**
1.1 Background 1
1.2 Literature Survey 2
1.3 Research Gap Analysis 4
1.4 Objectives of Current Research 5
1.5 Organization of the thesis 6

2 Sequence-based PPIs prediction by encoding the mutation rates of Amino acids **7**
2.1 Introduction 7
2.2 Methodology 8
2.2.1 Benchmark Data Preparation 8
2.2.1.1 Saccharomyces cerevisiae (*S.cerevisiae*) Dataset . . 8
2.2.1.2 Human protein Interaction Dataset $JUPPI_d$ . . . 9
2.2.2 Sequence Encoding 10
2.2.2.1 Colour Mapping Strategy for Sequences 10
2.2.2.2 Sub-images generation strategy 11
2.2.3 Deep learning Model 12
2.2.3.1 Architecture and hyperparameters 12
2.2.3.2 Classification strategy 13
2.3 Results 14
2.3.1 Evaluation Metrics 14
2.3.2 Performance on *S.cerevisiae* Dataset 14
2.3.3 Performance on $JUPPI_d$ Dataset 15
2.3.4 Application of EnMAS on Human-SARS-COV-2 interaction 17
2.4 Discussion 19

3 Prediction of PPIs using graphlet feature on HIPPIE dataset 21
3.1 Introduction . . . 21
3.2 Methodology . . . 22
3.2.1 Positive data selection . . . 22
3.2.2 Negative data selection . . . 22
3.2.3 Feature Extraction . . . 23
3.2.4 Training and Classification . . . 24
3.3 Results . . . 25
3.3.1 Performance of the proposed model GSPPI . . . 25
3.3.1.1 Effect of number of layers on the performance of our proposed model . . . 26
3.3.2 Comparison of the proposed model with other GNN variants 27
3.4 Discussion . . . 28

4 Prediction of PPIs using ensemble features on HIPPIE dataset 29
4.1 Introduction . . . 29
4.2 Methodology . . . 29
4.2.1 Data Selection . . . 30
4.2.2 Feature Extraction . . . 30
4.2.3 Deep Neural Network Model . . . 34
4.2.4 Classification Strategy . . . 35
4.3 Results . . . 35
4.3.1 Performance of the proposed model with different feature selection strategies . . . 36
4.3.2 Performance of the proposed model with different node features . . . 36
4.3.3 Comparison of the proposed model with other GNN variants 37
4.4 Discussion . . . 38

5 Prediction of PHIs using ensemble features on MPXV-Human PPIN 39
5.1 Introduction . . . 39
5.2 Methodology . . . 39
5.2.1 Formation of MPXV-Human PPIN . . . 39
5.2.2 Feature extraction of MPXV-Human PPIN . . . 40
5.2.3 Deep Neural Network Model . . . 40
5.2.4 Classification Strategy . . . 41
5.3 Results . . . 41
5.4 Discussion . . . 42

6 Computational Drug Repurposing Study on MPXV-Human PPIN 43
6.1 Introduction . . . 43
6.2 Methodology . . . 44
6.3 Results . . . 45
6.4 Discussion . . . 50

7 Computational Docking of potential drugs for MPXV protein structures **51**
7.1 Introduction 51
7.2 Methodology 51
7.3 Results 52
7.4 Discussion 56

8 Conclusions **57**
8.1 Limitations of the Work 58
8.2 Scope for Future Works 59

Bibliography **61**

List of Publications

1) JOURNAL

i) **Debarati Paul**, Sovan Saha, Subhadip Basu, and Tapabrata Chakraborti. "Can in-silico drug repurposing computationally suggest fast and low-risk therapies for emerging viral threats like Mpox?" ***Nature Scientific Reports (2024).*** **- Communicated (under review)**

2) CONFERENCE

i) **Debarati Paul**, Rupali Patua, Sovan Saha, Anup Kumar Halder, Subhadip Basu. "GSPPI: GraphSAGE-based Prediction of Protein-Protein Interactions using Graphlet features" In ***2nd International Conference On Data, Electronics And Computing (ICDEC-2023), December 15-16, 2023, Mizoram University, Aizawl, Mizoram.*** **- Accepted for Publication**

List of Figures

2.1 Overview of the process flow of *EnMAS*: the bio-inspired sequence-based encoding strategy for PPI prediction. 9

2.2 Performance comparison of *EnMAS* on Yeast (*S.cerevisiae*) PPI dataset and Human PPI dataset ($JUPPI_d$: C2). 17

2.3 Performance assessment of *EnMAS* with four variations (BLOSUM 45, BLOSUM 90, PAM 120, and PAM 250) on *SARS-COV-2: Human* interactions using C1, C2, and C3 type models. The top, middle, and bottom blocks represent the performances of the C1, C2, and C3-based models, respectively. 18

3.1 Basic workflow of the proposed model GSPPI. 23

3.2 Basic schematic diagram of link predictor of the GSPPI model. . . . 24

3.3 Performance metrics of three GNN variants using graphlet feature on HIPPIE dataset. 28

4.1 Basic workflow of the ensemble feature generation method on HIPPIE dataset. 31

4.2 a)Plotting AUC-ROC (in %) and number of composition-based features selected against each threshold variance using Variance Threshold Feature Selection Method in HIPPIE PPIN. b) Plotting AUC-ROC (in %) and number of ensembled features selected against each threshold variance using Variance Threshold Feature Selection Method in HIPPIE PPIN. 33

4.3 Basic schematic diagram of the ensemble feature-based model. . . . 34

4.4 Performance metrics of three GNN variants using ensemble (graphlet and composition-based) features on HIPPIE dataset. 38

5.1 a) Confusion matrix of the proposed ensemble-based feature model on MPXV-Human PPIN. b) AUC-ROC plot of the proposed model with ensemble feature, sequence-based feature, and graphlet feature with the AUC-ROC. 42

7.1 a) Crystal structure of MPXV methyltransferase VP39 in complex with inhibitor TO427 retrieved from PDB. b) Molecular structure of NADH retrieved from DrugBank. 53

7.2 NADH docked in the MPXV methyltransferase VP39 in complex with inhibitor TO427 (PDB ID: 8CEQ) a) best binding mode in the protein pocket. b) the binding interaction of NADH with amino acids with various bonds. 54

List of Tables

2.1 Sub-image and original image data size on *S.cerevisiae* and $JUPPI_d$ datasets. . . . 13
2.2 Performance comparison of DensePPI and EnMAS on *S.cerevisiae* dataset. The best values among the BLOSUM and PAM matrices are highlighted in bold. . . . 15
2.3 Performance comparison of JUPPI(S), DensePPI, BLOSUM, and PAM-based EnMAS on $JUPPI_d$ human dataset. The best values are highlighted in bold. . . . 16

3.1 Performance scores of our proposed model on the training dataset, validation dataset, and testing dataset . . . 26
3.2 Performance scores of our proposed model on the hold-out dataset for 5-fold CV. The best values are in bold. . . . 26
3.3 The results of our proposed model on test dataset with different number of layers for PPI datasets. The best values are in bold. . . . 27
3.4 The results of the average 5-fold CV of our proposed model with different numbers of layers on the PPI dataset. The best values are in bold. . . . 27

4.1 Performance of our proposed model with FastICA, PCA, and Variance threshold feature selection method on HIPPIE dataset. The best values are in bold. . . . 36
4.2 Performance scores of our proposed model with graphlet feature, sequence-based feature, and ensemble feature on the HIPPIE dataset. The best values are in bold. . . . 37

5.1 Performance scores of our proposed model with graphlet feature, sequence-based feature, and ensemble feature on MPXV-Human dataset. The best values are in bold. . . . 42

6.1 Domain, DrugBank Accession Number, Drug Name, DCS, and DrugBank Groups of the drugs identified through *ProcessDrugData* algorithm. . . . 47

7.1 Docking scores/Affinity values along with the distance from the best mode of the nine modes of NADH. The best values are in bold. 54
7.2 Docking scores/Affinity values along with the distance from the best mode of the nine modes of Fostamatinib. The best values are in bold. . . . 55

7.3 Docking scores/Affinity values along with the distance from the best mode of the nine modes of Glutamic acid. The best values are in bold. 55
7.4 Docking scores/Affinity values along with the distance from the best mode of the nine modes of Cannabidiol. The best values are in bold. 56
7.5 Docking scores/Affinity of top four repurposed drugs for 8CEQ. . . 56

List of Abbreviations

A

AAC	**A**mino **a**cid **C**omposition
AC	**A**uto-**c**ovariance
AI	**A**rtificial **I**ntelligence
AMP-AD	**A**ccelerating **M**edicines **P**artnership-**A**lzheimer's **D**isease
APAAC	**A**mphiphilic **P**seudo **A**mino **A**cid **C**omposition
ATC	**At**omic **C**omposition
AUC-ROC	**A**rea **U**nder the **C**urve of the **R**eceiver **O**perating **C**haracteristic
AUPRC	**A**rea **U**nder the **P**recision-**R**ecall **C**urve

B

BIND	**B**iomolecular **I**nteraction **N**etwork **D**atabase
BLOSUM	**BLO**cks **SU**bstitution **M**atrix
BTC	**B**ond **T**ype **C**omposition

C

CNN	**C**onvolutional **N**eural **N**etwork
COVID-19	**Co**rona**vi**rus **D**isease-20**19**
CT	**C**onjoint **T**riad
CTC	**C**onjoint **T**riad **C**alculation
CV	**C**ross **V**alidation

D

DCS	**D**rug **C**onsensus **S**core
DDR	**D**istance **D**istribution of **R**esidues

DIP	**D**atabase of **I**nteracting **P**roteins
DL	**D**eep **L**earning
DNN	**D**eep **N**eural **N**etwork
DPC	**Dip**eptide **C**omposition
DRC	**D**emocratic **R**epublic of the **C**ongo
DRIAD	**D**rug **R**epurposing **i**n **A**lzheimer's **D**isease
F	
FAD	**F**lavin **A**denine **D**inucleotide
FDA	**F**ood and **D**rug **A**dministration
FRN	**F**eature-Relational **R**easoning **N**etwork
G	
GAT	**G**raph **At**tention Network
GCN	**G**raph **C**onvolutional **N**etwork
GNN	**G**raph **N**eural **N**etwork
GO	**G**ene **O**ntology
GraphSAGE	**Graph SA**mple and aggre**G**at**E**
H	
HIPPIE	**H**uman **I**ntegrated **P**rotein-**P**rotein **I**nteraction r**E**ference
HPRD	**H**uman **P**rotein **R**eference **D**atabase
I	
ICA	**I**ndependent **C**omponent **A**nalysis
L	
LSTM	**L**ong **S**hort-**T**erm **M**emory
LUT	**L**ook-**u**p **T**able
M	
MCC	**M**atthews **C**orrelation **C**oefficient
MINT	**M**olecular **INT**eraction
MOS	**M**atrix **o**f **S**equence

Mpox	**M**onkey**pox**
MPXV	**M**onkey **P**o**x** **V**irus
MSE	**M**ean **S**quared **E**rror
N	
NADH	**N**icotinamide **A**denine **D**inucleotide **H**ydrogen
NME	**N**ew **M**olecular **E**ntity
O	
OMIM	**O**nline **M**endelian **I**nheritance in **M**an
ORCA	**OR**bit **C**ounting **A**lgorithm
P	
PAAC	**P**seudo **A**mino **A**cid **C**omposition
PAM	**P**oint **A**ccepted **M**utation
PCA	**P**rincipal **C**omponent **A**nalysis
PCP	**P**hysico-**c**hemical **p**roperties
PDB	**P**rotein **D**ata **B**ank
PHI	**P**athogen **H**ost **I**nteraction
PPI	**P**rotein **P**rotein **I**nteraction
PPIN	**P**rotein **P**rotein **I**nteraction **N**etwork
PPNI	**P**rotein-**P**rotein **N**on-**I**nteractions
Q	
QSO	**Q**uasi-**S**equence **O**rder
R	
RCSB	**R**esearch **C**ollaboratory for **S**tructural **B**ioinformatics
ReLU	**Re**ctified **L**inear **U**nit
RMSD	**R**oot **M**ean **S**quare **D**eviation
RRI	**R**epetitive **R**esidue **I**nformation
S	
SAE	**S**tacked **a**uto**e**ncoder

SDF **S**tructure **D**ata **F**ile

SEP **S**hannon **E**ntropy at **P**rotein Level

SER **S**hannon **E**ntropy at **R**esidue Level

SGD **S**tochastic **G**radient **D**escent

SOC **S**equence **O**rder **C**oupling Number

V

VACV **Vac**cinia **v**irus

W

WHO **W**orld **H**ealth **O**rganization

WSR **W**eighted **S**parse **R**epresentation

CHAPTER 1

Introduction

1.1 Background

Protein is one of the essential biological macromolecules of all living organisms. They are involved in a wide variety of cellular processes [1] including metabolism, signal transmission, hormone control, DNA transcription, and replication [2]. Proteins generally interact with other proteins to fulfill their biological activities. Uncovering their interactions, namely PPI data, aids in unveiling the fundamental structural traits of proteins [3] and the molecular machinery within cells [4]. Decoding PPIN is essential for comprehending the complexities of cellular machinery and gaining insights into the molecular foundation of health and illness.

Within the realm of infectious diseases, the interaction of proteins between the pathogen and the host organism becomes a central focus. Pathogens, including viruses, utilize their proteins to target human proteins to disrupt or manipulate biological circuits specifically [5]. The interactions cover a wide range, from attaching to host membrane receptors to enlisting host factors through viral proteins [6]. PHIs determine the result of an infection, affecting the host's reaction, the ability of the pathogen to avoid the immune system, and the course of the disease [7]. Therefore, studying the mechanisms of PHIs reveals the molecular foundations of infectious illnesses, facilitating the discovery of medicinal targets [8] and the diagnosis of diseases [9]. However, there is still a lack of clinically recognized drugs for effectively treating some emerging viral threats such as Mpox and COVID-19.

COVID-19 is an infectious disease caused by the SARS-COV-2 virus. The first known case was identified in Wuhan, China, in December 2019 [10]. While humanity was struggling with the COVID-19 pandemic, a Mpox outbreak in May 2022 in the United Kingdom emerged suddenly [11] and rapidly spread across Europe, the Americas, and subsequently all 6 WHO regions, with 110 countries reporting around 87000 cases and 112 deaths [12]. This outbreak caused by the Mpox virus

clade IIb, marked the first instance of widespread community transmission outside Africa, with subsequent cases confirmed in 111 countries by May 2023 [13].

Mpox [14] is an infectious viral disease caused by the MPXV, classified under the Orthopoxvirus genus under the Poxviridae family [15]. It manifests with symptoms like skin rash, fever, muscle pains, and swollen lymph nodes [16]. The first recorded human case dates back to 1970 in a 9-month-old boy in Basankusu, DRC [17]. The WHO declared it a 'Public Health Emergency of International Concern' from July 23, 2022, to May 10, 2023 [18]. Following the 2022-2023 outbreak, Mpox persisted very low in the human population outside Africa. In November 2023, the WHO reported a rise in Mpox cases (clade I) in the DRC, with 12,569 cases reported year-to-date and 651 fatalities. Till now, there is no scientifically validated and authorized vaccination for Mpox, despite extensive efforts by researchers to create one. Therefore, it is essential to comprehend and scrutinize the method by which this novel virus is transmitted.

1.2 Literature Survey

In recent years, the availability of large-scale sequencing data from high-throughput technologies has led to the development of many computational approaches for predicting intra-species and inter-species PPIs, using features derived from protein sequences. Accurate predictions of PPIs are achieved by identifying relevant features and applying suitable ML or DL techniques. Advancements in DL technologies have significantly contributed to this field.

Given the scarcity of structures, there has been an increased focus on utilizing sequences alone to forecast PPIs [19] [20] [21]. Sun et al. [22] developed a PPI predictor SAE where the input sequence had been encoded using AC and CT. For PPI prediction, DeepPPI [23] introduced various features from protein sequences, such as quasi-sequence, amino acid composition, and dipeptide composition. The research [24] had proposed a sequence-based DL system called DPPI employing a deep CNN that resembled a Siamese network combined with random projection and data augmentation techniques to provide precise predictions of PPIs. Li et al. have introduced a DNN framework, DNN-PPI [25] to predict PPIs by leveraging automatically learned features exclusively from protein primary sequences and utilizing CNN and LSTM neural network layers.

In recent years, there has been substantial advancement in GNNs, establishing them as essential tools in applications based on graph structures. In the work of [26], Singh et al. have introduced a methodology, Topsy-Turvy, which combines graph-theoretic (top-down) and sequence-based (bottom-up) approaches during training of their sequence-based predictor. Jha et al. [27] has proposed a frame-

work that combines graph-based techniques like GCN and GAT with language models (SeqVec [28] and ProtBert [29]) for the task of PPI prediction. Wu et al. have proposed DLPPI which has also utilized protein sequence data for PPI prediction with the help of FRN and GNNs [30]. Koca et al. [31] has proposed a GraphSAGE [32]-based model for PHI prediction from sequences using the Doc2Vec. A recent method, DensePPI [33] detects interactions by transforming amino acid sequences of interacting proteins into images using a randomized colour encoding scheme where the 20 amino acids have been mapped to the 26 most distinct colors in the RGB colour spectrum, as chosen in [34]. Halder et al. developed a multi-layered feature-based strategy called JUPPI [35] for PPI prediction, using sequence, domain, and GO features for feature generation. They incorporated a pair-input method in CV, with the test set divided into three difficulty levels: C1, C2, and C3. Thus sequence-based feature extraction approaches have proven their effectiveness in the prediction of PPIs or PHIs.

Several studies have combined multiple features for accurate PPI prediction. Jia et al. [36] has developed a predictor, called "iPPI-Emsl" by incorporating physicochemical attributes derived from the constituent amino acids of a protein, and wavelet transformations extracted from numerical series along a protein chain. In the work of [37], PPI was forecasted by the fusion of protein primary sequence and evolutionary data utilizing a novel protein sequence substitution matrix feature representation and an ensemble WSR model classifier for integration.

Investigating PPIs in the human and MPXV network is essential since it allows us to understand the target proteins through which infection gets mediated [5]. It can uncover prospective therapeutic targets. Researchers can accelerate the discovery of therapies for Mpox by repurposing current drugs [38]. Numerous computational approaches utilizing ML and DL techniques for drug repurposing continue to be developed and refined, addressing emerging challenges. Rodriguez et al. [39] have proposed the DRIAD method that employs an ML system to measure the association between the various stages of Alzheimer's disease and the biological processes or responses that a set of gene names can identify. Data from the AMP-AD datasets and gene names from RNA-seq have been utilized to create drug-associated gene lists. In a study [40], the researchers utilized a multitask deep learning model trained on data for 391 kinases to identify potential targets for both clinically approved and investigational kinase drugs. Li et al. [41] employed DL techniques with transcriptomic and molecular data to discover the repurposing potential of pimozide for non-small cell cancer which was subsequently tested *in vitro* (IC_{50} 19.54 μM) in A549 cells.

This survey has helped to identify research gaps highlighted in the subsequent section 1.3 and the scope of addressing them through the current study.

1.3 Research Gap Analysis

Randomized colour encoding strategy for Image-Based PPIs prediction from protein sequences

- Despite the novelty of DensePPI [33], the method faced a significant limitation due to the lack of a biological rationale for the colour scheme employed.
- Experimentally validated databases for high-quality negative data, such as PPNI data are limited [42] [43]. A refined approach is needed to curate negative PPIs effectively. Traditional CV, influenced by representational bias, may affect PPI prediction models. [44] [45]. Computational PPI prediction methods typically consider pairs of proteins as single data, termed *'pair-inputs'*. In this paired nature of input data, features from individual proteins of pair-input data can influence the predictive model, potentially introducing bias when it occurs in both the train and/or test sets simultaneously as an individual component. Hence, a novel robust CV technique introduced by Halder et al. [35] must validate this image-based PPI prediction task.

Sequence-based Method for prediction of PPIs utilizing variants of GNNs

- Interaction prediction using sequence-based features has been implemented in multiple works. However, sequence-based PPI site prediction relies entirely on features derived from analyzing amino acid sequences in proteins, ignoring structural or functional details. Moreover, it faces challenges when neither of the partner proteins has similar sequences represented in the training dataset, hindering accurate predictions. This limitation underscores the challenge of relying solely on sequence information for PPI predictions, as the effectiveness of such methods can be constrained by the absence of sufficiently similar sequence patterns in the training data. [46] [45] [47].

Effective PHIs detection and thereby anti-virals development for emerging viral threats

- Although vaccination efforts successfully eradicated smallpox globally, Mpox remains a sporadic and localized disease for which specific antiviral medication is still unavailable. In this context, understanding the interactions between MPXV proteins and human host proteins could shed light on how the virus evades the host's immune response. Predicted PHIs may reveal potential drug targets within the host-virus interaction network and aid in the identification of antigens that could be targeted for vaccine development.

1.4 Objectives of Current Research

Observing the issues discussed in section 1.3, this thesis addresses the following objectives to overcome the drawbacks of the existing approaches.

Bio-inspired colour encoding strategy for Image-Based PPIs and PHIs prediction from protein sequences

- This thesis presents an innovative bio-inspired DL model for identifying interacting protein pairs. Contrary to the random colour generation approach in DensePPI, the mutation rate of amino acids in a protein sequence over long evolutionary time scales from substitution matrices such as BLOSUM and PAM is used to generate 2D images from the amino acid sequences of the interacting proteins. This model has been leveraged to evaluate the interaction between SARS-COV-2 and humans.
- To counteract the representational bias, the substitution matrix-based colour encoding scheme is validated with the $JUPPI_d$ Dataset thereby reducing the component-level overlapping between the train and test sets. [35].

Incorporating topological information for prediction of PPIs utilizing variants of GNNs

- Appreciating the significance of graphlets in topological representation, a novel approach has been devised utilizing topological feature information as the node feature in a PPIN to predict PPI.
- To enhance the predictive capability of the model, a novel methodology, integrating graphlets as well as sequence-based features for the targeted objective of predicting PPIs through the application of the DL-based GraphSAGE model is proposed. The ensemble feature set has captured both the underlying organizational principles and the inherent information encoded in protein sequences providing a more versatile approach for PPIs prediction.

Leveraging Ensemble feature-based model for PHIs detection

- A demonstrative case study is conducted to predict the potential PHIs on the MPXV-Human PPIN using the similar graph-based DL approach and ensemble feature set. This has enabled us to comprehend the specific host proteins targeted by MPXV, which can further reveal potential treatment targets.

Drug repurposing case study on MPXV-Human PPIN

- A drug repurposing study is undertaken for Mpox. After conducting a comprehensive investigation, it has been found that several protein nodes in the MPXV-Human PPIN function as protein targets for various drugs approved by the FDA in DrugBank. The drugs included in this research have been determined using the consensus scoring strategy. The found repurposed drugs have been classified into three domains, namely Level I, Level II, and Level III, according to their derived DCS. A concurrent docking investigation has been conducted which provides further insights into the effectiveness of these newly found repurposed drugs.

1.5 Organization of the thesis

In the light of the discussion in the previous section, the thesis has been organized as follows:

i) **Chapter 2** describes a substitution matrix-based model called EnMAS whereby the evolutionary information has been harnessed in our sequence-to-image colour encoding strategy to predict protein interactions. This model is also used to predict the PHIs between SARS-COV-2 and humans.

ii) **Chapter 3** discusses a GraphSAGE-based PPI prediction model, GSPPI utilizing graphlet features (up to 5 nodes) of a protein node in HIPPIE PPIN.

iii) **Chapter 4** discusses a GraphSAGE-based PPI prediction model that uses an ensemble of graphlet features (up to 5 nodes) and amino acid composition-based features on HIPPIE PPIN.

iv) In **Chapter 5**, the ensemble feature-based model introduced in 4 has been leveraged to predict PHIs in MPXV-Human PPIN.

v) **Chapter 6** highlights an FDA-approved drug repurposing study on the MPXV-targeted human proteins found in 5.

vi) **Chapter 7** presents a computational docking study of the top four Level I repurposed drugs for MPXV protein structures, presented in 6 and

vii) **Chapter 8** concludes the thesis with limitations and future scopes.

CHAPTER 2

Sequence-based PPIs prediction by encoding the mutation rates of Amino acids

2.1 Introduction

Deciphering hidden messages in DNA and determining the interrelation between species is crucial in bioinformatics. Sequence alignment, an arrangement of DNA, or RNA segments, can play a crucial role in pinpointing regions of similarities that can result directly from structural, functional, or evolutionary relationships between the sequences. In this regard, scoring matrices can quantify this alignment's similarities [48]. They outline the rate at which different amino acid residues in protein sequences are replaced by other residues over evolutionary time [49]. Higher scores suggest that the two amino acids are evolutionary related and lower scores indicate a high likelihood that the two amino acids are evolutionary unrelated.

In this chapter, the aforementioned evolutionary relationship between two protein segments has been harnessed in our sequence-to-image colour encoding strategy to predict the interaction between two peptide sequences. In this proposed approach named "EnMAS", BLOSUM [50] and PAM [51] substitution matrices have been used for generating 2D images. BLOSUM matrices are usually used for scoring alignments between evolutionary divergent protein sequences whereas PAM substitution matrices are used for closely related homologs. Amino acid sequences of the interacting protein pairs are converted into their image representations using a colour encoding scheme derived from these substitution matrices, thus incorporating the biological and evolutionary information missing in

DensePPI[33]. The generation of square sub-images follows this colour encoding process of original images from each, which are then used to train the DL model (DenseNet201). Finally, the trained model is validated using the traditional CV on the Yeast Dataset and the 'pair-input based CV' on the $JUPPI_d$ Dataset. Our proposed EnMAS model has also been applied to predict interactions between human hosts and the SARS-COV-2 virus.

In the following, the chapter is organized as; Section 2.2 discusses the proposed methodology in three major steps of the EnMAS model; 1) Benchmark data preparation 2.2.1, 2) Sequence Encoding 2.2.2, and 3) Deep Learning model 2.2.3. Finally, the result and discussion are elaborated in section 2.3 and 2.4 respectively.

2.2 Methodology

The overall workflow of the EnMAS model *i.e.* Encoding the Mutation rates of Amino acids, has been depicted in Fig 2.1.

2.2.1 Benchmark Data Preparation

The process of preparing the PPI dataset for this study from the existing benchmarked data is outlined in the subsequent sections. Benchmark PPI Datasets of both Human and Yeast proteins are considered. Full-length sequences of interacting pairs taken with n and m number of amino acids are depicted in Fig 2.1A.

2.2.1.1 Saccharomyces cerevisiae (*S.cerevisiae*) Dataset

In our study, we employed the *S.cerevisiae* PPI dataset introduced by Du et al. [23], sourced from the DIP [52]. The dataset, originally consisting of 22,975 positive pairs, underwent refinement through various criteria, reducing the positive dataset to 17,257 pairs. Negative data pairs were formed by combining proteins from different subcellular locations [53], with a cap of $\sim 2,500$ pairs chosen from each localization. Pairing strategies ensured the exclusion of non-interacting pairs from the positive dataset and an equitable representation of proteins within the negative set [20] [19], which yielded a total of 48,594 negative pairs. An equal ratio of positive and negative sub-image samples is maintained to achieve balanced learning.

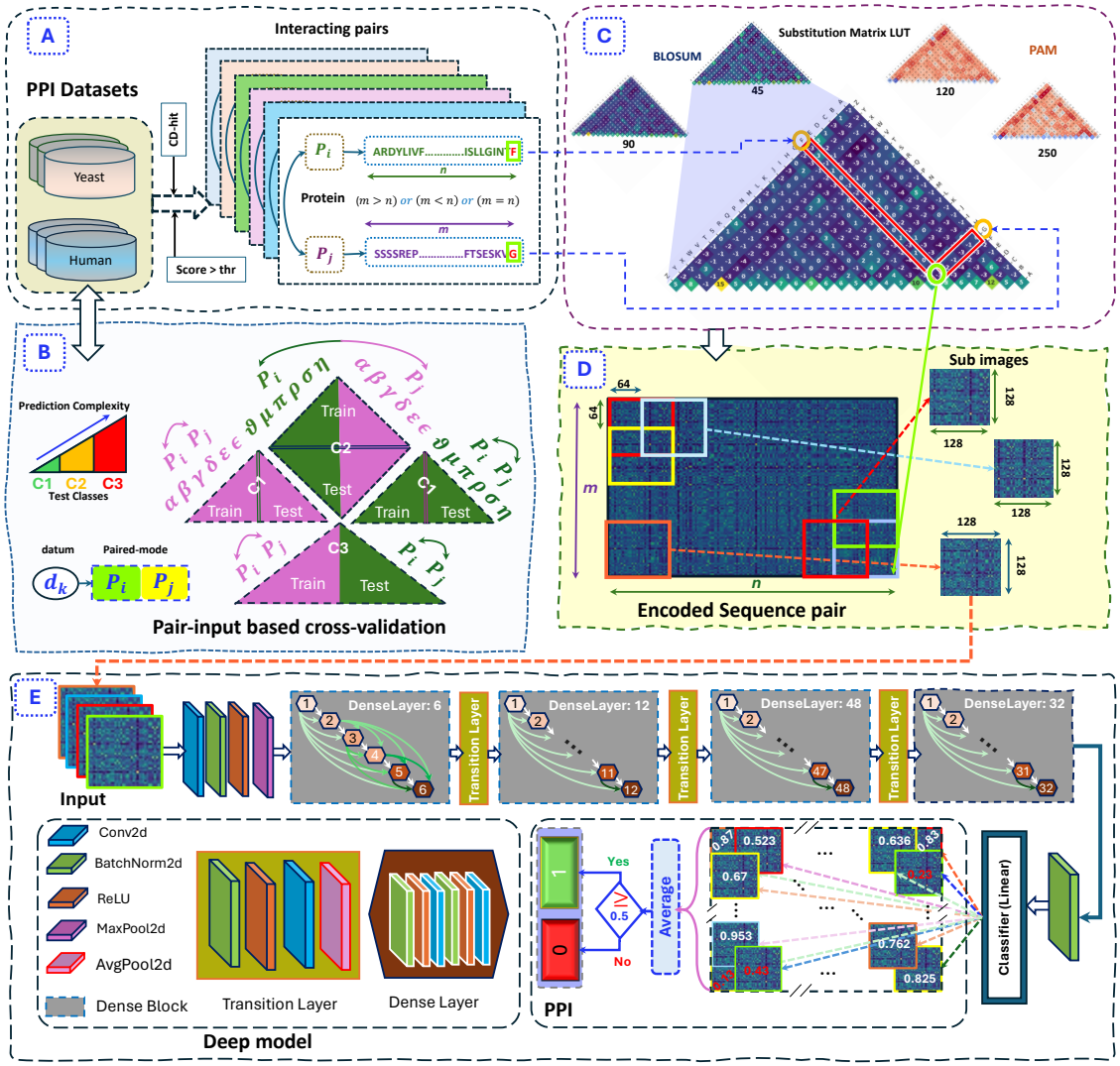


Figure 2.1: Overview of the process flow of *EnMAS*: the bio-inspired sequence-based encoding strategy for PPI prediction.

2.2.1.2 Human protein Interaction Dataset $JUPPI_d$

JUPPI dataset [35] is used in our study for predicting human protein interaction information. The JUPPI dataset was compiled from the HIPPIE v2.1 database, which contains the human PPIs sourced from seven major databases: DIP[54], MIPS[55], BIND[56], BioGRID[57], IntAct[58], MINT[59] and HPRD[60]. The selection of $JUPPI_d$ data was based on PPIs with HIPPIE confidence scores exceeding 0.75. CD-Hit [61] with less than 40% identity is applied to remove the homologous sequences from the redundant and homologous proteins between the train and test set and within the interacting or non-interacting protein pair. Consequently, out of 7,459 distinct proteins, $JUPPI_d$ has the most trustworthy 27,288 physical non-redundant positive interactions.

The dataset was enhanced by incorporating subcellular location information from the UniProt database [62]. Negative data were generated through a three-

step filtering method, excluding protein pairs not in the positive set, pairs from distinct subcellular locations, and those lacking interaction evidence with third-level neighbors. Following these criteria, the dataset included 40,467 negative interactions.

The $JUPPI_d$ dataset employed a pair-input-based CV approach to eliminate component-level overlapping between training and testing sets. The prediction difficulty of the test classes was categorized into three distinct types: C1, C2, and C3 [45]. C1 consists of test pairs where both proteins are found in the training set but do not occur together; C2 includes test pairs with only one protein shared with the training set; and C3 encompasses test pairs with no proteins in common with the training set. The pair-input-based CV is illustrated in Fig 2.1B.

2.2.2 Sequence Encoding

Sequence encoding is a crucial step in the prediction of PPIs, enabling computational models to process amino acid sequences and predict whether two proteins interact. This strategy transforms the sequence of amino acids into a format that ML algorithms can understand. This study introduces a new colour mapping strategy that utilizes substitution matrix-based LUT to assign unique colours based on the similarity and variation between two amino acids from interacting protein pairs, as illustrated in Fig 2.1C.

BLOSUM and PAM matrices are pivotal for sequence comparison, emphasizing evolutionary distances. BLOSUM matrices with higher numbers, such as BLOSUM 90, are tailored for comparing sequences that share a close evolutionary relationship, while lower-numbered matrices like BLOSUM 45 are aimed at comparing more distantly related sequences [50]. On the contrary, the PAM matrix system utilizes lower numbers for alignments of closely related sequences and higher numbers (e.g., PAM 250) for those sequences that are more distantly related, showcasing the inverse relationship in scaling compared to BLOSUM [51]. This dichotomy underscores the distinct scaling and intended applications of BLOSUM and PAM matrices in the field of bioinformatics [63]. In this work, we have introduced this substitution matrices-based sequence representation for sequence encoding for PPI prediction. This technique emphasizes the significance of alignments, simplifies the detection of conserved regions and mutations, and aids in pinpointing the binding positions of the interacting pairs.

2.2.2.1 Colour Mapping Strategy for Sequences

The commonly used BLOSUM and PAM substitution matrices are derived by analyzing sets of high-confidence alignments among many homologous proteins to

assess the frequencies of all possible substitutions. This substitution score from each matrix is incorporated for any interaction to construct the LUT for image generation.

The PAM [51] matrices were developed by analyzing mutations within the phylogenetic trees of protein families, each sharing specific sequence identities, to determine the probability of amino acid substitutions over evolutionary distances. The BLOSUM substitution matrix [50], calculates mutational probabilities between amino acids from numerous aligned protein clusters. It normalizes these probabilities and converts them into log odds (LOD) scores on a logarithmic scale for effective protein comparison.

In our colour encoding strategy, we use the scores from each cell of the substitution matrices to construct the LUTs that generate the colour map. Specifically, we use LUTs from BLOSUM 45, BLOSUM 90, PAM 120, and PAM250. Each LUT is specifically designed for the standard 20 amino acids; the PAM matrix is formatted as a 20×20 matrix. Conversely, the BLOSUM matrix incorporates additional symbols and is structured as a 24×24 matrix. This expansion includes the 20 standard amino acids, plus 'X' (representing any amino acid), 'B' (representing either 'D' or 'E'), 'Z' (representing either 'N' or 'Q'), and '*' (indicating a mutation or substitution to null) [64], as described at NCBI repository (`https://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/data/`).

For each amino acid pair, the designated LUT value specifies the pixel colour in RGBA format in the encoded 2D colour image, where 'R', 'G', 'B', and 'A' respectively represent Red, Green, Blue, and Alpha. For example, if the i^{th} amino acid and the j^{th} amino acid from the sequence pair of any two proteins correspond to a specific LUT score, this score is translated into a pixel value in the resultant image. This process is repeated for all possible amino acid pairs within the interacting proteins to construct the final encoded image. As illustrated in Fig 2.1C, consider protein pairs (P_i, P_j) of lengths n and m with sequences 'ARD....NT$\boldsymbol{F}$' and 'SSS.....KV$\boldsymbol{G}$', respectively. In this example, the amino acid pair Phenylalanine (F) and Glycine (G), the corresponding value in the BLOSUM 45 matrix is -3, which, through its LUT value, provides the pixel colour for that paired region. The resulting image dimensions are $n \times m$, aligning with the lengths of the interacting protein sequences.

2.2.2.2 Sub-images generation strategy

Once the sequence pair is encoded into a 2D image, contiguous sub-images of a fixed size are extracted at regular intervals. In this stage, a window of fixed size 128×128 is slid both horizontally and vertically to traverse the entire orig-

inal image with a fixed stride of 64 (see Fig 2.1D). The decision to adopt this method was made to generate sub-images of consistent dimensions, 128×128, rather than compressing them. This strategy ensures that crucial information is retained, especially for images derived from proteins with longer sequences. The stride specifies the distance the window moves horizontally and vertically as it scans across the image. This method ensures precise control over the overlap between adjacent windows, allowing each sub-image to maintain the same resolution and aspect ratio as the original image. The core principle of this technique is that all 2D sub-images derived from protein interactions have consistent predictive capabilities. By extracting sub-images, each interaction is given equal importance, regardless of the protein sizes involved. Subsequently, sub-images resulting from positive protein interactions have been labeled as positive datasets, while those from negative interactions have been labeled as negative datasets. These datasets have been further used for training our DL model. The cardinality of the main image sets derived from protein pairs, as well as the cardinality of the corresponding positive and negative sub-image sets used in the training and testing phases for each dataset, are detailed in Table 2.1.To ensure the original proportions of images are preserved during sub-image creation, protein pairs containing any sequence less than 128 in length have been excluded from the process.

2.2.3 Deep learning Model

The proposed model EnMAS, employs DenseNet201, a CNN architecture recognized for its dense connectivity pattern, as illustrated in Fig 2.1E. The network starts with a 2D convolution layer having 3 input channels and 64 output channels, followed by batch normalization and a ReLU activation function. Max pooling is conducted with a 3x3 kernel, reducing the feature map dimensions.

2.2.3.1 Architecture and hyperparameters

The architecture comprises four dense blocks separated by transition layers that initiate with batch normalization, ReLU activation layer and compress and down-sample spatial dimensions through 1x1 convolution and 2x2 average pooling, respectively. Each dense block consists of densely interconnected layers, with each layer receiving inputs from all preceding layers and passing outputs to all subsequent layers. The block configuration includes 6, 12, 48, and 32 dense layers sequentially.

Each dense layer begins with batch normalization to stabilize and expedite training, followed by ReLU activation to introduce non-linearity, and a 1x1 convolution to process features. This sequence is replicated with another round of

Table 2.1: Sub-image and original image data size on *S.cerevisiae* and $JUPPI_d$ datasets.

Dt.	TC	Set	Type	Positive	Negative
S.cerevisiae	C1	Training	Original image	13944	13944
			Sub-image	1022455	907305
		Test	Original image	1549	1549
			Sub-image	111143	102073
$Human : JUPPI_d$	C1	Training	Original image	21207	20480
			Sub-image	2241860	1533724
		Test	Original image	2372	2359
			Sub-image	237352	190008
	C2	Training	Original image	8906	7455
			Sub-image	990743	475625
		Test	Original image	676	592
			Sub-image	70417	38927
	C3	Training	Original image	10120	10333
			Sub-image	1106007	898345
		Test	Original image	2698	2874
			Sub-image	258501	222427

N.B.: Dt: DataSet, TC: Test class.

batch normalization, ReLU activation, and a 3x3 convolution. The output from the final dense block undergoes batch normalization before being processed by a linear classifier, finalizing the model structure.

Additionally, average pooling is utilized between layers with a momentum of 0.9 and a learning rate of 0.001. The categorical cross-entropy measure is chosen as the loss function due to the binary classification nature of the prediction task. The networks are optimized using mini-batch gradient descent with a batch size of 32 over 10 epochs, employing an SGD optimizer to enhance generalization.

2.2.3.2 Classification strategy

The EnMAS model trains on sub-images from an original image linked to a PPI. Utilizing a DenseNet201 architecture, it assigns each sub-image a confidence value between 0 and 1. These confidence values are averaged to determine the overall probability score for a specific PPI. If the average of all predicted probabilities for the sub-images of a particular PPI is 0.5 (threshold) or higher, the interaction is labeled as positive (1); otherwise, it's labeled as negative (0). This classification strategy is illustrated in Fig 2.1E.

2.3 Results

In this section, initially, the various performance metrics to be used in this and subsequent chapters are discussed. Then the effectiveness of the EnMAS model is tested using the *S.cerevisiae* dataset and the $JUPPI_d$ dataset employing a pair-input-based CV approach. Also, the performance of EnMAS on Human-SARS-COV-2 PPIN is noted.

2.3.1 Evaluation Metrics

There are many evaluation criteria available to evaluate the PPI prediction models such as accuracy, precision, recall (also called sensitivity), MCC Score, F1-score, AUPRC Score [65], and AUC-ROC Score [66] as discussed below -

$$Accuracy(ACC) = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.1}$$

$$Precision(PR) = \frac{(TP)}{(TP + FP)} \tag{2.2}$$

$$Recall/Sensitivity(RE) = \frac{(TP)}{(TP + FN)} \tag{2.3}$$

$$MCCscore(MCC) = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.4}$$

$$F1Score(F1) = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{2.5}$$

TP represents the count of positive examples correctly predicted as positive
FP represents the count of negative examples wrongly predicted as positive
TN represents the count of negative examples correctly estimated as negative
FN represents the count of positive examples wrongly estimated as negative.

2.3.2 Performance on *S.cerevisiae* Dataset

We evaluated the performance of EnMAS on the *S.cerevisiae* dataset [23] and compared it with DensePPI [33]. Our experimental study underscores that the encoding strategies based on the BLOSUM and PAM substitution matrices significantly outperform DensePPI. This improvement is evidenced in Comparison Table 2.2. Among all tested matrices, PAM 120 (*EnMAS:PAM120*) achieved the best results, showing a 1.35% increase in the AUC-ROC score compared to the DensePPI model (see Fig 2.2A). Conducting both training and testing on the same species reinforces the homology hypothesis, which significantly influences PPI

Table 2.2: Performance comparison of DensePPI and EnMAS on *S.cerevisiae* dataset. The best values among the BLOSUM and PAM matrices are highlighted in bold.

Model	AUC-ROC	AUPRC	ACC	PR	RE	F1
DensePPI [33]	0.9578	0.9421	0.8958	0.8548	**0.9513**	0.9005
EnMAS:BL45	0.9683	0.9620	0.9131	0.8864	0.9476	0.916
EnMAS:BL90	0.9698	0.9664	0.917	0.8909	0.9502	0.9196
EnMAS:PAM120	**0.9713**	**0.9671**	**0.9173**	**0.8982**	0.9412	**0.9192**
EnMAS:PAM250	0.9702	0.9647	0.9166	0.9041	0.9321	0.9179

prediction. The findings validate this hypothesis, with *EnMAS:PAM120* notably outperforming *EnMAS:PAM250*, and *EnMAS:BL45* exceeding the performance of *EnMAS:BL90*. Importantly, across the BLOSUM and PAM variants, EnMAS consistently exhibited superior performance compared to the previously leading DensePPI model [33].

2.3.3 Performance on $JUPPI_d$ Dataset

For performance evaluation on the human PPI dataset $JUPPI_d$, we employed the same CV strategy and data as previously reported in [35] for EnMAS. Notably, JUPPI(S) employs an *n-gram*-based sequence feature where the amino acid sequences of input pairs are encoded into seven groups [69]. We compared the performance of four variations of EnMAS (*EnMAS:BL45*, *EnMAS:BL90*, *EnMAS:PAM120*, and *EnMAS:PAM250*) with JUPPI(S) and DensePPI. We have also presented comparisons against two more DL-based approaches, M1 [67] and M2 [68], utilizing the $JUPPI_d$ dataset. M1 adopts a sequence-based approach, utilizing a DNN classifier with a 7×7 MOS descriptors derived from seven distinct groups of reduced amino acid alphabets for PPI prediction. M2 employs the CT and sequence-based AC individually, as well as a combination of both (CT+AC), obtained from reduced amino acid groups. Each type of feature descriptor was processed using a DNN-based classifier with the two proteins represented in concatenated mode. The results are summarized in Table 2.3. Among the models evaluated, *EnMAS:BL90* achieved the highest performance in terms of AUC-ROC, scoring 0.9501 for test class C1 on the $JUPPI_d$ human protein dataset. In the more stringent test class C2, *EnMAS:PAM120* outperformed the other models.

In the challenging test classes C1 and C2, the bio-inspired colour encoding strategy significantly outperformed DensePPI's [33] random colour encoding, with AUC-ROC score improvements of 5.86% and 6.67%, respectively. For test class C3,

Table 2.3: Performance comparison of JUPPI(S), DensePPI, BLOSUM, and PAM-based EnMAS on $JUPPI_d$ human dataset. The best values are highlighted in bold.

TC	Model	PR	RE	F1	AUC-ROC
C1	JUPPI (S)[35]	NA	NA	0.763	0.826
	M1(DNN-MOS) [67]	NA	NA	NA	0.842
	M2(DNN-AC) [68]	NA	NA	NA	0.825
	M2(DNN-CT) [68]	NA	NA	NA	0.850
	M2(DNN-CT+AC) [68]	NA	NA	NA	0.859
	DensePPI [33]	0.5271	0.9970	0.6897	0.8915
	EnMAS:BL45	0.5567	0.9966	0.7144	0.9310
	EnMAS:BL90	0.5699	0.9983	0.7256	**0.9501**
	EnMAS:PAM120	0.5459	0.9945	0.7049	0.9265
	EnMAS:PAM250	0.5538	0.9957	0.7117	0.9346
C2	JUPPI (S) [35]	NA	NA	0.68	0.696
	M1(DNN-MOS) [67]	NA	NA	NA	0.680
	M2(DNN-AC) [68]	NA	NA	NA	0.729
	M2(DNN-CT) [68]	NA	NA	NA	0.715
	M2(DNN-CT+AC) [68]	NA	NA	NA	0.739
	DensePPI [33]	0.6025	0.9748	0.7447	0.7683
	EnMAS:BL45	0.5665	0.9837	0.7190	0.7572
	EnMAS:BL90	0.5865	0.9792	0.7336	0.7691
	EnMAS:PAM120	0.5998	0.9792	0.7439	**0.8350**
	EnMAS:PAM250	0.5744	0.9940	0.7281	0.8321
C3	**JUPPI (S)**[35]	NA	NA	0.673	**0.641**
	M1(DNN-MOS) [67]	NA	NA	NA	0.589
	M2(DNN-AC) [68]	NA	NA	NA	0.510
	M2(DNN-CT) [68]	NA	NA	NA	0.534
	M2(DNN-CT+AC) [68]	NA	NA	NA	0.507
	DensePPI[33]	0.484	0.9985	0.6519	0.5016
	EnMAS:BL45	0.4839	0.9992	0.6520	0.5635
	EnMAS:BL90	0.484	0.9996	0.6522	0.5273
	EnMAS:PAM120	0.4841	0.9996	0.6523	0.6125
	EnMAS:PAM250	0.4839	0.9992	0.6520	0.6015

N.B.: S: Sequence. TC: Test class type. Bold scores represent the best results. NA: Not Available.

PAM 120 emerged as the leading variant of EnMAS (*EnMAS:PAM120*), achieving an AUC-ROC score of 0.6125, thereby outperforming DNN-MOS, DNN-AC, DNN-CT, DNN-CT+AC, and DensePPI. However, it fell short of surpassing JUPPI(S).

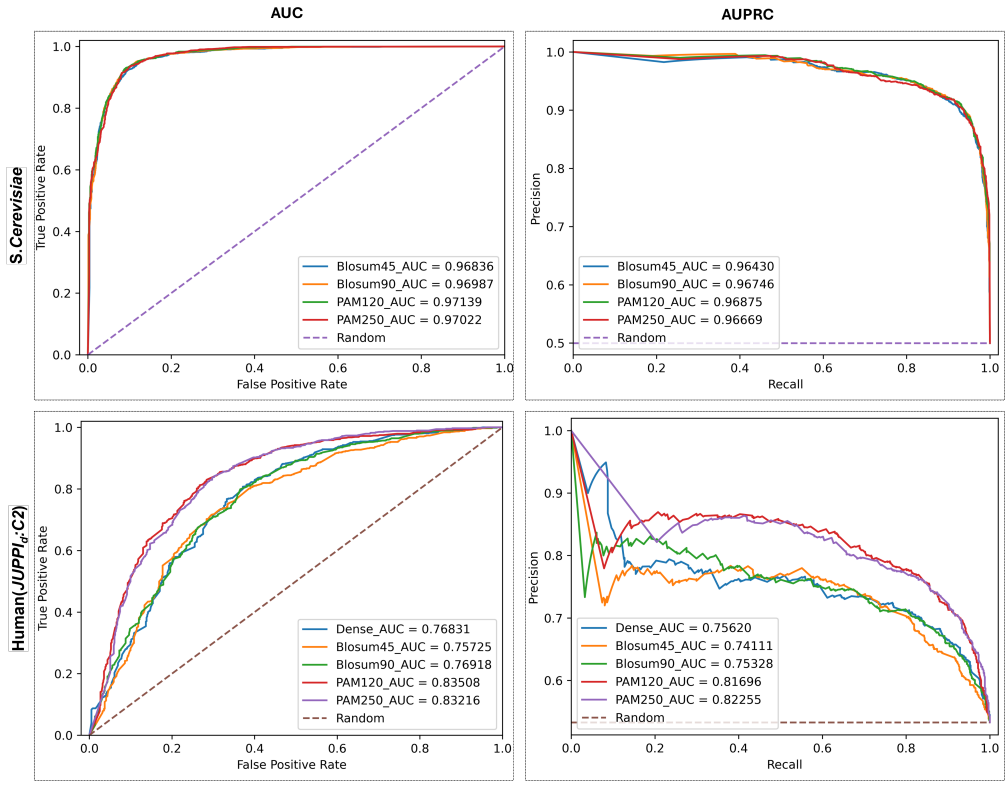


Figure 2.2: Performance comparison of *EnMAS* on Yeast (*S.cerevisiae*) PPI dataset and Human PPI dataset ($JUPPI_d$: C2).
N.B.:Row-wise plots represent the dataset-specific analyses and columns depict AUC-ROC curves and AUPRC on Yeast proteins (*S.cerevisiae*).

The AUC-ROC and AUPRC curves for C2 depicting the highest improvement are shown in Fig 2.2B. Moreover, in the domain of PPI prediction, the C2 type model plays a significant role in host-virus interaction prediction, particularly when information about the unknown virus is limited, as evidenced by the recent COVID-19 outbreak [70]. In this human PPI dataset $JUPPI_d$, EnMAS strongly reinforces the homology hypothesis, which significantly affects the accuracy of PPI predictions.

2.3.4 Application of EnMAS on Human-SARS-COV-2 interaction

We applied our EnMAS model to predict interactions between human hosts and the SARS-COV-2 virus. The reviewed SARS-COV-2 proteins and their interactions with humans are sourced from the UniProtKB database [71], which includes interactions substantiated by various existing experimental protocols. We retrieved data detailing 269 interactions that involved 15 SARS-COV-2 proteins and 252 human host proteins. We utilized 224 of these interactions for this experiment, excluding the remaining ones due to the insufficient sequence lengths of the

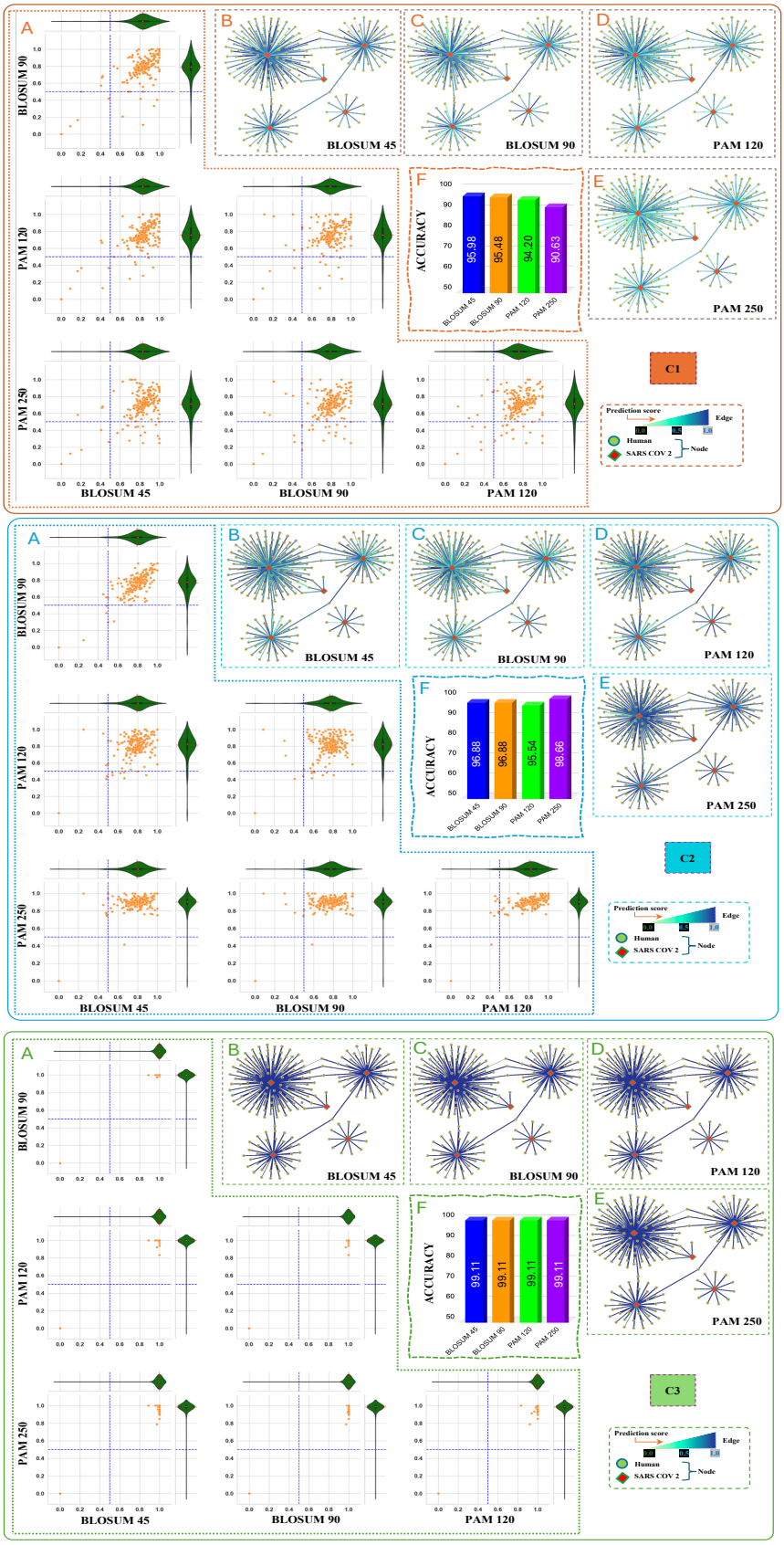


Figure 2.3: Performance assessment of *EnMAS* with four variations (BLOSUM 45, BLOSUM 90, PAM 120, and PAM 250) on *SARS-COV-2: Human* interactions using C1, C2, and C3 type models. The top, middle, and bottom blocks represent the performances of the C1, C2, and C3-based models, respectively.

proteins involved, from both the human hosts and SARS-COV-2. In this analysis, we applied all four EnMAS models to the test set for prediction purposes. The prediction performance details are illustrated in Fig 2.3. The C1-based EnMAS model with BLOSUM (45 and 90) shows better accuracy compared to PAM (120 and 250) (see Fig 2.3F, top block). In contrast, for C2, PAM 250 achieves a higher accuracy score (Fig 2.3F, middle block), surpassing the best C1 model. Notably, the highest accuracy score is observed with the C3-based model (Fig 2.3F, bottom block). This dataset is equally strict, similar to the C2 and C3 types, which is expected since the models are built with human protein data, and in the test set, one interactor (human) of the paired-input data point is known, while the other (SARS-COV-2) remains unknown. Consequently, our model is more suitable for this stricter dataset, as reflected in its performance.

We present a comparative performance assessment across all variations of EnMAS (BLOSUM 45, BLOSUM 90, PAM 120, and PAM 250) using JointGrid plots, which show the correlation-based distribution of predicted scores for each pairwise combination of models (see Fig 2.3A: top for C1, middle for C2, and bottom for C3). This assessment is further supported by a violin plot that provides a comprehensive view of the distribution of prediction scores for individual models. For example, in C1, the top plot in Fig 2.3A visualizes the comparative scores of BLOSUM 45 (x-axis) versus BLOSUM 90 (y-axis). A network representation of the predicted interactions from the C1, C2, and C3 models is shown in Fig 2.3 B to E for BLOSUM 45, BLOSUM 90, PAM 120, and PAM 250, respectively. In each network, the width and colour of the edges are proportional to the predicted scores, with darker edges indicating stronger interactions. Circular nodes represent human host proteins, while diamond-shaped nodes represent SARS-COV-2 proteins. From Fig 2.3F representing the bar plot of accuracy for each model, it is evident that the predictions in C3 are the most accurate, approaching closer to the value of 1, followed by C2, and then C1.

2.4 Discussion

This chapter introduces a novel approach EnMAS utilizing CNN to predict PPIs effectively through a unique sequence-to-image method incorporating biological and evolutionary information. Unlike the random colour encoding strategy used in DensePPI [33], bio-inspired evolutionary profile-specific substitution matrices such as BLOSUM 45, BLOSUM 90, PAM 120, and PAM 250 have been leveraged to generate the colour map in our proposed colour encoding scheme for sequence representation. Initially, the LUTs were constructed based on the LOD score values of the substitution matrices. For each pair of amino acids in both the

interacting and non-interacting protein sequences, the corresponding LUT value is utilized to generate the pixel value in RGBA. In this fashion, the protein pairs are transformed into 2D colour images using the substitution score from each cell of these matrices. Each original image is further divided into sub-images used to train the DenseNet model by the horizontal and vertical traversal of an optimum fixed-sized window over the original image.

We validated our bio-inspired colour encoding approach with two benchmark PPI datasets on the yeast (*S.cerevisiae*) and $JUPPI_d$. Table 2.2 highlights the significant performance enhancements achieved by EnMAS over DensePPI in predicting PPIs on the *S.cerevisiae* dataset. The evaluation on the $JUPPI_d$ human protein dataset is outlined in Table 2.3 depicting the notable performance achievements of EnMAS across various difficulty test classes C1, C2, and C3, showcasing significant improvements in AUC-ROC score over DensePPI's random colour encoding. Particularly, *EnMAS:BL90* demonstrated the highest AUC-ROC score in test class C1, while *EnMAS:PAM120* exhibited superior performance in the more stringent test class C2. Additionally, in test class C3, *EnMAS:PAM120* has surpassed all *state-of-the-art* models but JUPPI(S). It has also been observed in both yeast and human datasets that our bio-inspired sequence encoding strongly reinforces the homology hypothesis, significantly improving the accuracy of PPI predictions. In almost every case of prediction, when considering BLOSUM substitution matrices, BLOSUM 90 performs better than BLOSUM 45. Similarly, PAM 120 shows superior performance compared to PAM 250. This is expected, as higher thresholds in BLOSUM and lower thresholds in PAM indicate greater homogeneity among the sequences.

EnMAS offers a promising approach for identifying protein interactions, making it particularly valuable in scenarios such as virus outbreaks and drug development for diseases like COVID-19. To demonstrate its efficacy, EnMAS was also employed on the human-SARS-COV-2 dataset, as host-virus interactions belong to the C2 type in the pair-input-based complex test classes. More interestingly, it was observed that EnMAS was able to identify interactions with accuracy in the order of $C1 < C2 < C3$ across all four variations of the model.

The next study in chapter 3 aims to consider the efficacy of a novel topological feature to predict protein interactions contrary to the sequence-based methodology discussed in this chapter where a graphlet-based approach 'GSPPI' has been designed for PPI prediction from the HIPPIE PPIN.

CHAPTER 3

Prediction of PPIs using graphlet feature on HIPPIE dataset

3.1 Introduction

In the previous chapter 2, a novel bio-inspired approach for predicting PPIs using a CNN-based architecture by translating the protein interactions into images has been presented. This chapter presents a GNN-based graphSAGE model 'GSPPI' that has been leveraged to predict the interactions between proteins by utilizing protein graphlet features (up to 5 nodes), a protein node's local network topological property.

Systematic evaluations of the structure, also known as topology, in large networks, have been introduced in the works of [72]. These assessments rely on graphlets, which are small induced subgraphs of the larger networks [73]. Subsequently, a new graph-theoretic approach was introduced in a study conducted by [74] to analyze the relationship between local topology and function in real-world networks. The study provided evidence that the graphlet representation of a PPI network is highly relevant for predicting protein function. The study [75] has revealed the identification of similarities in the local topological structures of networks by analyzing repeating sub-patterns within them. This analysis utilizes a similarity measure generated from the distribution of graphlet frequencies. Specifically, they capture the local topology of proteins within a PPI network and represent it as a vector of graphlet degrees. This vector is a signature of a protein's node that describes the network similarities between all protein pairs.

Recognizing the importance of graphlets in topological representation, a novel

approach has been devised combining topological feature information and another inspiring GNN approach GraphSAGE [32] to predict PPI. The rest of the chapter is organized as follows. Section 3.2 describes the methodology like positive and negative data preparation from the HIPPIE database, feature extraction, and the classification strategy. The results of the experiments are detailed in section 3.3 which includes the evaluation metrics, the experimental setup, the performance of our proposed model GSPPI, and its comparison with other GNN variants. Section 3.4 concludes the chapter.

3.2 Methodology

This chapter focuses on a graph-based DL approach utilizing graphlet as a network topological feature. Graphlets, introduced by Pržulj [72] are small, highly linked induced subgraphs in a graph network. Each graphlet exhibits the automorphism property of a graph, indicating that the graph possesses an isomorphic structure inside itself. The orbit of an automorphism in a graph refers to the partitioning of the vertices that represent the symmetrical configuration of the graph. A 73-dimensional feature vector termed 'orbit-counts' has been considered to depict the frequency of a specific node's occurrence within all orbits across 5-node graphlets. This vector represents a node's positional characteristics within the local network structure, offering insights into its properties. In Fig 3.1, few graphlets out of 30 (G_0 to G_{29}) [72] [73] with 2 to 5 nodes have been displayed. It can be seen that G_0 comprises only one orbit, as does G_2 since all the node positions are equivalent. But G_1 exhibits two orbits since the node positions of the end nodes and middle node are not equivalent. In this fashion, overall 73 orbital positions are present. In our proposed methodology, these 73-dimensional orbit counts have been leveraged as a topological node feature vector for predicting PPIs.

3.2.1 Positive data selection

Our approach is implemented using a PPI network obtained from the HIPPIE database[76], containing 5,497 proteins linked by 13,163 interactions. This PPI network is represented as a graph, where proteins act as nodes and interactions between proteins form edges. The 13,163 interactions within the PPI network are utilized as positive data for the analysis.

3.2.2 Negative data selection

Obtaining a reliable negative dataset is crucial but more challenging than assembling a positive dataset because information about non-interacting proteins

is generally unavailable. In our selection of the negative dataset, negative edges have been selected by choosing all possible pairs of proteins that do not have an interacting edge between them. For example in Fig 3.1, as protein 'A' interacts with proteins 'B' and 'E' and proteins 'D' and 'C' do not have any link with proteins 'A', 'B' and 'E', therefore 'A-C', 'A-D' forms negative interactions. With this concept of negative sampling, an equal number of negative edges for training the model have been generated.

Subsequently, the pairs of interacting proteins and those of non-interacting proteins are combined to form an all-encompassing dataset. This comprehensive dataset comprises a total of 2×13163 protein pairs, maintaining a balanced ratio of 1:1 between the negative and positive datasets.

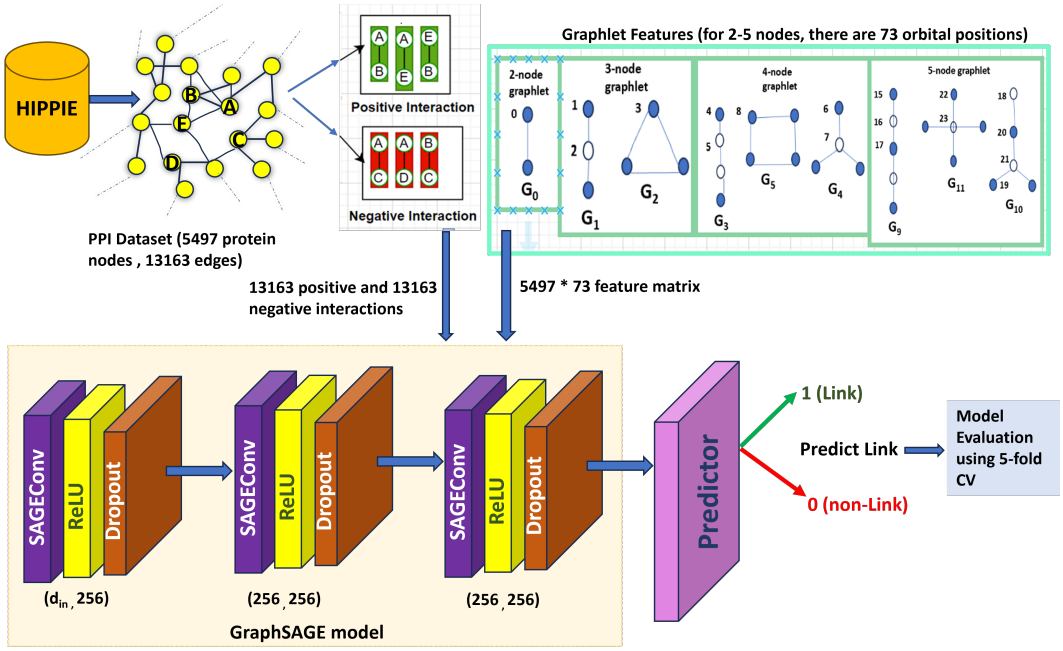


Figure 3.1: Basic workflow of the proposed model GSPPI.

A) Benchmark protein interactions are taken from the HIPPIE database. B) From the PPI network graph (where A, B, etc. are proteins) formed, positive and negative datasets are generated. C) Graphlet features: The graphlet orbit-counts for each node have been computed, resulting in a 73-dimensional feature vector concerning all 30 graphlets from G_0 to G_{29}which involves up to 5 nodes. D) Positive and negative interactions with graphlet orbit-counts as feature vector is trained using the graphSAGE model and tested using 5-fold CV.

3.2.3 Feature Extraction

The computation of the graphlet properties considering the 5 different nodes at orbital positions resulting in 73 combinations of sub-networks or graphlets is done using the ORCA[77] which is an efficient algorithm for computing orbit counts of

5-node graphlets for each node in the PPIN. It is designed as a command-line tool, accepting arguments such as orbit type and graphlet size (set to 5 in our scenario), along with an input file specifying the network in a basic text format. The 73 orbit-counts have been computed for each node, considering 5-node graphlets resulting in a feature matrix of size 5497 $\times$ 73 for the 5,497 proteins where the rows correspond to protein nodes within the PPI network and the columns represent the elements of the 73-dimensional orbit-counts feature. The two-node combinations are used as training samples. For each combination, 73-dimensional orbit-count features are extracted for each associated protein node resulting in a 146-dimensional feature vector (2 $\times$ 73) representing each protein pair interaction/ non-interaction.

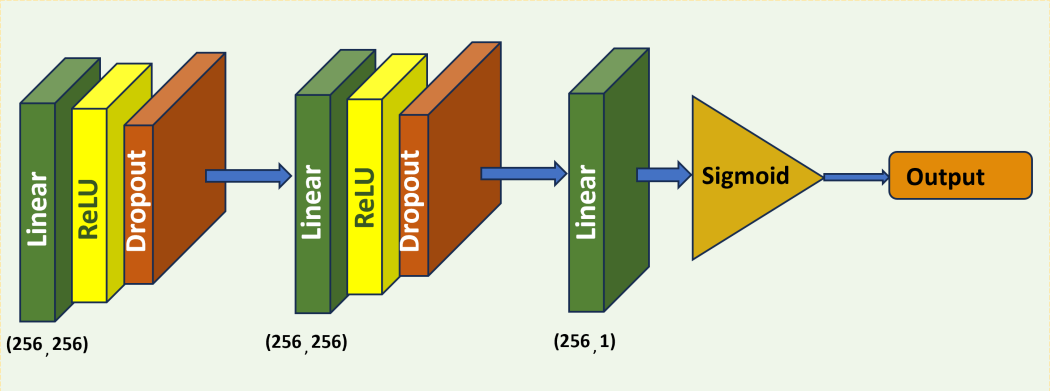


Figure 3.2: Basic schematic diagram of link predictor of the GSPPI model.

3.2.4 Training and Classification

GraphSAGE, an inductive representation learning technique is used for training purposes. It aims to transform nodes in a graph into a low-dimensional vector embedding while preserving their structural properties. Firstly, it performs random neighbor sampling for each node by sampling a fixed number of neighbors which helps in capturing local neighborhood information. Secondly, it aggregates information from the sampled neighbors to compute the final node embedding. It can predict embeddings for new nodes using learned aggregator functions without re-training. This makes it particularly suitable for graphs with diverse sizes and architectures. It has shown promising results in many graph-based applications, particularly in link prediction [78]. The training module consists of three layers of convolution with 256 units each and uses the activation function as ReLU in each layer. The SageConv (graphSAGE variant) aggregate function is the mean aggregator. It utilizes the average of neighbors' representations, normalized by each neighbor's degree, to form the aggregate representation. During the model's training phase, the Adam optimizer with a learning rate of 0.002 has been employed and a weight decay of 0.00001 is used to minimize the MSE loss. After

each layer, a dropout layer with a dropout rate of 0.002 has been incorporated to avoid over-fitting. The selection of these hyperparameters, including dropout rate, learning rate, the number of hidden layers, and the number of units in each layer, has been determined based on previous research findings.

Determining the existence of interaction between two protein nodes can be considered a binary classification problem. Link predictor has been designed using three linear layers with ReLU activation function and a dropout for minimizing over-fitting in each layer. Finally, it uses the sigmoid activation function to predict the edges. Positive edges are labeled as "1" and negative edges as "0". The model is trained using the aforementioned training data with both positive and negative datasets. The final class of the hold-out data is determined by model output with a threshold value of 0.5. If the probability output is greater than 0.5, then it is considered as "1" or belonging to the positive class otherwise belonging to the negative class. Fig 3.1 gives a schematic diagram of how the model with a 5-node graphlet node feature is used to predict PPI and the design of the link predictor has been shown in Fig 3.2.

3.3 Results

This section covers the performance of the proposed model GSPPI, the performance of GSPPI by varying the number of SageConv layers, and subsequently, a comparison with other GNN models.

3.3.1 Performance of the proposed model GSPPI

To gauge the model's performance, the dataset has been split in a way that allocates 70% of it for training, 20% for the validation set, and reserving the remaining 10% for testing. The various performance statistical metrics of our proposed model in the training dataset, validation, and test dataset have been illustrated in Table 3.1. The 5-fold CV procedure has been used to train our proposed model on the 70% positive and negative dataset i.e. 9,214 positive interactions and 9,214 negative interactions. Five predictive models have been generated through the training phase, and their performances have been assessed by validating the hold-out dataset. Results of the 5-fold CV of the GraphSAGE model for every fold using Graphlet node features for the HIPPIE dataset have been compiled in Table 3.2. Based on the comprehensive metric scores presented in both Table 3.1 and Table 3.2, it is evident that our classification model's performance in predicting PPI within the HIPPIE dataset, utilizing graphlet features, is auspicious.

Table 3.1: Performance scores of our proposed model on the training dataset, validation dataset, and testing dataset

Dataset	**PR(%)**	**RE(%)**	**F1(%)**	**ACC(%)**	**MCC(%)**	**AUC-ROC(%)**
Training Set	92.42 ± 2.91	96.55 ± 2.2	95.47 ± 1.39	94.75 ± 2.04	90.86 ± 2.8	94.75 ± 2.04
Validation Set	83.01 ± 0.74	84.92 ± 0.35	83.94 ± 0.54	83.17 ± 0.59	67.53 ± 1.97	53.76 ± 0.59
Test Set	84.36 ± 0.87	83.36 ± 2.36	79.87 ± 1.56	80.90 ± 1.33	62.13 ± 2.57	80.90 ± 1.33

Table 3.2: Performance scores of our proposed model on the hold-out dataset for 5-fold CV. The best values are in bold.

Fold No.	**PR(%)**	**RE(%)**	**F1(%)**	**ACC(%)**	**MCC(%)**	**AUC-ROC(%)**
1	90.85	91.10	90.98	90.97	81.93	90.97
2	93.21	93.87	93.54	93.52	87.03	93.52
3	94.45	97.02	95.72	95.66	91.35	95.66
4	94.65	95.12	94.88	94.87	89.75	94.87
5	**96.45**	**97.29**	**96.86**	**96.85**	**93.71**	**96.85**

3.3.1.1 Effect of number of layers on the performance of our proposed model

The number of layers is a critical parameter that impacts the performance of DL models [79]. Generally, it is assumed that having more GNN layers enables the model to capture a richer set of information from both node and edge features. However, more layers can lead to performance degradation due to over-smoothing [80] and vanishing gradients. To illustrate the impact of the number of layers, experiments are conducted with up to four GraphSAGE layers including 5-fold CV within our proposed PPI framework, and the results are presented in Table 3.3 and Table 3.4 respectively. It can be observed that the performance metrics scores are improving from 1 to 3 layers and slightly decreasing post that except for recall. The best results have been achieved considering 3 layers. Hence, the number of convolutional layers considered in our proposed model is 3.

Table 3.3: The results of our proposed model on test dataset with different number of layers for PPI datasets. The best values are in bold.

# **Layers**	**PR(%)**	**RE(%)**	**F1(%)**	**ACC(%)**	**MCC(%)**
1	82.01	60.59	69.69	73.65	49.01
2	**77.99**	76.69	77.34	77.52	55.06
3	77.94	83.98	**80.85**	**80.11**	**60.39**
4	71.81	**89.75**	79.78	77.26	56.30

Table 3.4: The results of the average 5-fold CV of our proposed model with different numbers of layers on the PPI dataset. The best values are in bold.

# **Layers**	**PR(%)**	**RE(%)**	**F1(%)**	**ACC(%)**	**MCC(%)**
1	90.12	90.33	90.20	90.19	80.43
2	89.44	93.31	91.30	91.11	82.36
3	**91.94**	**95.52**	**93.69**	**93.54**	**87.16**
4	83.86	93.92	88.14	86.74	75.04

3.3.2 Comparison of the proposed model with other GNN variants

GNNs represent a notable advancement in the realm of DL. This advancement has garnered significant attention from researchers in recent years. Our experimental study has been aimed to benchmark our work against the prevailing *state-of-the-art* GNN variants: GCN and GAT. A comparative analysis has been performed using the 5-node graphlet feature to assess their relative performances keeping the experimental setup identical to that of GSPPI. Fig 3.3 shows the analysis of different performance metrics such as PR, RE, F1 Score, ACC, AUC-ROC Score, and MCC Score for the three models GCN, GAT, and GraphSAGE. The colour blue, amber, and green represent the GCN, GAT, and the proposed graphSAGE-based model respectively. It is apparent from the results that the proposed GraphSAGE model for the HIPPIE dataset improves the results significantly more than its GAT and GCN counterparts in terms of all evaluation metrics. But our proposed model using graphSAGE has out-performed these GNN variants as evidenced by the results of the hold-out set with PR, RE, F1, ACC, AUC-ROC Score, and MCC of 0.8436, 0.8336, 0.7987, 0.809, 0.809 and 0.6213 respectively.

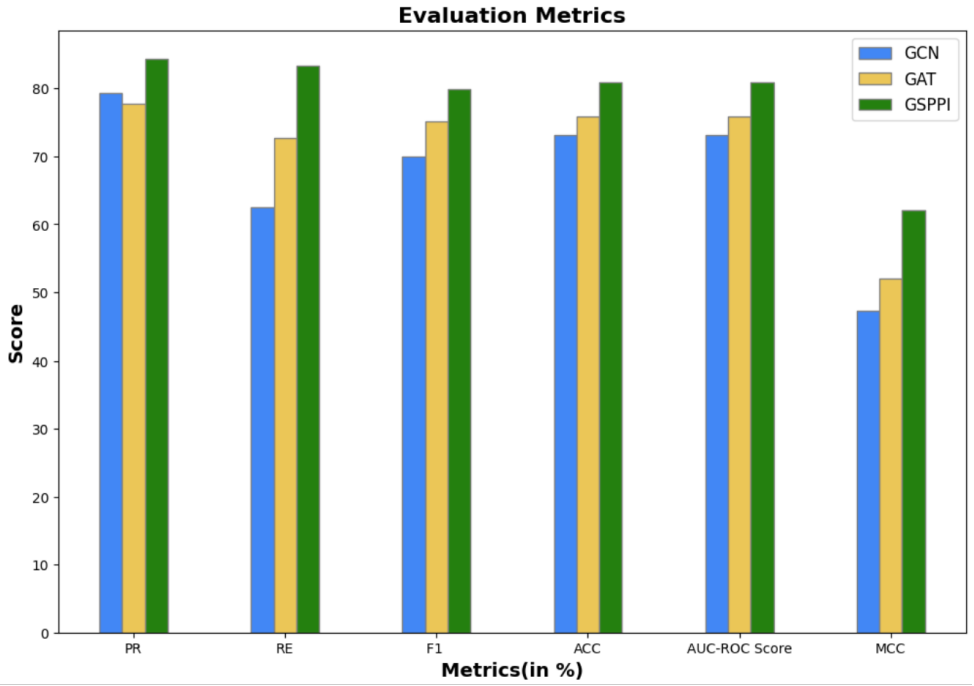


Figure 3.3: Performance metrics of three GNN variants using graphlet feature on HIPPIE dataset.

3.4 Discussion

This chapter introduces a novel inductive graphSAGE-based DL methodology called GSPPI for predicting PPI pairs using protein's discriminative graphlet features (2-5 nodes). All available PPI data has been obtained from the HIPPIE dataset. GraphSAGE, a variation of the GNN model, functions by selectively extracting and combining data from a node's immediate surroundings. It leverages the 73-dimensional orbit counts as a topological property for each protein node in a PPIN to predict PPIs. After completing the training process, the model utilizes sigmoid activation for predictions, classifying test data as positive or negative interaction based on a threshold of 0.5. Performance of GSPPI along with a 5-fold CV demonstrating the model's efficiency have been illustrated in Table 3.1 and Table 3.2. Comparative experiments with other GNN variants like GCN and GAT using similar graphlet characteristics in Fig 3.3 highlight the proposed model's robustness and effectiveness of a topological feature in distinguishing between interacting and non-interacting protein pairs. The next chapter discusses an ensemble method combining graphlet features with protein sequence composition-based features for PPI prediction on the HIPPIE dataset.

CHAPTER 4

Prediction of PPIs using ensemble features on HIPPIE dataset

4.1 Introduction

In the chapter3, we discussed the graphlet feature (up to 5 nodes) for PPI prediction. Appreciating the significance of graphlets in topological representation, we have suggested a novel methodology, integrating graphlets and sequence-based features for the targeted objective of predicting PPIs. The hybrid representation combines the topological context recorded by graphlet features with the biochemical insights provided by amino acid composition features. The goal is to effectively utilize the distinct information from each feature set, leading to a stronger and more distinguishable representation of features. The rest of the chapter is organized as follows. Details of the methodology including data selection, feature extraction techniques, deep neural network model, and the classification strategy are described in Section 4.2. Section 4.3 details the results. Finally, the overall discussion on this chapter is given in Section 4.4.

4.2 Methodology

This work uses the ensemble feature to construct a strong classifier for predicting PPIs. The ensemble feature comprises graphlet property and a range of composition-based characteristics, as detailed in the subsequent feature extraction section 4.2.2. The general process of data selection and feature creation is illustrated in Fig 4.1.

4.2.1 Data Selection

The positive and negative data selection is similar to that of section 3.2.1 and 3.2.1 where the PPI network data is acquired from the HIPPIE database consisting of 5,497 proteins connected by 13,163 interactions. The 13,163 interactions observed among the protein nodes in the described PPI network are regarded as positive data. Negative edges are generated by selecting potential protein pairs lacking interaction, employing negative sampling to create an equal number of negative edges for model training.

4.2.2 Feature Extraction

Our research presents a novel method for combining graphlet features with sequence-based features for feature extraction to improve the model's ability of distinguishing between different options by combining local structure information using graphlet features and the underlying physicochemical properties of proteins using amino acid composition features. Therefore, the process of generating features may be stated in three steps:

1. **Graphlet Feature extraction**

   The concept of a network's graphlet feature and its extraction have already been discussed in the previous chapter 3.2.3. The node characteristics are extracted by calculating 73-dimensional orbit counts for each node, considering 2-5 node graphlets using ORCA. With 5,497 protein nodes in the HIPPIE PPIN, the resulting feature vector size is 5497 $\times$ 73, representing the frequency of appearance for every 5,497 proteins, acting as a node in the protein interaction network, in every orbital position (0 - 72) across 5-node graphlets in the feature matrix.

2. **Sequence-based Feature Extraction**

   The amino acid composition features provide a thorough overview of the protein content by calculating the percentage of each residue type from the amino acid sequence. The Pfeature tool [81], a web server capable of calculating a broad spectrum of protein/peptide features based on their amino acid sequences, is utilized to produce 976 distinct features based on composition. The utilization of this numerical encoding of a protein allows us to represent proteins in sequences of different lengths using a standardized collection of features. The amino acid sequences for the 5,497 proteins of the HIPPIE network are acquired from the UniProt [82] database. These

sequences were then inputted into the Pfeature server to construct the following composition-based features:

- Simple compositions: AAC, DPC, ATC, and BTC.
- Physico-chemical compositions: PCP.
- Compositions involving repeats and distribution of amino acid: RRI and DDR.
- Shanon Entropy-based Compositions: SEP and SER.
- Miscellaneous Compositions: CTC, PAAC, APAAC, QSO, and SOC.

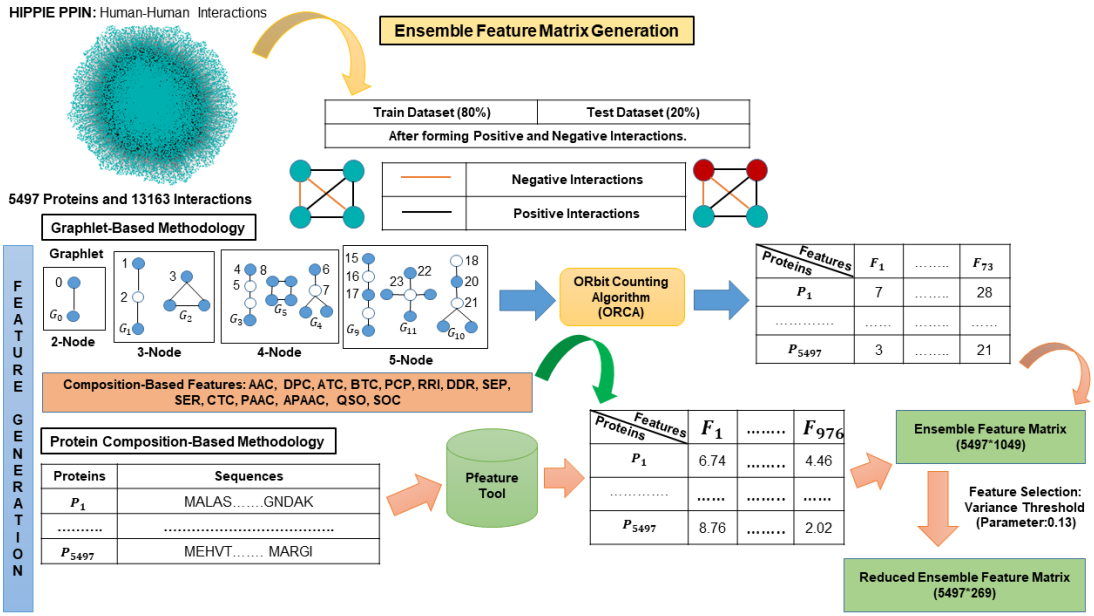


Figure 4.1: Basic workflow of the ensemble feature generation method on HIPPIE dataset.

A) Benchmark PPIN data have been taken from the HIPPIE Database comprising 5,497 nodes and 13,163 interactions. B) Positive Interactions: Protein pairs that are directly interacting in the PPIN. Negative Interactions: Combination of protein pairs that do not have an edge between them. C) Graphlet feature: Depending on the interactions in the PPI database, the 73-dimensional graphlet orbit-count features have been computed for 5,497 protein nodes using the ORCA algorithm. D) Composition-based feature: Based on the peptide sequences of the proteins, 976 different compositions have been generated for 5,497 protein nodes. E) Ensemble feature: The graphlet feature matrix has been combined with the sequence-based feature matrix to generate a 1,049-dimensional feature matrix for all 5,497 protein nodes. F) Feature selection: Variance threshold feature selection method has been applied on the ensemble feature matrix with threshold parameter as 0.13 to generate a reduced feature matrix sized at 5497 $\times$ 269.

The Variance Threshold feature selection approach is applied to exclude features (variables) in the dataset that exhibit low variance. The reason

for using this strategy is that features with low variance typically exhibit minimal fluctuation throughout the dataset and may not make a substantial contribution to the predictive capability of a model. The variance threshold parameter has been determined using the equation shown below.

$$\text{Variance threshold (i)} = \max_{i \in [0.05, 0.20]} \text{AUC-ROC}(i) \tag{4.1}$$

Put simply, we have chosen the variance threshold 'i' within the range 0.05 to 0.20, based on the condition that it yields the highest AUC-ROC [66] on the test set.

To determine the most effective threshold variance, an experiment is conducted to assess the performance metrics of the model. The threshold is varied from 0.05 to 0.20, as seen in Fig 4.2-(a). Here, blue coloured bars represent the number of selected sequence-based features and the brown coloured line represents the AUC-ROC in percentage for various thresholds in feature selection. It is evident from the figure that a threshold of 0.13 yields the maximum AUC-ROC of 79.89% with a selection of 196 features. Therefore, the features with variance below the stated threshold are excluded and the rest are retained for further analysis or model training. Finally, since we have 5,497 protein nodes in our HIPPIE PPIN, the resulting amino acid composition feature matrix is sized at 5497×196.

3. **Ensemble Feature Extraction**

In the ensemble feature extraction procedure, the 73-dimensional 2-5 nodes graphlet features are merged, as previously stated, with the 976-dimensional composition-based features. This combination results in a 1049-dimensional feature vector for each protein node. Afterward, the Variance threshold feature selection approach is employed to choose relevant features from the pool of 1,049 features. The initial range is set from 0.05 to 0.20 and the AUC-ROC attained for each threshold value is plotted. The variance threshold has been chosen based on the same technique as the prior feature selection.

$$\text{Choose } i \text{ such that } i \in [0.05, 0.20] \text{ and AUC-ROC}_i \text{ is maximized.}$$

The plot of the AUC-ROC (in %) and number of features selected against each variance threshold within the range 0.05 to 0.20 has been represented in Fig 4.2-(b). Here, the colour blue in bars represents the number of selected features and the colour red in line represents the AUC-ROC in percentage for various thresholds in feature selection. It is apparent from this figure that

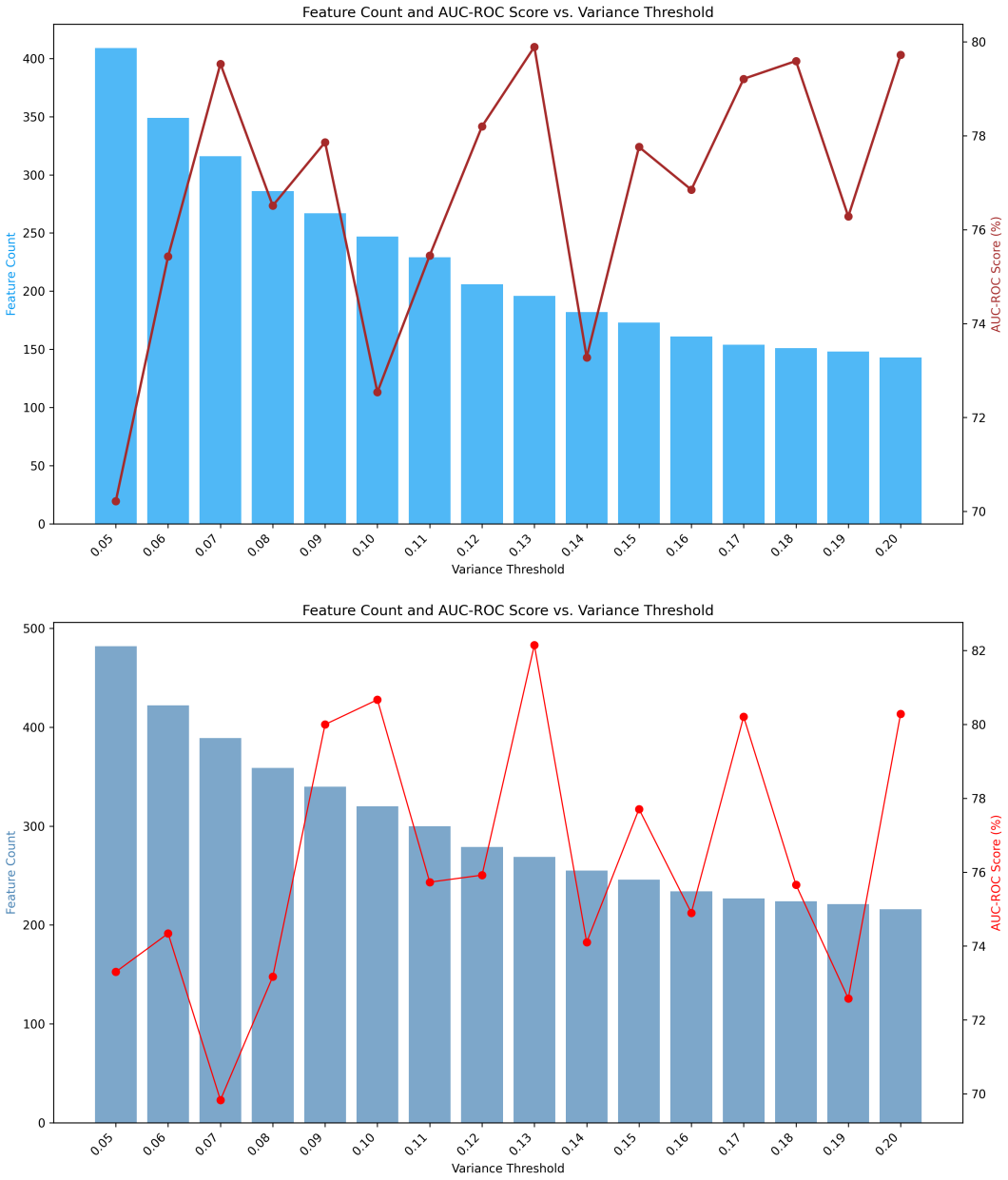


Figure 4.2: a)Plotting AUC-ROC (in %) and number of composition-based features selected against each threshold variance using Variance Threshold Feature Selection Method in HIPPIE PPIN. b) Plotting AUC-ROC (in %) and number of ensembled features selected against each threshold variance using Variance Threshold Feature Selection Method in HIPPIE PPIN.

a threshold of 0.13 gives the best AUC-ROC of 82.15% with 269 features being selected by the feature selection method. At last, a feature vector is obtained that consists of 269 dimensions. Thus, for 5,497 protein nodes in our PPIN, a feature matrix of size 5497 $\times$ 269 is generated to train our model.

4.2.3 Deep Neural Network Model

The GraphSAGE has already been introduced in the previous chapter 3.2.4. The SageConv architecture, a variant of the GraphSAGE architecture is employed, in our proposed work. The model has a convolutional operator that is more expressive, allowing it to capture more complex characteristics. The SageConv aggregate function that we have considered is the mean aggregator. It utilizes the average of neighbors' representations, normalized by each neighbor's degree, to form the aggregate representation. Fig 4.3 illustrates the suggested model and its corresponding classifier.

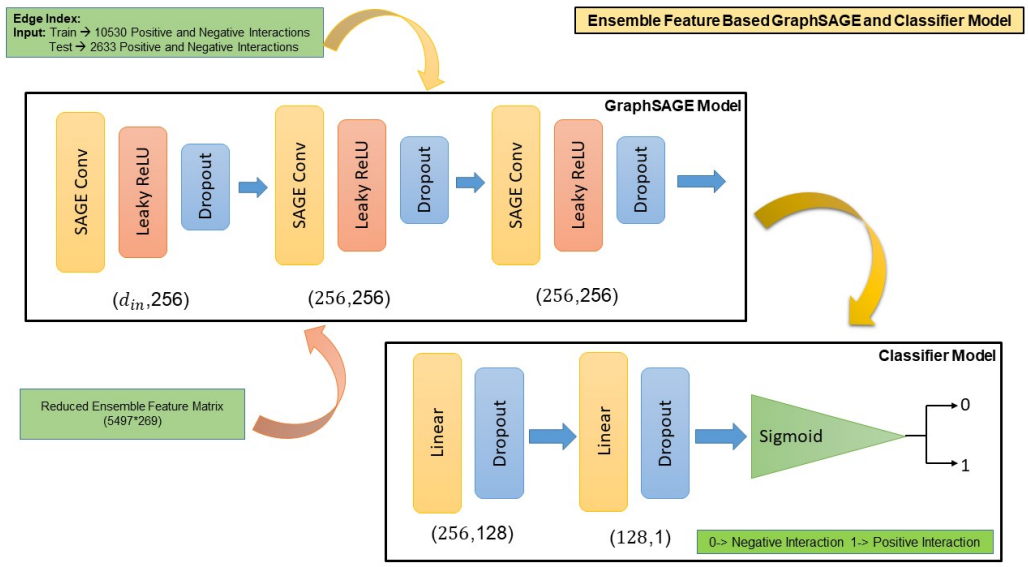


Figure 4.3: Basic schematic diagram of the ensemble feature-based model.

A) The generated feature matrix along with the edge indices have been given as input to our model. B) The graphSAGE-based model comprises three SAGEConv layers along with the Leaky ReLU activation function and dropout mechanism. The d_{in} in our proposed ensemble feature-based training model is 269. C) The output of the model has been provided to a classifier comprising two linear layers with dropout layers ending with a sigmoid activation function. D) Based on the confidence score, the likelihood of being in the positive or negative class has been determined.

Our model incorporates three SageConv layers, each consisting of 256 units. Each layer is equipped with a Leaky ReLU activation function and utilizes the dropout technique with a dropout rate of 0.003. The Adam optimizer with a learning rate of 0.002 is utilized throughout the model's training phase to reduce the MSE loss. This optimizer efficiently adjusts the model's parameters based on the gradient of the loss function, which is computed during the training phase. This contributes to enhancing the model's accuracy and reducing overall loss. The

optimizer implements a weight decay of 0.0001 as a regularization strategy, gradually diminishing the weights in each epoch and progressively approaching zero. This contributes to enhancing the model's AUC-ROC and reducing overall loss. The initial selection of these hyperparameters, such as the dropout rate, learning rate, weight decay rate, the number of hidden layers, and the number of units in each layer, is determined based on prior research and fine-tuned experimentally. The dataset consisting of 13,163 positive edges and 13,163 negative edges is partitioned into an 80% training set and a 20% test set for our experiments. Therefore, the training set has 10,530 positive edges and 10,530 negative edges, whereas the test set comprises 2,633 positive edges and 2,633 negative edges.

4.2.4 Classification Strategy

The predictor has two linear layers with 256 units and 128 units, each incorporating a dropout mechanism, to address the issue of overfitting. In the last layer, the sigmoid function is employed as the activation function to forecast the test edges. The predictor's output has been categorized based on the confidence values for each test edge, which range from 0 to 1, as either 0 or 1 as per Eq (4.2). A threshold of 0.5 was selected to assign the final class label to these edges. In the equation provided, $FC_i(t)$ represents the final class assignment for the i^{th} test edge, and the $Pred_i(t)$ represents the probability output score of the sigmoid activation function of the classifier. If the probability output score exceeds 0.5, it is assigned as "1" (classified as the positive class; otherwise, it is assigned as "0" (classified as the negative class).

$$FC_i(t) = \begin{cases} 1 & \text{if } Pred_i(t) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

4.3 Results

This section covers the performance of the proposed prediction model with different feature selection strategies and node features, and subsequently, a comparison with other graph-based DL models is also presented.

4.3.1 Performance of the proposed model with different feature selection strategies

Feature selection is a crucial step in data pre-processing that eliminates unnecessary and duplicated information that has little impact on enhancing the prediction performance of the predictor and lowering the model's complexity. A thorough examination of three widely used feature selection techniques: Fast ICA, PCA, and Variance Threshold with a threshold parameter of 0.13, is performed. The goal is to identify the most effective strategy in terms of feature selection performance. Given that the Variance threshold feature selection with a threshold value of 0.13 and 269 features has outperformed other threshold values, the same number of features is also selected for comparison purposes in all other feature selection methods. Although all approaches have used the same set of 269 features, Variance Threshold with a value of 0.13 showed higher discriminatory ability compared to FastICA and PCA in terms of selecting effective features, as seen in Table 4.1.

Table 4.1: Performance of our proposed model with FastICA, PCA, and Variance threshold feature selection method on HIPPIE dataset. The best values are in bold.

	FastICA	**PCA**	**Variance Threshold (0.13)**
PR	0.7721	0.7270	**0.8647**
RE	0.3988	0.6988	**0.7622**
MCC	0.3211	0.4367	**0.6476**
$F1$	0.5259	0.7126	**0.8103**
$AUC - ROC$	0.6405	0.7182	**0.8215**

4.3.2 Performance of the proposed model with different node features

The selected input characteristics greatly influence the performance of any model. To conduct a comparative study of our model, it is trained and tested using three separate feature sets. Initially, only graphlet features are used followed by only sequence-based features. Finally, both graphlet and sequence features are included in an ensemble mode. Table 4.2 demonstrates the performance of the proposed model with different node attributes. As reflected, the performance of the proposed model using the ensemble feature is superior to the performances

achieved by any individual type of feature, such as a graphlet or sequence-based feature.

Table 4.2: Performance scores of our proposed model with graphlet feature, sequence-based feature, and ensemble feature on the HIPPIE dataset. The best values are in bold.

	Graphlet Feature	**Sequence-based Feature**	**Ensemble Feature**
Precision	**0.9047**	0.8431	0.8647
RE	0.5735	0.7345	**0.7622**
MCC	0.5514	0.6028	**0.6476**
*F*1	0.702	0.7851	**0.8103**
AUC − ROC	0.7566	0.7989	**0.8215**

4.3.3 Comparison of the proposed model with other GNN variants

Our experimental study aimed to compare our work with the current *state-of-the-art* GNN variants, specifically GCN and GAT using 5-node graphlet features. Fig 4.4 shows the analysis of different performance metrics such as PR, RE, F1, AUC-ROC, and MCC for the three models GCN, GAT, and GraphSAGE represented by colours purple, yellow, and green respectively. The results indicate that our proposed model, utilizing graphSAGE, has shown significantly better performance compared to other variants of GNN. This is evident from the evaluation results of GraphSAGE where the PR, RE, F1, AUC-ROC, and MCC are 0.8647, 0.7622, 0.8103, 0.8215, and 0.6476, respectively.

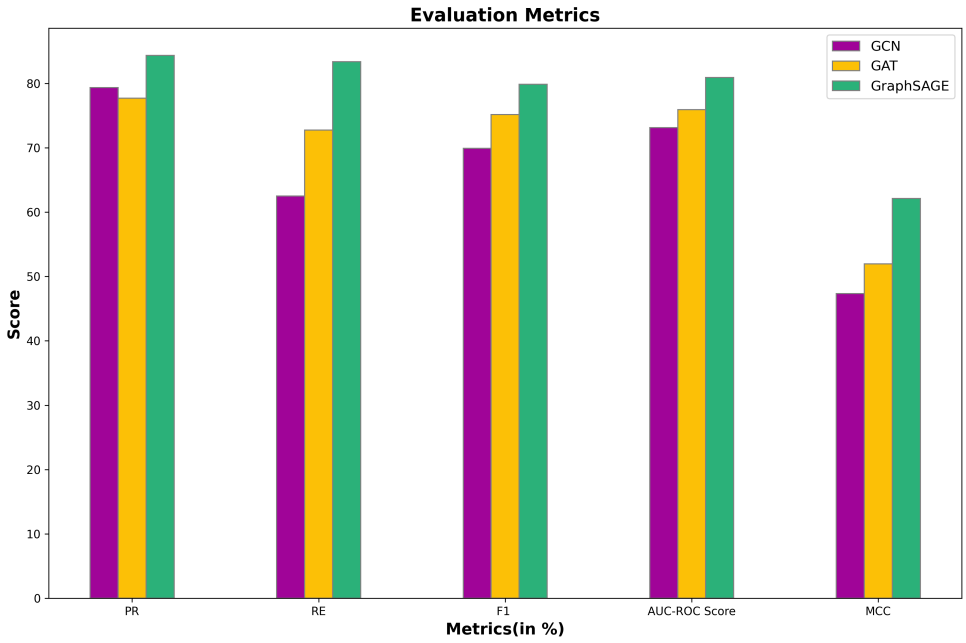


Figure 4.4: Performance metrics of three GNN variants using ensemble (graphlet and composition-based) features on HIPPIE dataset.

4.4 Discussion

This chapter presents yet another extension of the research conducted in the previous chapter whereby the proposed method emphasizes the significance of synergistic integration of the network topological property which is graphlet feature(up to 5 nodes) and inherent composition, physico-chemical sequence-based features of proteins for the prediction of PPI. The graphSAGE-based DL model has been trained using the ensemble feature set in the HIPPIE dataset. Using graphlet features enables the identification of complex connections among adjacent nodes, facilitating the detection of specific structural patterns. On the other hand, sequence-based features enhance the model's understanding of the biological properties of proteins, assisting in the differentiation between interacting and non-interacting pairs based on their biochemical characteristics. Table 4.2's ablation study reveals that the ensemble feature set, when combined with the GraphSAGE model, outperforms the sequence-based and graphlet features individually on the HIPPIE dataset. In the subsequent chapter 5, this ensemble feature-based model and deduction obtained from the HIPPIE dataset has been extended to validate the MPXV-Human PPIN as a case study.

CHAPTER 5

Prediction of PHIs using ensemble features on MPXV-Human PPIN

5.1 Introduction

In this chapter, a demonstrative case study is performed to forecast the PHIs on the interacting protein network of Human and MPXV utilizing the ensemble feature-based model discussed in chapter 4. Studying the interaction between viral proteins and host proteins yields rich information on the dynamics of host-pathogen interactions, which is crucial for developing efficient vaccines. Since the GraphSAGE-based model with ensemble feature has provided satisfying results in the HIPPIE dataset compared to graphlet and sequence-based feature alone, we have extended our research on MPXV-Human PPIN sourced from [38]. In the following section 5.2, the methodology like dataset creation, feature extraction, training of DL model, and prediction is discussed followed by experimental results 5.3 and discussion 5.2.

5.2 Methodology

The same methodology 4.2 of feature generation technique, model training, and classification of HIPPIE PPIN is also used for MPXV-Human PPIN.

5.2.1 Formation of MPXV-Human PPIN

The MPXV-Human PPIN is created by merging the Human-human interactions obtained from the HIPPIE database with the projected interaction network

of MPXV-Human from [38]. The interaction network may be conceptualized as a graph of 7,576 nodes and 17,647 edges. The dataset consists of 7,576 nodes, which includes 10 nodes corresponding to the MPXV protein[38] and 7,566 nodes corresponding to human proteins. The 17,647 edges represent the protein interactions in the MPXV-Human PPIN. The positive dataset consists of these positive edges or interactions. The negative dataset is created by using negative sampling to the positive interactions in the PPIN, specifically targeting pairs of proteins that lack an interacting edge between them. Therefore, an equivalent number of negative edges has been produced in this manner.

5.2.2 Feature extraction of MPXV-Human PPIN

In our case study, experiments with MPXV-Human PPIN are conducted considering 2-5 nodes graphlet feature alone, features based on solely amino acid composition, and an ensemble of the two aforementioned features. The generation of the characteristic is threefold.

1. The graphlet features (up to 5 nodes) for MPXV-Human PPIN have been generated using ORCA similar to previous chapter 1. Since we had 7,576 protein nodes, orbital counts have been generated for each of the protein nodes resulting in a 7576 $\times$ 73-dimensional feature matrix.

2. The amino acid composition-based features that have been extracted for both MPXV (UniProt Organism ID: 619591) and human (UniProt Organism ID: 9606) using Pfeature tool is similar to section 2. The Variance Threshold Feature Selection method with a threshold of 0.13 has reduced the feature count from 976 to 196 resulting in a feature matrix of dimension of 7576 $\times$ 196 for the 7,576 protein nodes in the PPIN.

3. The final step involved combining two sets of features 1) 2-5 nodes 73-dimensional graphlet features and 2) 976-dimensional amino acid composition features summing up to a 1049-dimensional feature vector to leverage the node's topological property and amino acid composition-based features. From this 1049-dimensional feature vector, a 269-dimensional feature vector has been generated using the Variance threshold feature selection algorithm with a threshold of 0.13 to generate an ensemble feature matrix of size 7576 $\times$ 269.

5.2.3 Deep Neural Network Model

The dataset, consisting of 17,647 positive edges and 17,647 negative edges, has been split into an 80% training set and 20% test set. Therefore, the training

set has 14,117 positive and negative edges each, whereas the test set comprises 3,530 positive and negative edges each. The DL GraphSAGE model has been trained using the provided training data. The training model is composed of three SAGEConv layers, each employing a Leaky-ReLU activation function and a dropout layer. The experimental setup has been maintained identical to the one utilized during the training of the HIPPIE dataset. The model's output has been provided to the classifier for the ultimate classification task.

5.2.4 Classification Strategy

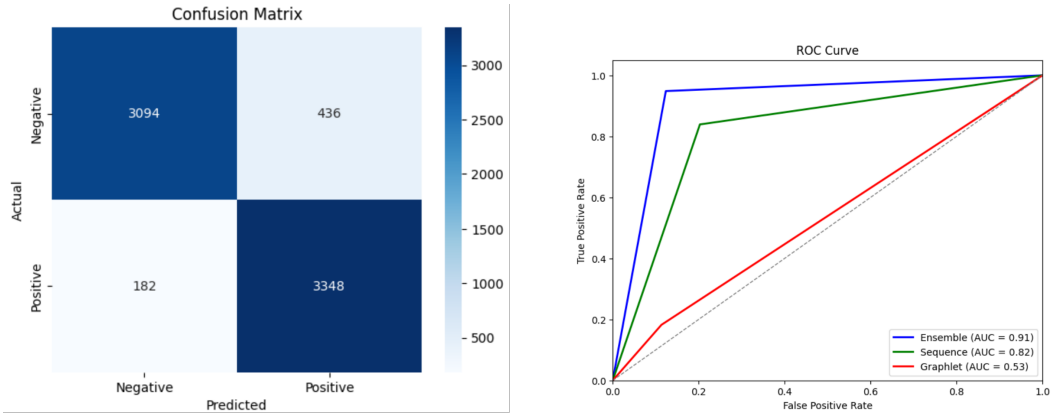
The classifier's architecture resembles that used for the HIPPIE dataset 4.2.4 employing two linear layers, each incorporating a dropout mechanism, with a final sigmoid activation to predict edges, labeling positive edges as "1" and negative edges as "0". The model is trained on both positive and negative edges, and test edges are classified based on a threshold of 0.5, where probabilities above 0.5 are labeled "1" (positive class) and below are labeled "0" (negative class).

5.3 Results

A comprehensive experiment is implemented utilizing MPXV-Human PPIN data to evaluate the efficacy of graphlet, sequences, and ensembled feature-based approaches. The findings from Table 5.1 demonstrate a significant advantage of the ensemble technique in capturing and leveraging essential patterns within the Mpox data compared to both graphlet-based and sequence-based features in terms of all assessment measures. A summary of the proposed ensemble model's predictions is also presented, highlighting the counts of TP, TN, FP, and FN instances in Fig 5.1(a). It can be seen from the figure that the model has performed well in correctly identifying positive instances (TP: 3348) and negative instances (TN: 3094). However, it has made some minor errors, incorrectly classifying some instances as positive when they are negative (FP: 436) and vice versa (FN: 182). Moreover, AUC-ROC comparison plots are plotted for our proposed model using three distinct features: ensemble feature, sequence-based feature, and graphlet feature along with their scores, which have been reported in Fig 5.1(b). In the figure, the colour blue, green and red represent the ensemble feature, sequence-based feature, and graphlet feature respectively. The ensemble feature achieved the highest AUC-ROC of 0.91, outperforming the sequence-based and graphlet features which scored 0.82 and 0.53 respectively.

Table 5.1: Performance scores of our proposed model with graphlet feature, sequence-based feature, and ensemble feature on MPXV-Human dataset. The best values are in bold.

	Graphlet Feature	**Sequence-based Feature**	**Ensemble Feature**
PR	0.6165	0.8053	**0.8848**
RE	0.1822	0.8391	**0.9484**
MCC	0.097	0.6368	**0.8271**
*F*1	0.2812	0.8219	**0.9155**



(a) (b)

Figure 5.1: a) Confusion matrix of the proposed ensemble-based feature model on MPXV-Human PPIN. b) AUC-ROC plot of the proposed model with ensemble feature, sequence-based feature, and graphlet feature with the AUC-ROC.

5.4 Discussion

This chapter demonstrates the efficacy of the ensemble-feature-based model, introduced in chapter 4, on MPXV-Human PPIN to predict PHIs. Table 5.1 demonstrates that the GraphSAGE model did not achieve satisfactory performance with 73-dimensional graphlet features (up to 5 nodes) alone but improves with 196-dimensional sequence-based features. The optimal performance was obtained by combining graphlet features with the sequence-based features, resulting in an ensemble 269-dimensional feature vector for each protein node. The suggested model accurately predicts 3,348 positive PHIs involving the MPXV-Human virus. In the next chapter 6, a drug repurposing investigation is undertaken on these accurately identified MPXV-targeted host proteins to identify existing treatments that may disrupt the interaction between the host and the virus.

CHAPTER 6

Computational Drug Repurposing Study on MPXV-Human PPIN

6.1 Introduction

This chapter presents a case study of drug repurposing on the MPXV-targeted human proteins identified in the previous chapter5. Drug repurposing [83], also known as drug repositioning, is the process of discovering novel therapeutic applications for pharmaceuticals that have previously been authorized or are currently being tested in clinical trials. The use of de-risked compounds in this approach is particularly appealing since it has the potential to lower development costs and shorten timescales [83] [84]. Since the drugs already exist, the costs and time associated with preclinical and early clinical development phases are often lower compared to *de novo* drug discovery. In addition, repurposing drugs already approved by the FDA for a different indication is more cost-effective, as it involves previously established potential toxicities and can boast a higher success rate (30%) compared to developing an NME [84]. Additionally, each unsuccessful clinical trial of an NME requires significant time and resources. It can therefore offer a quick response to MPXV, an emerging viral threat.

With the progress of research and medicine, data on drug targets and their associated disorders began to be stored in several databases, such as PubMed [85], DrugBank [86], ChemBank [87], OMIM [88], and GeneCards [89]. This has motivated us to investigate more and develop a cohesive strategy to find new medicinal therapies for treating Mpox.

In our drug repurposing study, the computationally active drugs have been de-

veloped using a consensus scoring strategy [90] and mapping of the drug-protein target downloaded from DrugBank. Drugs and their corresponding protein targets are retrieved from DrugBank which are biologically validated and approved. Thus drugs identified using this methodology 6.2 are *a priori* aligned with biological understanding. Hence there is a higher scope of success, reliability, and explainability. The result of this study is compiled in section 6.3 followed by a discussion in section 6.4.

6.2 Methodology

Our proposed model successfully predicts 3,348 positive PPIs in the MPXV-Human PPIN 5.3. This achievement has inspired us to evaluate current medications to uncover new options that can disrupt the connections between the host and the virus. The identified human protein nodes in the MPXV-Human PPIN serve as protein targets for different FDA drugs in DrugBank [86]. DrugBank, an extensive online repository, provides comprehensive data on pharmaceuticals, their biological targets, and potential interactions between different medications. In this study, the drugs have been identified using the DCS [90] and selected using the *ProcessDrugData* algorithm 1. These drugs might be the potential contenders for treating MPXV. The algorithm takes the Mapping of Drug ID, Drug Name, and corresponding protein targets for each drug downloaded from the DrugBank. The MPXV-targeted human protein nodes in the test PPI of MPXV-Human PPIN correctly identified by our ensemble feature-based model also served as the target proteins for the algorithm's input. The program retrieves the drug IDs that match our protein targets from the provided mapping and outputs the drug names. The DCS for each of these drugs is determined. The metric quantifies the frequency of occurrence of each drug throughout the whole MPXV-Human PPIN. Therefore, it is precisely defined according to Eq (6.1) below :

$$DCS(i) = \frac{\text{frequency}(i)}{N} \tag{6.1}$$

Here, DCS(i) represents the score for the drug 'i' linked with a human protein in the MPXV-Human PPIN, frequency(i) represents the total occurrences of drug 'i' in the MPXV-Human PPIN, and 'N' denotes the total number of drugs associated with human proteins in the MPXV-Human PPIN.
Once the DCS has been computed for each drug, they are arranged in descending order according to their determined DCS. Finally, the drug names are categorized into three distinct levels or domains based on their DCS. The drug data processing approach is outlined in the algorithm provided below:

Algorithm 1 ProcessDrugData

Input: path_set ▷ Path to the directory of the below input files

- Mapping of Drug ID, Drug Name, and corresponding protein targets for each drug
- List of Human-protein targets

Output: Generates a three-level categorization of drugs based on their frequency associated with protein targets

procedure ProcessDrugData(path_set)
    Read Drug ID, Drug Name, and corresponding protein targets for each drug.
    Read list of Human-protein targets.
    Retrieve Drugs IDs for the target proteins from the mapping.
    Retrieve Drug names for the drug IDs found in *Step 4*.
    Calculate the DCS of each drug (d) per Eq (6.1) retrieved in *Step 5*.
    Arrange the drugs in descending order according to the DCS computed in *Step 6*.
    // Group the drugs according to the DCS:
    for each drug d **do**
        if $\text{DCS}(d) \geq 9$ **then** assign level I to d
        else if $3 \leq \text{DCS}(d) < 9$ **then** assign level II to d
        else if $1 \leq \text{DCS}(d) < 3$ **then** assign level III to d
    Return the Drug names along with their levels and DQ score.

6.3 Results

In this section, repurposed drugs and sufficient supporting evidence from existing literature to support the potential effectiveness of these selected drugs in the treatment of MPXV are presented.

Categorization of the repurposed drugs

The drug components identified through *ProcessDrugData* algorithm have been listed in Table 6.1 along with their DrugBank Accession Number, Groups in DrugBank, and computed DCS. The drugs are ranked in descending order based on their DCS, with NADH [91] being the most computationally efficient treatment for MPXV, with a score of 35. The repurposed pharmaceuticals are categorized into three domains, namely Level I, Level II, and Level III, based on their DQ Scores. The medications that get the best scores in computational efficiency for treating Mpox have been classified as Level I. Subsequently, repurposed drugs, which exhibit lower efficacy (better scores) compared to Level I treatments, are categorized as Level II. Ultimately, drugs that get good scores (in comparison to Level II drugs) in computational efficiency for treating Mpox have been categorized

under Level III. The pharmacological results have been supported by many pieces of evidence documented in the subsequent portion of the available literature.

Level I Repurposed Drugs evidence

1. Shang et al. [92] emphasize the importance of NAD+ and the enzymes that use NAD+ in immune responses against viral infections.

2. Kataria et al. [93] analyzed the genomes of various Mpox strains and the human proteome to identify potential therapeutic candidates against MPXV. Among these, fostamatinib was highlighted as a prospective option.

3. Fontaine et al. [94] highlighted that an extensive metabolic analysis of primary human foreskin fibroblasts infected with VACV unveiled a shift in glutamine metabolism. Glutamine, a derivative of glutamic acid, plays a vital role as a carbon source in mammalian cells. Deprivation of exogenous glutamine led to a substantial reduction in infectious virus production in VACV-infected cells.

4. Zach J. Leinen et al. [95] conducted a review indicating that cannabinoids, including Cannabidiol and Tetrahydrocannabinol, play a substantial role in treating various conditions that span neurological disorders, inflammatory bowel disease, cancer, and emerging viral diseases like Mpox.

5. Kampf, Günter. [96] said that copper (99.9%) has demonstrated comparable effectiveness against both VACV and MPXV when applied for 3 minutes, thereby highlighting its usefulness in treating MPXV.

6. Clinical trials [97] have shown that applying a topical solution containing zinc oxide to the affected skin areas offers several advantages for individuals with MPXV infection. Zinc aids in drying up the pox, alleviating itching, and safeguarding against bacterial superinfection.

Level II Repurposed Drugs evidence

1. Lam et al., using homology modeling and molecular docking proposed FAD as a potential compound for suppressing the reproduction of the virus [98].

2. Arndt et al.'s study delineated the antiviral attributes of *Sarracenia purpurea* extract against viruses like vaccinia virus and monkeypox virus. Freshly grown plants were combined with 190-proof ethanol in a 1:3 ratio, demonstrating efficacy in impeding poxvirus reproduction during the initial stage of viral replication [99].

Table 6.1: Domain, DrugBank Accession Number, Drug Name, DCS, and DrugBank Groups of the drugs identified through *ProcessDrugData* algorithm.

Domain	Accession Number	Drug Name	DCS	Group
Level I	DB00157	NADH	35	A, N
	DB12010	Fostamatinib	27	A, I
	DB00142	Glutamic acid	25	A, N
	DB09061	Cannabidiol	9	A, I
	DB09130	Copper		A, I
	DB01593	Zinc		A, I
Level II	DB03147	FAD	8	A
	DB00898	Ethanol	7	A
	DB00273	Topiramate	6	A
	DB00128	Aspartic acid		A, N
	DB04272	Citric acid	5	A, N, V
	DB00252	Phenytoin	4	A, V
	DB00174	Asparagine		A, I, N
	DB00184	Nicotine		A
	DB09462	Glycerin		A, I
Level III	DB00316	Acetaminophen	2	A
	DB00141	N-Acetylglucosamine		A, I, N
	DB00688	Mycophenolate mofetil		A, I
	DB00166	Lipoic acid		A, I, N
	DB00819	Acetazolamide		A, V
	DB00119	Pyruvic acid		A, I, N
	DB00570	Vinblastine		A
	DB00309	Vindesin	1	A, I
	DB01229	Paclitaxel		A, V
	DB01248	Docetaxel		A, I
	DB06772	Cabazitaxel		A
	DB02709	Resveratrol		I
	DB00281	Lidocaine		A, V
	DB09213	Dexibuprofen		A, I
	DB01050	Ibuprofen		A
	DB01098	Rosuvastatin		A
	DB00811	Ribavirin		A
	DB01024	Mycophenolic acid		A, I
	DB01086	Benzocaine		A, I
	DB00193	Tramadol		A, I
	DB00338	Omeprazole		A, I, V
	DB11672	Curcumin		A, I
	DB12965	Silver		A, I
	DB06736	Aceclofenac		A, I
	DB00655	Estrone		A
	DB03904	Urea		A, I

N.B.: A: Approved, N: Nutraceutical, I: Investigational and V: Vet approved.

3. Topiramate was administered as a therapeutic intervention for headache, neck stiffness, and diplopia, which are neurological manifestations of Mpox [100], as stated in the review study [101].

4. Singh et al. [102] identified Asp331, Asn312, Lys314, and Asn329 as the specific docking sites for the p37 protein, pivotal for viral infection propagation. These peptides could serve as reliable starting points for subsequent in vitro and in vivo experiments to develop peptide-based therapies for Mpox treatment.

5. The publication [103] details the retrieval of the Mpox viral antigen utilizing an anti-Vaccinia virus rabbit antibody. The retrieval procedure entailed autoclaving the antigen at 121°C and a pH of 6.0 in a citric acid solution for 15 minutes under specific conditions.

6. In a case study [104], the neurological symptom of Mpox, specifically seizures, was effectively managed with the use of phenytoin and diazepam.

7. Pal et al. [105] conducted an experiment using *Nicotiana clevelandii* and *Nicotiana benthamiana* plants containing nicotine, revealing their ability to control the temporary buildup of VACV from day 20, preventing symptom occurrence. They also found that combining botanical extracts from Sarracenia with vegetable glycerin effectively suppressed orthopoxvirus replication.

Level III Repurposed Drugs evidence

1. The authors in [106] have outlined the production of glycovariants of 7D11, a neutralizing monoclonal IgG antibody (mAb) targeting the L1 transmembrane protein of the closely related MPXV and displaying strong neutralization activity against MPXV. The p7D11 mAb expressed in wild-type (WT) glycoengineered plants (p7D11WT mAb) was observed to possess complex-type N-glycans, with N-acetylglucosamine (GlcNAc) identified as the terminal glycan residue on at least one of the two termini.

2. The paper [107] identifies Mycophenolic acid (MPA) and mycophenolate mofetil (MMF) as important drugs that can be used as reference points for the development of targeted anti-MPVX drugs. These drugs act as inhibitors of Inosine 5'-monophosphate dehydrogenase (IMPDH), an enzyme necessary for viral activity. MPA exhibits direct inhibition of IMPDH, whereas MMF functions as a prodrug that releases MPA [108].

3. The authors of this study [109] have reported a groundbreaking finding that two redox-modulating drugs, ethacrynic acid, and lipoic acid, impede the in-vitro proliferation of VACV.

4. Acetazolamide was administered as a therapeutic intervention for neurological symptoms such as headache, neck stiffness, and diplopia associated with Mpox [100], as stated in [101].

5. The inhibition of viral replication by cidofovir (CDV) requires the formation of CDV's diphosphate form (CDVpp). Pyruvate kinase, creatine kinase, and nucleoside diphosphate kinase have the ability to convert CDV monophosphate into its diphosphate form [110].

6. The study conducted by [111] emphasizes the significant involvement of the microtubule network in the replication cycle of Mpox virions. Consequently, compounds targeting microtubules, derived from natural sources or synthesized chemicals such as Vinblastine, Vindesine, Paclitaxel, Docetaxel, Cabazitaxel, etc., exhibit strong potential as antiviral candidates.

7. The study [112] has shown that several investigational drugs, including adamantane derivatives, monoterpenoid derivatives, and resveratrol, have exhibited promising antiviral properties in in-vitro trials.

8. Hallo et al. [113] reported that Metamizole 500 mg q6h, Lidocaine 50 mg/g topical, and Ibuprofen can relieve pain caused by Mpox [114]. Dexibuprofen, the active enantiomer of racemic ibuprofen, is also effective. Acetaminophen has also shown efficacy in a separate study [115]

9. Wang et al. [116] identified the top 20 putative inhibitors of MPXV E12. Among these potential inhibitors are Rutin, Quercitrin, Epigallocatechin, Rosuvastatin, 5-hydroxy-L-Tryptophan, and Deferasirox, among others.

10. Baker et al. [117] demonstrated that ribavirin and seven other drugs effectively inhibit orthopoxvirus replication at pharmacologically relevant doses. Separate research [118] identified several promising antiviral drugs for MPXV, including ribavirin, tiazofurin, protease inhibitors, and hydrolase inhibitors, which have demonstrated effectiveness against Mpox.

11. Hans et al. [119] recommended topical benzocaine/lidocaine gels to relieve skin rash associated with Mpox temporarily. Tapentadol or tramadol are suggested as second-line therapies for severe pain and Omeprazole for Dyspepsia is recommended as part of a comprehensive supportive management approach for acute Mpox infections.

12. Banik et al. [120] discovered that curcumin had the highest binding affinity with the pox protein compared to other plant compounds, indicating its potential as a virus inhibitor.

13. The study [121] highlights the use of silver nanoparticles in viral vaccine development due to their shown antiviral effectiveness against MPXV.

14. Periorbital edema has been observed in various ocular symptoms of Mpox infections [122]. A successful outcome has been achieved by treating patients with aceclofenac and paracetamol.

15. Molecular docking investigations were done in the publication [123] using Mpox-associated targets and selected putative active molecules generated from the PPI network. The study confirmed that estrone (MOL008204), the main active chemical, binds to the AR target (PDB ID: 2QPY) and is the most effective therapy choice for Mpox.

16. 16. Lee et al. have discovered that Hydroxyurea acts as a Ribonucleotide reductase inhibitor and has potential as an anti-orthopoxvirus agent. This finding is based on a preclinical study [124].

6.4 Discussion

In this chapter, a drug repurposing study is introduced which seeks to utilize existing treatments for alternative purposes, rather than creating wholly new drugs. By leveraging existing pharmaceuticals, this approach reduces the risks of developing new compounds and offers a viable answer to the time constraints of traditional drug development. In our methodology, the drugs and their corresponding protein targets mapping are downloaded from DrugBank. Using the *ProcessDrugData* algorithm, the potential drugs linked to the human proteins of MPXV targets and involved in direct positive interactions are retrieved from the DrugBank mapping. The DCS is calculated for each such drug. These scores are organized in descending order based on their value. Subsequently, the most promising drugs that might potentially alleviate the consequences and symptoms of Mpox have been grouped into three levels I, II, and III based on their DCS as shown in Table 6.1. In addition, many existing pieces of supporting data are highlighted, from successful studies conducted on these drugs, in treating Mpox. In the next chapter 7, a molecular docking study has been performed between the top four level I repurposed drugs and the target MPXV protein receptor to provide further insights into the effectiveness of these newly found repurposed drugs.

CHAPTER 7

Computational Docking of potential drugs for MPXV protein structures

7.1 Introduction

In the previous chapter 6, the drug repurposing study highlights the computationally potential drugs segregated into three tiers that could treat Mpox. This chapter presents a docking study that sheds further light on the efficacy of these drugs. Molecular docking is a highly effective method for structure-based drug discovery. It seeks to anticipate the experimental binding modes and affinities of small molecules within the binding sites of specific receptor targets [125]. In our study, docking is used to predict the interaction between the top four repurposed drug molecules (ligand) 6.1 and Mpox proteins (target receptor). The rest of the chapter is organized as follows. Section 7.2 describes the docking procedure. The docking results of the top four repurposed drugs are detailed in section 7.3 and section 7.4 concludes the chapter.

7.2 Methodology

Protein-ligand docking is performed using AutoDock Vina [126] version 1.5.6 on the top four level I prospective drugs listed in Table 6.1 of the previous chapter and BIOVIA Discovery Studio Visualizer 4.5 [127] is used for the visualization of the crystal structures. The optimization of the MPXV protein involved removing water and other atoms, followed by adding a polar hydrogen group. This molecular modeling is employed in the proposed methodology to assess the

binding potential of repurposed drugs with MPXV. The available Crystal structure of MPXV having PDB ID: 8CEQ [128] is downloaded from the RCSB PDB (`https://www.rcsb.org/`) and the molecules of the repurposed drugs are downloaded from DrugBank in SDF/PDB format. The docking process yields docking scores which aid in identifying the most suitable molecules docked at the chosen target site. Subsequently, all molecules are ranked according to these scores. A lower score suggests that a smaller amount of energy is needed for the ligand-receptor interaction and thus a stronger protein-ligand complex.

7.3 Results

In this section, the docking results of the top four repurposed drug molecules NADH, Fostamatinib, Glutamic acid, and Cannabidiol with MPXV protein structure is presented.

Docking results of NADH with respect to MPXV protein structure

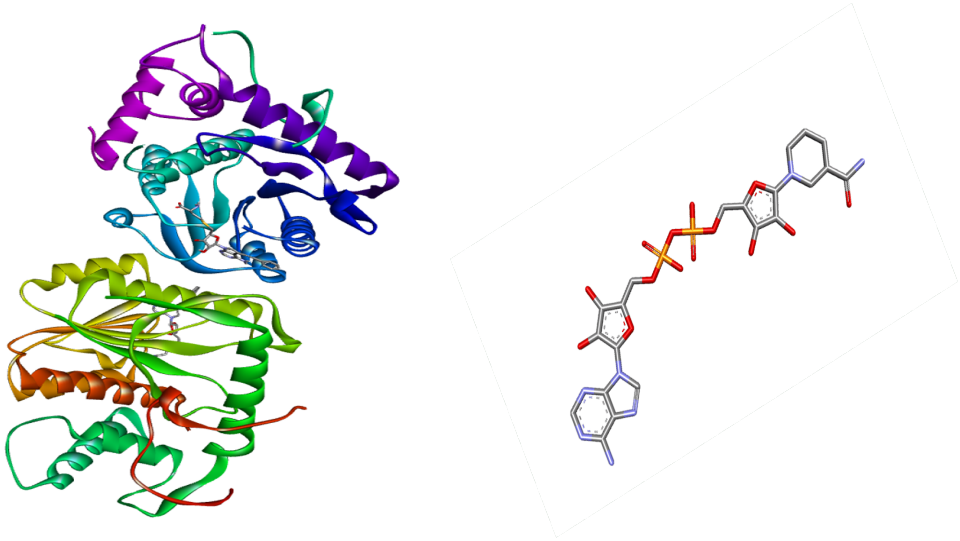
A detailed analysis of molecular docking between NADH having the highest DCS (35) and MPXV methyltransferase VP39 in complex with inhibitor TO427 (PDB ID: 8CEQ) has been done. MPXV protein crystal structure is obtained from RCSB PDB in .pdb format. Fig 7.1a presents the crystal structures of MPXV 2'-O-RNA methyltransferase (MTase) VP39 in complex with the pan-MTase inhibitor sinefungin. The molecular structure of NADH obtained from DrugBank is shown in Fig 7.1b. After successful docking, nine modes of drug-protein interactions are generated, each with specific docking scores indicating binding energy. The mode with the lowest binding energy is identified as the optimal binding mode, suggesting the most stable interaction between the ligand and the protein. The best affinity score obtained is -9.7 (kcal/mol) as evident from Table 7.1. RMSD values are computed relative to the best mode and exclusively involve movable heavy atoms. Two forms of RMSD metrics are offered: rmsd/lb (RMSD lower bound) and rmsd/ub (RMSD upper bound), which differ in the method of matching atoms during distance calculation: [129].

- rmsd/ub matches each atom in one conformation with its corresponding atom in the other conformation, disregarding any symmetry.

- rmsd/lb is described as follows:

$$\text{rmsd/lb}(c1, c2) = \max(\text{rmsd}'(c1, c2), \text{rmsd}'(c2, c1)) \tag{7.1}$$

where rmsd' aligns each atom in one conformation with the nearest atom of the same element type in the other conformation.



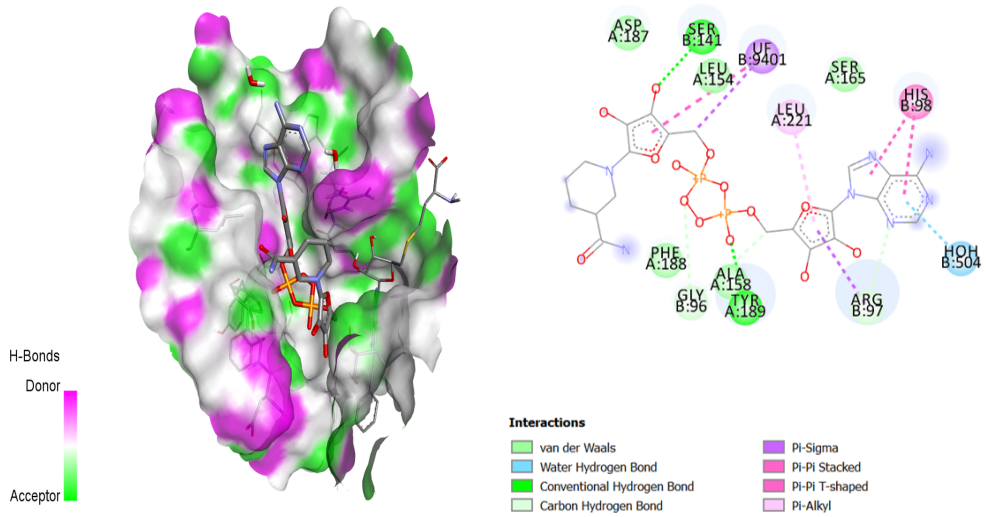
(a) (b)

Figure 7.1: a) Crystal structure of MPXV methyltransferase VP39 in complex with inhibitor TO427 retrieved from PDB. b) Molecular structure of NADH retrieved from DrugBank.

The interaction of NADH with the MPXV protein shows a high-affinity interaction in Fig 7.2a as the ligand fits inside the core pocket region of the protein. This is further supported by the hydrogen bonding observed between the oxygen atom of NADH with TYR 189 and SER 141. The details of the various interactions along with the active sites such as ASP 187, SER 141, LEU 154, LEU 221, SER 165, HIS 98, ARG 97, ALA 158, TYR 189, GLY 96, PHE 188 have been shown in Fig 7.2b. This further highlights that this widely available natural compound holds promise as a potential anti-Mpox medication, warranting further investigation, which can benefit both patients and public health initiatives.

Docking results of Fostamatinib, Glutamic acid, and Cannabidiol with respect to MPXV protein structure

The virtual interaction of the MPXV methyltransferase VP39 in complex with inhibitor TO427 (PDB ID: 8CEQ) with the other Level I repurposed drugs i.e. Fostamatinib, Glutamic acid, and Cannabidiol has been measured using the AutoDock Vina suite. A similar docking study as NADH has been performed on these remaining top three level I drugs. The summarized results of the binding energy observations of the nine modes along with their distances from the best orientation for Fostamatinib, Glutamic acid, and Cannabidiol can be found in



(a) (b)

Figure 7.2: NADH docked in the MPXV methyltransferase VP39 in complex with inhibitor TO427 (PDB ID: 8CEQ) a) best binding mode in the protein pocket. b) the binding interaction of NADH with amino acids with various bonds.

Table 7.1: Docking scores/Affinity values along with the distance from the best mode of the nine modes of NADH. The best values are in bold.

Mode	Affinity (kcal/mol)	rmsd l.b.	rmsd u.b.
1	**-9.7**	**0.000**	**0.000**
2	-9.5	1.956	3.317
3	-9.4	1.942	2.869
4	-9.4	1.820	7.256
5	-9.3	2.592	7.838
6	-9.2	2.438	7.647
7	-9.1	1.969	7.598
8	-8.9	3.071	4.788
9	-8.9	2.313	7.636

Table 7.2, Table 7.3, and Table 7.4 respectively. The best binding energy results obtained from the docking of 8CEQ with these ligands are summarized in Table 7.5 along with their DCS. It is observed from the result that while NADH registers the highest DCS of 35 in the drug repurposing study, it still attains the highest position in comparison to the other repurposed drugs in the molecular docking study with an Affinity score of -9.7 kcal/mol. The docking scores of Fostamatinib, Glutamic acid, and Cannabidiol are found to be -8.1, -4.9, and -8.7 kcal/mol respectively. The docking result suggests that these drug molecules

have the capability to inhibit MPXV since they have demonstrated high-affinity interactions with the MPXV methyltransferase VP39 in complex with inhibitor TO427. This indicates opportunities for pharmaceutical companies to leverage existing drug libraries for expedited Mpox treatment development.

Table 7.2: Docking scores/Affinity values along with the distance from the best mode of the nine modes of Fostamatinib. The best values are in bold.

Mode	**Affinity (kcal/mol)**	**rmsd l.b.**	**rmsd u.b.**
1	**-8.1**	**0.000**	**0.000**
2	-7.9	2.439	7.739
3	-7.7	5.725	8.532
4	-7.7	5.851	8.541
5	-7.4	4.868	7.521
6	-7.3	2.745	7.394
7	-7.3	2.877	7.739
8	-7.1	5.347	8.857
9	-7.0	5.409	8.662

Table 7.3: Docking scores/Affinity values along with the distance from the best mode of the nine modes of Glutamic acid. The best values are in bold.

Mode	**Affinity (kcal/mol)**	**rmsd l.b.**	**rmsd u.b.**
1	**-4.9**	**0.000**	**0.000**
2	-4.8	2.088	2.451
3	-4.7	2.203	4.119
4	-4.7	1.644	2.019
5	-4.7	2.149	2.660
6	-4.6	1.736	4.214
7	-4.6	2.205	4.145
8	-4.4	2.298	3.341
9	-4.4	2.295	2.882

Table 7.4: Docking scores/Affinity values along with the distance from the best mode of the nine modes of Cannabidiol. The best values are in bold.

Mode	**Affinity (kcal/mol)**	**rmsd l.b.**	**rmsd u.b.**
1	**-8.7**	**0.000**	**0.000**
2	-8.5	3.003	4.881
3	-8.5	2.196	4.627
4	-8.4	3.147	4.075
5	-8.4	2.442	7.076
6	-8.3	1.232	3.648
7	-8.3	2.843	7.749
8	-7.9	1.915	6.962
9	-7.6	1.973	7.771

Table 7.5: Docking scores/Affinity of top four repurposed drugs for 8CEQ.

Drugs	**Drug ID**	**Affinity (kcal/mol)**	**DCS**
NADH	DB00157	**-9.7**	**35**
Fostamatinib	DB12010	-8.1	27
Glutamic acid	DB00142	-4.9	25
Cannabidiol	DB09061	-8.7	9

7.4 Discussion

To further strengthen our findings, a molecular docking study has been performed on the MPXV methyltransferase VP39 in complex with inhibitor TO427 (PDB ID: 8CEQ) (target protein receptor) with ligands such as NADH, Fostamatinib, Glutamic acid, and Cannabidiol. The docking affinity score (in kcal/mol) of these four drug molecules is reported in Table 7.5. The docking result suggests that these drug molecules have a strong capability to inhibit MPXV since they have demonstrated a high binding affinity with the target receptor 8CEQ. The likelihood of these drug molecules in combating Mpox can be further reinforced through clinical trials in the future. Consequently, this creates opportunities for pharmaceutical businesses to capitalize on existing drug libraries thereby speeding up the introduction of groundbreaking Mpox treatments.

CHAPTER 8

Conclusions

PPIs are fundamental to the functioning of biological systems, regulating a myriad of cellular processes. Therefore, accurate prediction of PPIs is crucial for disease diagnosis and the development of novel therapeutic interventions. In prior approaches, the majority of research on PPI has predominantly concentrated on sequence information. This approach disregards other types of information, such as evolutional, structural, or functional details, and solely depends on sequence data to predict interaction sites between proteins.

Additionally, pathogens, including viruses, utilize their proteins to specifically target human proteins to disrupt or manipulate biological circuits. PHIs determine the result of an infection, affecting the host's reaction, the ability of the pathogen to avoid the immune system, and the course of the disease. Therefore, studying the mechanisms of PHIs reveals the molecular foundations of infectious illnesses, offering potential opportunities for treatments.

Keeping the above facts in mind, the research work embodied in this thesis mainly focuses on three primary directions: 1) prediction of PPIs using different DL methodologies, 2) prediction of PHIs in MPXV-Human dataset considering Mpox as a case study, 3) detailed drug repurposing study on MPXV-Human PPIN, and 4) Computational docking of potential drugs for MPXV protein structures.

In chapter 2, an innovative bio-inspired substitution matrix-based CNN approach "EnMAS" for identifying interacting protein pairs has been proposed. In addition to achieving excellent performance in yeast and human datasets, the model has successfully predicted interactions between humans and SARS-COV-2.

In chapter 3, a topological feature-based PPI prediction method has been developed on the HIPPIE dataset. It introduces a GNN-based GraphSAGE model "GSPPI" designed to forecast protein interactions utilizing the graphlet feature (up to 5 nodes) where each protein acts as a node in PPIN. The proposed model has demonstrated noteworthy performance by exploiting the local network topo-

logical properties of protein nodes.

In chapter 4, a comprehensive approach for predicting PPIs by integrating graphlet-based features and amino acid composition-based features has been proposed. The graphlet features encapsulate local structural patterns within protein interaction networks, capturing intricacies in the network topology up to 5 nodes. Simultaneously, amino acid composition-based features encode information about the sequence composition of proteins, providing insights into their biochemical properties. By combining these two types of features, the proposed model can effectively capture both structural and sequence-related characteristics of proteins, enabling a more holistic representation of the underlying biology governing PPIs. This integration improves the predicting capability of protein interactions with high accuracy and robustness.

In chapter 5, the ensemble feature-based approach discussed in 4 has been leveraged to predict PHIs within the MPXV-Human PPIN. The application of the integrated approach has uncovered novel interactions between MPXV proteins and human host proteins shedding light on the dynamics of MPXV-host interactions, offering scope for therapeutic intervention and drug discovery in combating Mpox infection.

In chapter 6, a drug repurposing study is utilizing the identified MPXV-targeted human proteins. It involves screening FDA-approved drugs against the identified protein targets within the MPXV-Human PPIN and identifying potential candidate drugs capable of treating Mpox using a drug consensus scoring strategy. The identified drugs have been categorized into three tiers based on their computed DCS. Furthermore, the efficacy of the potential drugs has been supported by various literature pieces of evidence.

In chapter 7, following the selection of candidate drugs through the drug repurposing study, the top four level I drugs have undergone molecular docking studies to assess their potential binding affinity and interaction with the available crystal structure of MPXV from PDB. The results of molecular docking studies serve as valuable guidance in the selection and optimization of repurposed drugs for Mpox treatment.

8.1 Limitations of the Work

Although this work achieves its main goal of predicting interactions between host (human) and pathogens (SARS-COV-2 and MPXV), it is important to acknowledge some limitations of the proposed methods.

- The EnMAS model has performed well on the yeast dataset but the substitu-

tion matrix-based colour encoding strategy has failed to prove its superiority over the randomized colour encoding strategy on the Pan et al. dataset [130]. The proposed EnMAS model has achieved a PR of 98.97%, RE of 98.53%, MCC of 97.44%, F1 of 98.75%, and AUC-ROC of 99.87% which is slightly less than DensePPI [33].

- In the GSPPI and subsequent ensemble feature-based model, one possible limitation could be the reliance on the HIPPIE dataset for obtaining PPI data. While HIPPIE is a valuable resource, it may not encompass the entire spectrum of PPIs present in biological systems, which might limit the model's generalizability.

- Although computational drug repurposing study and docking have been successful in identifying three potential drugs such as NADH, Fostamatinib, and Cannabidiol in combating MPXV, it is not sufficient to reach any logical end to the drug discovery process since docking compounds into a static protein structure can be misleading as the chosen conformation may not be representative of the conformation capable of binding the ligand.

8.2 Scope for Future Works

- This work can be extended to residue-level prediction to enhance a deeper understanding of PPIs. This direction broadens the applicability of the method in active-site prediction, drug-target identification, drug design, and functional enrichment.

- In ensemble feature-based PPIs / PHIs prediction, GO annotation [131] which measures functional similarity between proteins can be incorporated along with graphlet features (up to 5 nodes) for learning intricate interactions. Since similar functions indicate a higher likelihood of interaction, this can predict potential interactions.

- In drug repurposing, DL and mechanistic AI can be integrated with our consensus scoring strategy to identify prospective drugs as anti-viral medications. Further for any emerging viral diseases like Mpox and COVID-19 where PHI network information is not known or fully established initially, this computational study can bridge the gap *in-silico* and still predict some potential repurposable drugs through high interactivity scores in the future.

- In molecular docking studies, compounds demonstrating promising binding affinities and interaction patterns can be progressed to subsequent stages of

pre-clinical and clinical evaluation, paving the way for the development of effective therapeutics against this infectious disease.

Bibliography

[1] B. Alberts, "The cell as a collection of protein machines: Preparing the next generation of molecular biologists," *cell*, vol. 92, no. 3, pp. 291–294, 1998.

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

[3] P. Braun and A.-C. Gingras, "History of protein–protein interactions: From egg-white to complex networks," *Proteomics*, vol. 12, no. 10, pp. 1478–1498, 2012.

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

[5] S. Saha, P. Chatterjee, M. Nasipuri, and S. Basu, "Detection of spreader nodes in human-sars-cov protein-protein interaction network," *PeerJ*, vol. 9, e12117, 2021.

[6] X. Zhou, B. Park, D. Choi, and K. Han, "A generalized approach to predicting protein-protein interactions between virus and host," *BMC genomics*, vol. 19, no. 6, pp. 69–77, 2018.

[7] D. P. Ryan and J. M. Matthews, "Protein–protein interactions in human disease," *Current opinion in structural biology*, vol. 15, no. 4, pp. 441–446, 2005.

[8] I. Petta, S. Lievens, C. Libert, J. Tavernier, and K. De Bosscher, "Modulation of protein–protein interactions for the development of novel therapeutics," *Molecular Therapy*, vol. 24, no. 4, pp. 707–718, 2016.

[9] U. Kuzmanov and A. Emili, "Protein-protein interaction networks: Probing disease mechanisms using model systems," *Genome medicine*, vol. 5, no. 4, pp. 1–12, 2013.

[10] J. Page, D. Hinshaw, and B. McKay, "In hunt for covid-19 origin, patient zero points to second wuhan market–the man with the first confirmed infection of the new coronavirus told the who team that his parents had shopped there," *The Wall Street Journal*, vol. 26, 2021.

[11] "Multi-country monkeypox outbreak: Situation update." (2022), [Online]. Available: `https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON393`.

[12] "Who factsheet – mpox (monkeypox)." (2023), [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/monkeypox`.

[13] P. Joi. "Mpox is no longer officially a public health emergency; here's why we shouldn't let down our guard." (2023), [Online]. Available: `https://www.gavi.org/vaccineswork/mpox-no-longer-officially-public-health-emergency-heres-why-we-shouldnt-let-down`.

[14] "Who recommends new name for monkeypox disease." (2022), [Online]. Available: `https://www.who.int/news/item/28-11-2022-who-recommends-new-name-for-monkeypox-disease`.

[15] Y. Xiang and A. White, "Monkeypox virus emerges from the shadow of its more infamous cousin: Family biology matters," *Emerging microbes & infections*, vol. 11, no. 1, pp. 1768–1777, 2022.

[16] N. Chen, G. Li, M. K. Liszewski, *et al.*, "Virulence differences between monkeypox virus isolates from west africa and the congo basin," *Virology*, vol. 340, no. 1, pp. 46–63, 2005.

[17] F.-M. Lum, A. Torres-Ruesta, M. Z. Tay, *et al.*, "Monkeypox: Disease epidemiology, host immunity and clinical interventions," *Nature Reviews Immunology*, vol. 22, no. 10, pp. 597–613, 2022.

[18] C. Wenham and M. Eccleston-Turner, "Monkeypox as a pheic: Implications for global health governance," *The Lancet*, vol. 400, no. 10369, pp. 2169–2171, 2022.

[19] J. Shen, J. Zhang, X. Luo, *et al.*, "Predicting protein-protein interactions based only on sequences information," *Proc Natl Acad Sci*, vol. 104, no. 11, pp. 4337–4341, 2007.

[20] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic acids research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[21] Y. Z. Zhou, Y. Gao, and Y. Y. Zheng, "Prediction of protein-protein interactions using local description of amino acid sequence," in *Advances in Computer Science and Education Applications: International Conference, CSE 2011, Qingdao, China, July 9-10, 2011. Proceedings, Part II*, Springer, 2011, pp. 254–262.

[22] T. Sun, B. Zhou, L. Lai, and J. Pei, "Sequence-based prediction of protein protein interaction using a deep-learning algorithm," *BMC bioinformatics*, vol. 18, pp. 1–8, 2017.

[23] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, "Deepppi: Boosting prediction of protein–protein interactions with deep neural networks," *Journal of chemical information and modeling*, vol. 57, no. 6, pp. 1499–1510, 2017.

[24] S. Hashemifar, B. Neyshabur, A. A. Khan, and J. Xu, "Predicting protein–protein interactions through sequence-based deep learning," *Bioinformatics*, vol. 34, no. 17, pp. i802–i810, 2018.

[25] H. Li, X.-J. Gong, H. Yu, and C. Zhou, "Deep neural network based predictions of protein interactions using primary sequences," *Molecules*, vol. 23, no. 8, p. 1923, 2018.

[26] R. Singh, K. Devkota, S. Sledzieski, B. Berger, and L. Cowen, "Topsy-turvy: Integrating a global view into sequence-based ppi prediction," *Bioinformatics*, vol. 38, no. Supplement_1, pp. i264–i272, 2022.

[27] K. Jha, S. Saha, and H. Singh, "Prediction of protein–protein interaction using graph neural networks," *Scientific Reports*, vol. 12, no. 1, p. 8360, 2022.

[28] M. Heinzinger, A. Elnaggar, Y. Wang, *et al.*, "Modeling aspects of the language of life through transfer-learning protein sequences," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–17, 2019.

[29] A. Elnaggar, M. Heinzinger, C. Dallago, *et al.*, "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 7112–7127, 2021.

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

[31] M. B. Koca, E. Nourani, F. Abbasoğlu, İ. Karadeniz, and F. E. Sevilgen, "Graph convolutional network based virus-human protein-protein interaction prediction for novel viruses," *Computational Biology and Chemistry*, vol. 101, p. 107 755, 2022.

[32] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[33] A. A. Halsana, T. Chakroborty, A. K. Halder, and S. Basu, "Denseppi: A novel image-based deep learning method for prediction of protein-protein interactions," *IEEE Transactions on NanoBioscience*, 2023.

[34] P. Green-Armytage, "A colour alphabet and the limits of colour coding," *JAIC-Journal of the International Colour Association*, vol. 5, 2010.

[35] A. K. Halder, S. S. Bandyopadhyay, P. Chatterjee, M. Nasipuri, D. Plewczynski, and S. Basu, "Juppi: A multi-level feature based method for ppi prediction and a refined strategy for performance assessment," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 531–542, 2020.

[36] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "Ippi-esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into pseaac," *Journal of theoretical biology*, vol. 377, pp. 47–56, 2015.

[37] Z.-H. You, W.-Z. Huang, S. Zhang, Y.-A. Huang, C.-Q. Yu, and L.-P. Li, "An efficient ensemble learning approach for predicting protein-protein interactions by integrating protein primary sequence and evolutionary information," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 3, pp. 809–817, 2018.

[38] S. Saha, P. Chatterjee, M. Nasipuri, S. Basu, and T. Chakraborti, "Computational drug repurposing for viral infectious diseases: A case study on monkeypox," *Briefings in Functional Genomics*, elad058, 2024.

[39] S. Rodriguez, C. Hug, P. Todorov, *et al.*, "Machine learning identifies candidates for drug repurposing in alzheimer's disease," *Nature communications*, vol. 12, no. 1, p. 1033, 2021.

[40] X. Li, Z. Li, X. Wu, *et al.*, "Deep learning enhancing kinome-wide polypharmacology profiling: Model construction and experiment validation," *Journal of medicinal chemistry*, vol. 63, no. 16, pp. 8723–8737, 2019.

[41] B. Li, C. Dai, L. Wang, *et al.*, "A novel drug repurposing approach for non-small cell lung cancer using deep learning," *PLoS One*, vol. 15, no. 6, e0233112, 2020.

[42] L. G. Trabuco, M. J. Betts, and R. B. Russell, "Negative protein–protein interaction datasets derived from large-scale two-hybrid experiments," *Methods*, vol. 58, no. 4, pp. 343–348, 2012.

[43] P. Blohm, G. Frishman, P. Smialowski, *et al.*, "Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis," *Nucleic Acids Res*, vol. 42, no. D1, pp. D396–D400, 2013, ISSN: 1362-4962.

[44] Y. Park and E. M. Marcotte, "Revisiting the negative example sampling problem for predicting protein–protein interactions," *Bioinformatics*, vol. 27, no. 21, pp. 3024–3028, 2011.

[45] Y. Park and E. M. Marcotte, "Flaws in evaluation schemes for pair-input computational predictions," *Nature methods*, vol. 9, no. 12, pp. 1134–1136, 2012.

[46] T. Hamp and B. Rost, "Evolutionary profiles improve protein–protein interaction prediction from sequence," *Bioinformatics*, vol. 31, no. 12, pp. 1945–1950, 2015.

[47] T. Hamp and B. Rost, "More challenges for machine-learning protein interactions," *Bioinformatics*, vol. 31, no. 10, pp. 1521–1525, 2015.

[48] A. Widjaja, "Point accepted mutation (pam) and block substitution matrix (blosum) interactive programs on kelleybioinfo. org," Ph.D. dissertation, San Diego State University, 2017.

[49] R. Trivedi and H. A. Nagarajaram, "Substitution scoring matrices for proteins-an overview," *Protein Science*, vol. 29, no. 11, pp. 2150–2163, 2020.

[50] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, 1992.

[51] M. O. Dayhoff, "A model of evolutionary change in proteins," *Atlas of protein sequence and structure*, vol. 5, pp. 89–99, 1972.

[52] G. Rédei, "Dip (database of interacting proteins)," in *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, Springer Netherlands, 2008, pp. 511–511.

[53] A. Bairoch and B. Boeckmann, "The swiss-prot protein sequence data bank," *Nucleic acids research*, vol. 19, no. Suppl, p. 2247, 1991.

[54] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D449–D451, 2004.

[55] H.-W. Mewes, D. Frishman, C. Gruber, *et al.*, "Mips: A database for genomes and protein sequences," *Nucleic acids research*, vol. 28, no. 1, pp. 37–40, 2000.

[56] G. D. Bader, D. Betel, and C. W. Hogue, "Bind: The biomolecular interaction network database," *Nucleic acids research*, vol. 31, no. 1, pp. 248–250, 2003.

[57] A. Chatr-Aryamontri, R. Oughtred, L. Boucher, *et al.*, "The biogrid interaction database: 2017 update," *Nucleic acids research*, vol. 45, no. D1, pp. D369–D379, 2017.

[58] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, *et al.*, "Intact: An open source molecular interaction database," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D452–D455, 2004.

[59] L. Licata, L. Briganti, D. Peluso, *et al.*, "Mint, the molecular interaction database: 2012 update," *Nucleic acids research*, vol. 40, no. D1, pp. D857–D861, 2011.

[60] T. Keshava Prasad, R. Goel, K. Kandasamy, *et al.*, "Human protein reference database—2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–D772, 2008.

[61] W. L. Ã and A. Godzik, "Cd-hit : a fast program for clustering and comparing large sets of protein or nucleotide sequences," vol. 22, no. 13, pp. 1658–1659, 2006. DOI: 10.1093/bioinformatics/btl158.

[62] E. Boutet, D. Lieberherr, M. Tognolli, *et al.*, "Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: How to use the entry view," in *Plant Bioinformatics*, Springer, 2016, pp. 23–54.

[63] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[64] W. R. Pearson, "Selecting the right similarity-scoring matrix," *Current protocols in bioinformatics*, vol. 43, no. 1, pp. 3–5, 2013.

[65] R. Draelos, *Measuring performance: Auprc and average precision*, 2020.

[66] A. J. Bowers and X. Zhou, "Receiver operating characteristic (roc) area under the curve (auc): A diagnostic measure for evaluating the accuracy of predictors of education outcomes," *Journal of Education for Students Placed at Risk (JESPAR)*, vol. 24, no. 1, pp. 20–46, 2019.

[67] X. Wang, Y. Wu, R. Wang, Y. Wei, and Y. Gui, "A novel matrix of sequence descriptors for predicting protein-protein interactions from amino acid sequences," *PloS one*, vol. 14, no. 6, e0217312, 2019.

[68] Y. Gui, R. Wang, Y. Wei, and X. Wang, "Dnn-ppi: A large-scale prediction of protein-protein interactions based on deep neural networks," *Journal of Biological Systems*, vol. 27, no. 01, pp. 1–18, 2019.

[69] A. K. Halder, P. Chatterjee, M. Nasipuri, D. Plewczynski, and S. Basu, "3gclust: Human protein cluster analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 6, pp. 1773–1784, 2019. DOI: 10.1109/TCBB.2018.2840996.

[70] S. Saha, A. K. Halder, S. S. Bandyopadhyay, P. Chatterjee, M. Nasipuri, and S. Basu, "Computational modeling of human-ncov protein-protein interaction network," *Methods*, vol. 203, pp. 488–497, 2022.

[71] *Covid-19 uniprotkb 2021.* https://covid-19.uniprot.org/, Last accessed on May 23, 2024.

[72] N. Pržulj, D. G. Corneil, and I. Jurisica, "Modeling interactome: Scale-free or geometric?" *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[73] N. Pržulj, "Biological network comparison using graphlet degree distribution," *Bioinformatics*, vol. 23, no. 2, e177–e183, 2007.

[74] T. Milenković and N. Pržulj, "Uncovering biological network function via graphlet degree signatures," *Cancer informatics*, vol. 6, CIN–S680, 2008.

[75] D. Bhattacharjee, S. M. M. Hossain, R. Sultana, and S. Ray, "Topological inquisition into the ppi networks associated with human diseases through graphlet frequency distribution," in *International Conference on Pattern Recognition and Machine Intelligence*, Springer, 2017, pp. 431–437.

[76] G. Alanis-Lobato, M. A. Andrade-Navarro, and M. H. Schaefer, "Hippie v2. 0: Enhancing meaningfulness and reliability of protein–protein interaction networks," *Nucleic acids research*, gkw985, 2016.

[77] T. Hočevar and J. Demšar, "Combinatorial algorithm for counting small induced graphs and orbits," *PloS one*, vol. 12, no. 2, e0171428, 2017.

[78] S. Bhatkar, P. Gosavi, V. Shelke, and J. Kenny, "Link prediction using graphsage," in *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, IEEE, 2023, pp. 1–5.

[79] M. Uzair and N. Jamil, "Effects of hidden layers on the efficiency of neural networks," in *2020 IEEE 23rd international multitopic conference (INMIC)*, IEEE, 2020, pp. 1–6.

[80] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun, "Measuring and relieving the over-smoothing problem for graph neural networks from the topological view," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 3438–3445.

[81] A. Pande, S. Patiyal, A. Lathwal, *et al.*, "Computing wide range of protein/peptide features from their sequence and structure," *BioRxiv*, p. 599 126, 2019.

[82] U. Consortium, "Uniprot: A worldwide hub of protein knowledge," *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.

[83] T. T. Ashburn and K. B. Thor, "Drug repositioning: Identifying and developing new uses for existing drugs," *Nature reviews Drug discovery*, vol. 3, no. 8, pp. 673–683, 2004.

[84] S. Pushpakom, F. Iorio, P. A. Eyers, *et al.*, "Drug repurposing: Progress, challenges and recommendations," *Nature reviews Drug discovery*, vol. 18, no. 1, pp. 41–58, 2019.

[85] *Pubmed*, `https://pubmed.ncbi.nlm.nih.gov/`, Last accessed on May 23, 2024.

[86] D. S. Wishart, C. Knox, A. C. Guo, *et al.*, "Drugbank: A knowledgebase for drugs, drug actions and drug targets," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D901–D906, 2008.

[87] K. P. Seiler, G. A. George, M. P. Happ, *et al.*, "Chembank: A small-molecule screening and cheminformatics resource database," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D351–D359, 2007.

[88] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders," *Nucleic acids research*, vol. 43, no. D1, pp. D789–D798, 2015.

[89] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "Genecards: A novel functional genomics compendium with automated data mining and query reformulation support.," *Bioinformatics (Oxford, England)*, vol. 14, no. 8, pp. 656–664, 1998.

[90] S. Saha, A. K. Halder, S. S. Bandyopadhyay, *et al.*, "Drug repurposing for covid-19 using computational screening: Is fostamatinib/r406 a potential candidate?" *Methods*, vol. 203, pp. 564–574, 2022.

[91] *Nadh-drugbank*, `https://go.drugbank.com/drugs/DB00157`, Last accessed on May 23, 2024.

[92] J. Shang, M. R. Smith, A. Anmangandla, and H. Lin, "Nad+-consuming enzymes in immune defense against viral infection," *Biochemical Journal*, vol. 478, no. 23, pp. 4071–4092, 2021.

[93] R. Kataria, S. Kaur, and R. Kaundal, "Deciphering the complete human-monkeypox virus interactome: Identifying immune responses and potential drug targets," *Frontiers in Immunology*, vol. 14, p. 1 116 988, 2023.

[94] K. A. Fontaine, R. Camarda, and M. Lagunoff, "Vaccinia virus requires glutamine but not glucose for efficient replication," *Journal of virology*, vol. 88, no. 8, pp. 4366–4374, 2014.

[95] Z. J. Leinen, R. Mohan, L. S. Premadasa, A. Acharya, M. Mohan, and S. N. Byrareddy, "Therapeutic potential of cannabis: A comprehensive review of current and future applications," *Biomedicines*, vol. 11, no. 10, p. 2630, 2023.

[96] G. Kampf, "Efficacy of biocidal agents and disinfectants against the monkeypox virus and other orthopoxviruses," *Journal of Hospital Infection*, vol. 127, pp. 101–110, 2022.

[97] S. Noe, S. Zange, M. Seilmaier, *et al.*, "Clinical and virological features of first human monkeypox cases in germany," *Infection*, vol. 51, no. 1, pp. 265–270, 2023.

[98] T.-P. Lam, V.-H. Tran, T. T. Mai, *et al.*, "Identification of diosmin and flavin adenine dinucleotide as repurposing treatments for monkeypox virus: A computational study," *International Journal of Molecular Sciences*, vol. 23, no. 19, p. 11 570, 2022.

[99] W. Arndt, C. Mitnik, K. L. Denzler, *et al.*, "In vitro characterization of a nineteenth-century therapy for smallpox," *PLoS One*, vol. 7, no. 3, e32610, 2012.

[100] A. İ. Sofuoğlu, N. Akçay, E. Şevketoğlu, and G. Bektaş, "Pseudotumor cerebri syndrome as a neurologic involvement of multisystem inflammatory syndrome in children: A case report," *Journal of Tropical Pediatrics*, vol. 67, no. 4, fmab075, 2021.

[101] N. Deb, P. Roy, A. Biswakarma, *et al.*, "Neurological manifestations of coronavirus disease 2019 and mpox in pediatric patients and their management: A state-of-the-art systematic review.," 2023.

[102] S. Singh, V. Srivastava, P. Godara, *et al.*, "An in-silico-based study identified peptide inhibitors that can block the egression of the monkeypox virus by inhibiting the p37 protein target," *Peptide Science*, vol. 115, no. 5, e24325, 2023.

[103] N. Nagata, M. Saijo, M. Kataoka, *et al.*, "Pathogenesis of fulminant monkeypox with bacterial sepsis after experimental infection with west african monkeypox virus in a cynomolgus monkey," *International Journal of Clinical and Experimental Pathology*, vol. 7, no. 7, p. 4359, 2014.

[104] M. Kamali Aghdam, H. Bakhtiari, D. N. Diaz, and K. Eftekhari, "Neurological manifestations of novel coronavirus disease in a 2-month-old infant: A case report," *Journal of Pediatric Neurology*, vol. 19, no. 03, pp. 210–212, 2020.

[105] D. Pal and K. Raj, "Poxvirus and the role of medicinal plants in controlling infections: Ethnopharmacology, chemistry, clinical and preclinical studies, and future perspectives," in *Anti-Viral Metabolites from Medicinal Plants*, Springer, 2023, pp. 373–395.

[106] A. Esqueda, H. Sun, J. Bonner, *et al.*, "A monoclonal antibody produced in glycoengineered plants potently neutralizes monkeypox virus," *Vaccines*, vol. 11, no. 7, p. 1179, 2023.

[107] A. A. Rabaan, A. H. Abas, T. E. Tallei, *et al.*, "Monkeypox outbreak 2022: What we know so far and its potential drug targets and management strategies," *Journal of Medical Virology*, vol. 95, no. 1, e28306, 2023.

[108] Q. Shu and V. Nair, "Inosine monophosphate dehydrogenase (impdh) as a target in drug discovery," *Medicinal research reviews*, vol. 28, no. 2, pp. 219–232, 2008.

[109] M. Spisakova, Z. Cizek, and Z. Melkova, "Ethacrynic and α-lipoic acids inhibit vaccinia virus late gene expression," *Antiviral research*, vol. 81, no. 2, pp. 156–165, 2009.

[110] P. Beaulieu, M. Desai, and N. Meanwell, "Successful strategies for the discovery of antiviral drugs," in *RSC Drug Discovery Series*, RSC Cambridge, 2013, p. 248.

[111] R. Sharma, K.-T. Chen, and R. Sharma, "Emerging evidence on monkeypox: Resurgence, global burden, molecular insights, genomics and possible management," *Frontiers in Cellular and Infection Microbiology*, vol. 13, p. 1 134 712, 2023.

[112] E. Khani, B. Afsharirad, and T. Entezari-Maleki, "Monkeypox treatment: Current evidence and future perspectives," *Journal of Medical Virology*, vol. 95, no. 1, e28229, 2023.

[113] A. Hallo-Carrasco, C. L. Hunt, C. C. Prusinski, *et al.*, "Pain associated with monkeypox virus: A rapid review," *Cureus*, vol. 15, no. 2, 2023.

[114] F. Pfäfflin, D. Wendisch, R. Scherer, *et al.*, "Monkeypox in-patients with severe anal pain," *Infection*, vol. 51, no. 2, pp. 483–487, 2023.

[115] A. Patel, J. Bilinska, J. C. Tam, *et al.*, "Clinical features and novel presentations of human monkeypox in a central london centre during the 2022 outbreak: Descriptive case series," *bmj*, vol. 378, 2022.

[116] D.-P. Wang, R. Zhao, H.-F. Wang, *et al.*, "Crystal structure of mrna cap (guanine-n7) methyltransferase e12 subunit from monkeypox virus and discovery of its inhibitors," *International Journal of Biological Macromolecules*, vol. 253, p. 127 565, 2023.

[117] R. O. Baker, M. Bray, and J. W. Huggins, "Potential antiviral therapeutics for smallpox, monkeypox and other orthopoxvirus infections," *Antiviral research*, vol. 57, no. 1-2, pp. 13–23, 2003.

[118] H. M. Al-Kuraishy, A. I. Al-Gareeb, H. F. Hetta, A. Alexiou, M. Papadakis, and G. E.-S. Batiha, "Monkeypox epidemic at the door: Should we remain idly by or prepare strongly?" *AMB Express*, vol. 13, no. 1, pp. 1–6, 2023.

[119] G. H. Hans, D. Wildemeersch, and I. Meeus, "Integrated analgesic care in the current human monkeypox outbreak: Perspectives on an integrated and holistic approach combining old allies with innovative technologies," *Medicina*, vol. 58, no. 10, p. 1454, 2022.

[120] A. Banik, S. R. Ahmed, S. B. Shahid, T. Ahmed, H. K. Tamanna, H. Marma, *et al.*, "Therapeutic promises of plant metabolites against monkeypox virus: An in silico study," *Advances in Virology*, vol. 2023, 2023.

[121] N. A. Mohamed, L. Zupin, S. I. Mazi, H. A. Al-Khatib, and S. Crovella, "Nanomedicine as a potential tool against monkeypox," *Vaccines*, vol. 11, no. 2, p. 428, 2023.

[122] B. Gurnani, K. Kaur, S. Chaudhary, and H. Balakrishnan, "Ophthalmic manifestations of monkeypox infection," *Indian Journal of Ophthalmology*, vol. 71, no. 5, pp. 1687–1697, 2023.

[123] Y. Jiao, C. Shi, and Y. Sun, "The use of xuanbai chengqi decoction on monkeypox disease through the estrone-target ar interaction," *Frontiers in Microbiology*, vol. 14, 2023.

[124] M. B. Slabaugh and C. K. Mathews, "Hydroxyurea-resistant vaccinia virus: Overproduction of ribonucleotide reductase," *Journal of virology*, vol. 60, no. 2, pp. 506–514, 1986.

[125] I. A. Guedes, C. S. de Magalhães, and L. E. Dardenne, "Receptor–ligand molecular docking," *Biophysical reviews*, vol. 6, pp. 75–87, 2014.

[126] O. Trott and A. J. Olson, "Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of computational chemistry*, vol. 31, no. 2, pp. 455–461, 2010.

[127] B. Gopal Samy and L. Xavier, "Molecular docking studies on antiviral drugs for sars," *International Journal*, vol. 5, no. 3, 2015.

[128] J. Silhan, M. Klima, T. Otava, *et al.*, "Discovery and structural characterization of monkeypox virus methyltransferase vp39 inhibitors reveal similarities to sars-cov-2 nsp14 methyltransferase," *Nature Communications*, vol. 14, no. 1, p. 2259, 2023.

[129] *Autodock vina manual*, `https://vina.scripps.edu/manual/`, Last accessed on May 23, 2024.

[130] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, "Large-scale prediction of human protein- protein interactions from amino acid sequence based on latent topic features," *Journal of proteome research*, vol. 9, no. 10, pp. 4992–5001, 2010.

[131] X. Wu, L. Zhu, J. Guo, D.-Y. Zhang, and K. Lin, "Prediction of yeast protein–protein interaction network: Insights from the gene ontology and annotations," *Nucleic acids research*, vol. 34, no. 7, pp. 2137–2150, 2006.