Computational Analysis of Large and Heterogeneous Biological Networks

Thesis submitted by

Soumyendu Sekhar Bandyopadhyay

Doctor of Philosophy(Engineering)

Department of Computer Science and Engineering
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata-700032, India
2023

JADAVPUR UNIVERSITY
KOLKATA 700032, INDIA

Index No: 99/19/E

1. Title of the Thesis:

   Computational Analysis of Large and Heterogeneous Biological Networks

2. Name, Designation & Institute of the supervisor:

   Prof. Subhadip Basu
   Professor,
   Department of Computer Science and Engineering,
   Jadavpur University, Kolkata-700032

3. List of Publications:

   a) **Journal**

      i) Sovan Saha, Anup Kumar Halder, **Soumyendu Sekhar Bandyopadhyay**, Piyali Chatterjee, Mita Nasipuri, Debdas Bose, and Subhadip Basu. "Drug repurposing for COVID-19 using computational screening: Is Fostamatinib/R406 a potential candidate?." ***Methods (2021) (IF:4.647)***.

      ii) **Soumyendu Sekhar Bandyopadhyay**, Anup Kumar Halder, Monika Zareba-Kozioł, Anna Bartkowiak- Kaczmarek, Aviinandaan Dutta, Piyali Chatterjee, Mita Nasipuri, Tomasz Wojtowicz, Jakub Wlodarczyk, and Subhadip Basu. "RFCM-PALM: In-Silico Prediction of S-Palmitoylation Sites in the Synaptic Proteins for Male/Female Mouse Data." ***International Journal of Molecular Sciences (2021) (IF:5.8)***.

      iii) Sovan Saha, Anup Kumar Halder, **Soumyendu Sekhar Bandyopadhyay**, Piyali Chatterjee, Mita Nasipuri, and Subhadip Basu. "Computational modeling of human-nCoV protein-protein interaction network." ***Methods (2021) (IF:4.647)***

      iv) **Soumyendu Sekhar Bandyopadhyay**, Anup Kumar Halder, Sovan Saha, Piyali Chatterjee, Mita Nasipuri, and Subhadip Basu. "Assessment of GO-Based Protein Interaction Affinities in the Large-Scale Human–Coronavirus Family Interactome." ***Vaccines (2023) (IF:7.8)***

b) **Preprint Archive**

   i. Anup Kumar Halder, **Soumyendu Sekhar Bandyopadhyay**, Witold Jedrzejewski, Subhadip Basu, and Jacek Sroka. ”FuzzyPPI: Human Proteome at Fuzzy Semantic Space.” ***BioRxiv (2023) https://doi.org/10.1101/2023.05.24.541959*** (Under review in IEEE Transactions on Big Data)

c) **Conference**

   i) **Soumyendu Sekhar Bandyopadhyay**, Anup Kumar Halder, Piyali Chatterjee, Jacek Sroka, Mita Nasipuri, and Subhadip Basu. ”Analysis of Large-Scale Human Protein Sequences Using an Efficient Spark-Based DBSCAN Algorithm.” In ***Proceedings of International Conference on Frontiers in Computing and Systems. Advances in Intelligent Systems and Computing, vol 1255. Springer, Singapore.*** , Organized by Jalpaiguri Government Engineering College, India, 2020 `https://doi.org/10.1007/978-981-15-7834-2_56`

   ii) **Soumyendu Sekhar Bandyopadhyay**, Anup Kumar Halder, Kaustav Sengupta, Piyali Chatterjee, Mita Nasipuri, Dariusz Plewczynski, and Subhadip Basu. ”Multi-Level Feature Based Sub-Cellular Location Prediction of Apoptosis Proteins.” In ***International Conference on Data, Electronics and Computing (ICDEC-2022)*****, North-Eastern Hill University Shillong, Meghalaya, India, 2022 (Accepted and Presented)**

4. List of Patents: None

5. List of Presentations in National/International/Conference/Workshops:

   i) **Soumyendu Sekhar Bandyopadhyay**, Anup Kumar Halder, Piyali Chatterjee, Jacek Sroka, Mita Nasipuri, and Subhadip Basu. ”Analysis of Large-Scale Human Protein Sequences Using an Efficient Spark-Based DBSCAN Algorithm.” In ***Proceedings of International Conference on Frontiers in Computing and Systems. Advances in Intelligent Systems and Computing, vol 1255. Springer, Singapore.*** , Organized by Jalpaiguri Government Engineering College, India, 2020 `https://doi.org/10.1007/978-981-15-7834-2_56`

   ii) **Soumyendu Sekhar Bandyopadhyay**, Anup Kumar Halder, Kaustav Sengupta, Piyali Chatterjee, Mita Nasipuri, Dariusz Plewczynski, and Subhadip Basu. "Multi-Level Feature Based Sub-Cellular Location Prediction of Apoptosis Proteins." In ***International Conference on Data, Electronics and Computing (ICDEC-2022)*****, North-Eastern Hill University Shillong, Meghalaya, India, 2022 (Accepted and Presented)**

Statement of Originality

I, Soumyendu Sekhar Bandyopadhyay registered on **03/06/2019** do hereby declare that this thesis entitled **"Computational Analysis of Large and Heterogeneous Biological Networks"** contains literature survey and original research work done by the undersigned candidate as part of Doctoral studies.

All information in this thesis have been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct. I have fully cited and referred all materials and results that are not original to this work.

I also declare that I have checked this thesis as per the "Policy on Anti Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software is 7%.

Soumyendu Sekhar Bandyopadhyay

Signature of Candidate

Date: 21.09.23

Certified by Supervisor:

Subhadip Basu.

21.09.2023

(Signature with date and seal)

Subhadip Basu, Ph.D.
Professor
Computer Sc. & Engg. Department
Jadavpur University
Kolkata-700032

PROFORMA-2

CERTIFICATE FROM THE SUPERVISOR

Index No: 99/19/E

This is certify that the thesis entitled **"Computational Analysis of Large and Heterogeneous Biological Networks"** submitted by Shri Soumyendu Sekhar Bandyopadhyay, who got his name registered on 03/06/2019, for the award of Ph.D. (Engg.) degree of Jadavpur University is absolutely based upon his own work under the supervision of Prof. Subhadip Basu, Department of Computer Science and Engineering, Jadavpur University, Kolkata and that neither his thesis nor any part of it has been submitted for any degree/diploma or any other academic award anywhere before.

Subhadip Basu.

Signature of the Supervisor
and date with Office Seal

21.09.2023

Subhadip Basu, Ph.D.
Professor
Computer Sc. & Engg. Department
Jadavpur University
Kolkata-700032

Dedicated to my family

ACKNOWLEDGMENT

Pursuing for the PhD degree has been a truly life-changing experience for me and it would have not been possible to achieve this goal without the support and guidance that I received from many people.

It is a great pleasure for me to express my respect and deep sense of gratitude to my advisor Dr. Subhadip Basu, Professor, Computer Science and Engineering Department, Jadavpur University for his continuous support during my PhD study and related research, his patience, motivation, and immense knowledge. Without his guidance and constant feedback this PhD research work would not have been successful.

I would also like to express my sincere regards to Dr. Mita Nasipuri, Former Professor, Computer Science, and Engineering Department, Jadavpur University for her constant inspiration Without her persistent help and optimism this dissertation would not have been possible. I am also grateful to her, as a coordinator of CMATER, Department of Computer Science and Engineering, Jadavpur University for allowing me to use the infrastructures of the laboratory.

I am thankful to Dr. Piyali Chatterjee, Professor and HoD, Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India. Her valuable input, suggestions, and constant feedback have made an invaluable contribution to my PhD work.

I express my sincere thanks to Prof. Mahantapas Kundu, Prof. Debotosh Bhattacharjee, Prof Ram Sarkar, Prof Nibaran Das for their support, and suggestions. It was a great learning environment.

I would like to thank Prof. Nandini Mukherjee, HoD, Department of Computer Science and Engineering, Jadavpur University, for making available all the departmental facilities and his valuable comments.

Words will be short to thank Dr. Anup Kumar Halder, Faculty of Mathematics and Information Sciences, Warsaw University of Technology, Warsaw, Poland, and also with the Centre of New Technologies, University of Warsaw, Poland, for his valuable input, suggestions, and constant feedback have made an invaluable contribution towards my PhD work.

I would like to thank Prof. Sujoy Bhattacharya, Dean, School of Engineering and Technology, Adamas University, Kolkata, and Prof. Sajal Saha, HoD, Dept. of CSE, Adamas University, Kolkata for their continuous support towards my Ph.D.

I would like to thank Prof. Moutushi Singh, HoD, Dept. of IT, CSE(IoT), CSE(IoTCSBT), Institute of Engineering and Management, Kolkata for her continuous co-operation and support towards my Ph.D.

I gratefully acknowledge the funding received towards my PhD work from the Department of Biotechnology, Govt. of India.

I am also profoundly grateful for the hard work of my co-authors and their substantial contribution to uplift the studies presented in this thesis. Thanks to Prof. Debdas Bose, Dr. Sovan Saha, Dr. Jacek Sroka, Prof. Dariusz Plewczynski, Prof. Jakub Wlodarczyk, Dr. Monika Zareba-Koziol, Dr. Tomasz Wójtowicz, Anna Bartkowiak-Kaczmarek, Kaustav Sengupta, Aviinandaan Dutta for all the brainstorming sessions. I am delighted to have worked with them and I look forward to working with them again.

I would like to thank Mr. Pabak Indu, Assistant Professor, Dept. of CSE(AIML) Institute of Engineering and Management, Kolkata, and Ms. Tanaya Das, Assistant Professor, Dept. of CSE, Adamas University, Kolkata for their huge co-operation and support towards my Ph.D.

I thank my fellow labmates Dr. Sourav Pramanik, Dr. Samir Malakar, Dr. Soumi Paul, Dr. Koushik Dutta, Koushiki Roy, Mr. Nirmal Das, Dr. Debapriyo Banik, Mr. Ratnadeep Dey, Dr. Neelotpal Chakraborty, Dr. Swarnendu Ghosh for all their continuous support and for all the fun we have had in the laboratory. It was a great experience learning with them.

I would like to express my deepest gratitude to my parents for their unwavering love, encouragement, and support throughout my academic journey. Without their constant belief in me and their sacrifices, this thesis would not have been possible.

I would like to express my thanks to my in-laws, my wife Sumita, my lovely daughter Aadita for their unwavering support and boundless love have been my constant source of strength throughout this challenging academic journey.

Last but not least, I would like to thank my almighty, my guruma for the blessings provided to me to make the journey successful.

Soumyendu Sekhar Bandyopadhyay

(Soumyendu Sekhar Bandyopadhyay)
Registration No: 1021904005 of 2019-2020
Department of Computer Science & Engineering
Jadavpur University

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

A paradigm that describes the flow of genetic information across a living organism is called the Central Dogma of Molecular Biology. In its most fundamental form, it may be broken down into three distinct processes: transcription, translation, and replication. Converting a section of Deoxyribonucleic acid (DNA) into messenger Ribonucleic acid (mRNA) is referred to as transcription. The process by which mRNA is converted into the protein it encodes is called translation. The process of making duplicates of DNA is referred to as replication. The genetic material in the majority of species is made up of DNA; however, there are other organisms, such as retroviruses like HIV, that use RNA as their genetic material instead. The creation of DNA from RNA in these kinds of organisms takes place via a process known as reverse transcription. The process of convering a protein from DNA has been described in Figure 1.1

1.1.1 DNA to RNA to Protein

DNA is a molecule with a double helix structure made up of chains of nitrogenous bases and a sugar-phosphate backbone. The Base pairing between the nitrogenous bases in DNA molecules where Guanine interacts with Cytosine with three hydrogen bonds and Thymine interacts with Adenine with two hydrogen bonds. The double

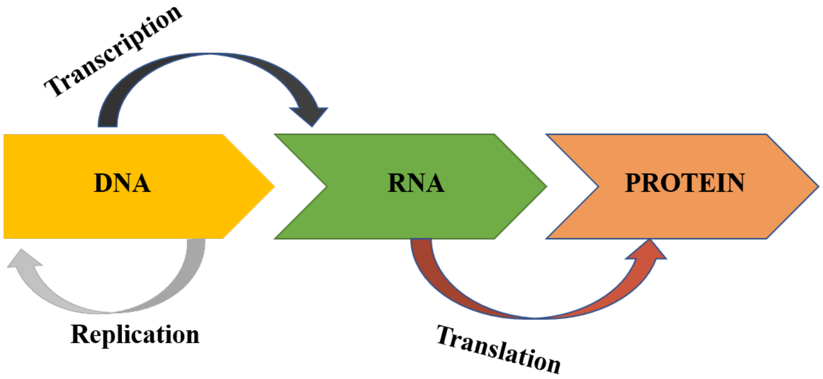


Figure 1.1: Central Dogma of Molecular Biology. The figure depicts the scheme for the construction of proteins from DNA molecules.

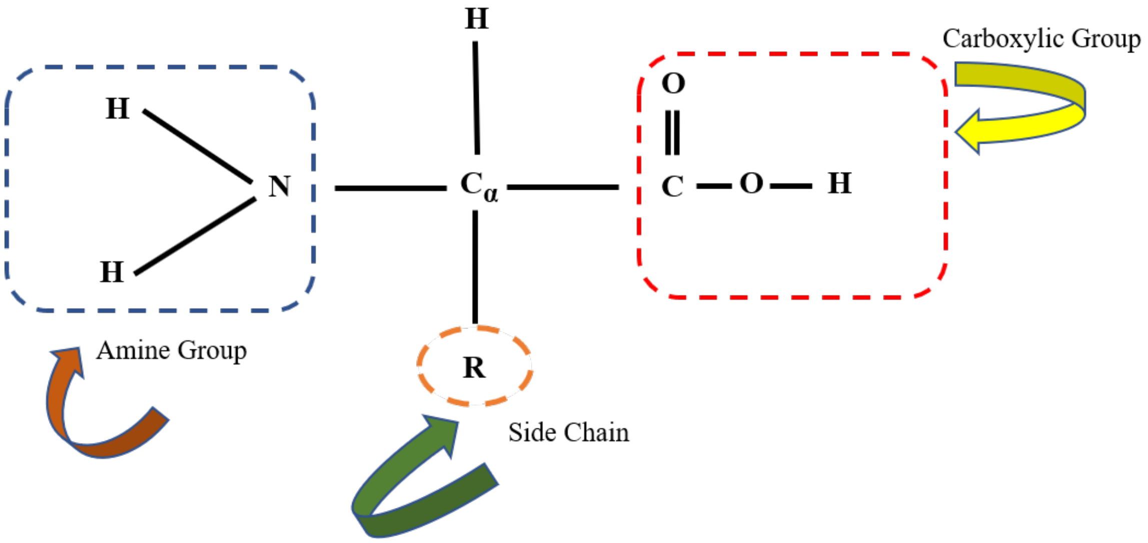


Figure 1.2: The structure of a single amino acid where alpha carbon is attached with a carboxylic group-$COOH$, an amine group (-NH_2), and a side chain R which differentiate the chemical properties of different amino acid

helix structure aids in self-replication since it is self-complementary. Evolution is the result of flaws in the replicating process.

Proteins and RNA are synthesized to carry out instructions from DNA. In most organisms, replication into RNA accounts for less than 1% of the total DNA. Most of the remaining 99% of our genetic material are either used for regulatory processes or is completely useless which is also known as "junk" DNA.

The next step, translation, is when the mRNA is actually "translated" into a protein. Triplets of nitrogen bases are the building blocks of the amino acid code. There are 4^3=64 potential triplets in RNA due to its four bases. Each of the twenty amino acids may be represented by a unique triplet or codon. As is clear, amino acids may be encoded by more than one codon. In addition, there is a single codon that indicates the final translation location.

1.1.2 Amino Acid

Amino acids are chemical molecules of biological significance; they are made up of an amine (-NH_2) and a carboxylic acid (-COOH) functional group (*see* Figure 1.2). Codons are what determine which amino acids may be produced. The involvement of proteins in almost all cellular processes makes amino acids crucial to survival. There may be hundreds of amino acids in the natural world, yet only 20 are used to make proteins. A protein is the result of the aggregation of a number of amino acids.

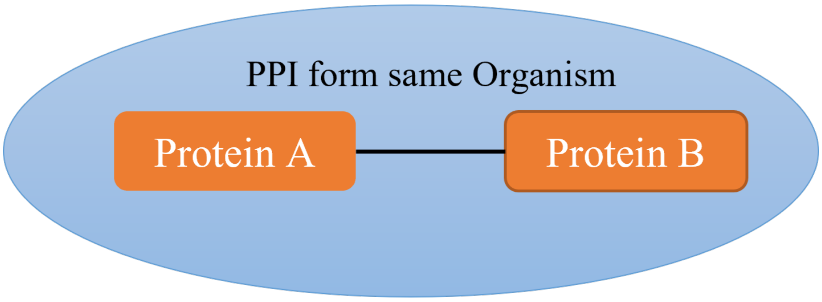


Figure 1.3: Intra-species PPI where both the proteins are from the same organism.

1.1.3 Proteins

Most of the cellular function is performed by proteins, which are linear polymers of amino acids. Proteins may have anything from several hundred to several thousand individual amino acids held together by peptide bonds. The nitrogen atom at the carboxyl terminus of one amino acid forms a peptide bond with the carbon atom at the amino group of another amino acid, as shown in Figure 1.3. Enzymes are proteins that speed up chemical processes in the body. *Actin* and *myosin* are two more proteins that help with structural or mechanical tasks. There are additional proteins that handle things like cell signaling, cell adhesion, immunological responses, and the cell cycle.

1.1.4 Protein-Protein Interaction Network

The physical relationship of two or more proteins is referred to as protein-protein interaction (PPI) and plays a critical role in the regulation of cellular activities, signaling pathways, and disease mechanisms [1, 2]. Modulating protein function, localization, stability, and activity may be achieved by PPIs involving enzyme-substrate, receptor-ligand, and protein-protein complex forms. PPIs are crucial for cellular homeostasis and function since they are involved in a wide variety of biological activities such as signal transduction, gene regulation, protein trafficking, and cell signaling [1–3].

To decipher the intricate chemical interactions that control cellular processes, knowledge of the mechanisms behind PPIs is essential. Protein domains, motifs, and surfaces all play important roles in PPIs by allowing proteins to recognize and interact with one another [4, 5]. PPIs originate and remain stable because to electrostatic interactions including hydrogen bonding, salt bridges, and Van der Waals forces, which mediate electrostatic complementarity between interacting proteins [5, 6]. Protein structural changes, flexibility, and dynamics may also alter PPIs, either promoting or inhibiting PPI [7, 8].

The connections between proteins in a cell or organism are represented by protein-

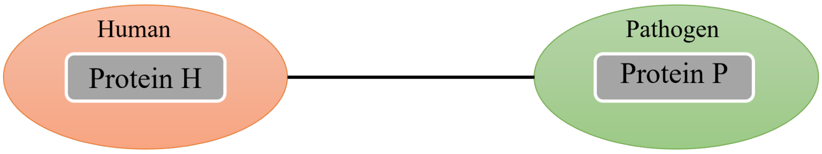


Figure 1.4: Inter-species PPI where two different proteins are from two different organisms.

protein interaction networks (PPINs), which are complex and ever-changing networks. To put it simply, PPINs are essential for controlling many different cellular activities, signaling pathways, and responses to stimuli [9, 10]. They help to understand cellular behavior, disease causes, and drug development by revealing how proteins are organized and coordinated within a cellular environment. Functional modules, hubs, and essential regulatory components may be uncovered by analyzing PPINs, which are often represented as nodes or proteins linked by edges or interactions [11, 12]. PPINs are produced using a range of experimental methodologies, such as yeast two-hybrid (Y2H), co-immunoprecipitation (Co-IP), mass spectrometry (MS), and fluorescence resonance energy transfer (FRET), among other techniques [13, 14]. The utilization of experimental methodologies results in the production of extensive datasets pertaining to PPIs, which can subsequently be scrutinized to construct all-encompassing PPINs [15]. The analysis of PPINs frequently involves the application of computational techniques, including network-based methodologies, graph theory, and machine learning algorithms. These approaches have been instrumental in uncovering valuable information pertaining to protein function, protein complexes, signaling pathways, and networks associated with the disease [11, 12, 16].

The examination of PPINs has resulted in the recognition of pivotal proteins, commonly referred to as hubs or network nodes, that serve as intermediaries between various constituents of the network and exert a significant impact on its global operation [17]. The aforementioned hubs have been identified as plausible therapeutic targets for drug discovery and have been associated with a range of illnesses, such as cancer, neurodegenerative disorders, and infectious ailments [18]. Additionally, the examination of PPINs has demonstrated the significance of modular arrangement, wherein proteins exhibiting comparable functions have a tendency to aggregate, creating functional modules or complexes that regulate cellular processes [19]. A way of expressing systems as complicated sets of binary connections or relations between distinct biological components is called a biological network. Biological networks can be classified based on their type and detection method. Based on the type, a biological network can be classified as follows. Figure 1.5 shows different classifications of biological networks.

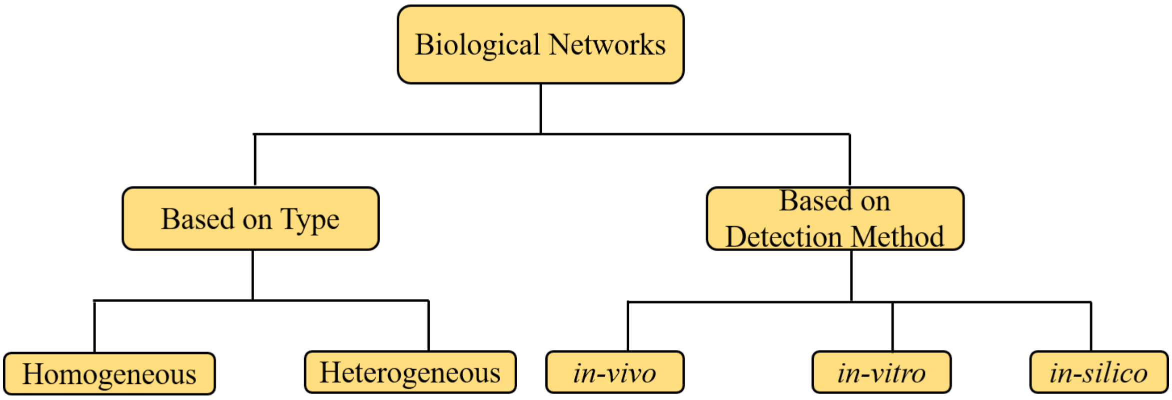


Figure 1.5: Classification of different biological networks.

- **Homogeneous Biological networks:** In the case of homogeneous protein interactions, the interacting proteins are both from the same species or origin, meaning that their amino acid sequences and overall structures are almost identical. As a result of their shared primary, secondary, and tertiary structures, the interacting proteins may be said to belong to the same protein family. The dimerization of transcription factors, proteins that bind to DNA and control gene expression, is a well-studied example of homogenous protein interaction [20].

- **Heterogeneous Biological Networks:** The term heterogeneous protein interactions pertains to a specific category of protein-protein interaction wherein the proteins involved are of diverse species or lineages. This means that the proteins are distinct entities originating from different protein families, organisms, or cellular compartments. To clarify, it can be observed that the proteins that interact with each other do not possess identical or substantially similar amino acid sequences and structures. Additionally, these proteins may exhibit distinct functions or perform varying roles within the cellular environment. Heterogeneous protein interactions are essential in various biological processes, including but not limited to cell signaling, metabolism, and immune responses. The phenomenon under consideration pertains to the interactions that occur between proteins originating from distinct cellular compartments. These interactions may involve membrane receptors and cytoplasmic signaling proteins, or nuclear proteins and cytoplasmic factors. The significance of heterogeneous protein interactions lies in their ability to facilitate cellular communication, coordination, and regulation. This enables diverse proteins to collaborate in a synchronized manner, thereby accomplishing intricate cellular processes. One instance of heterogeneous protein interaction can be observed in the interaction between enzymes and their respective cofactors or substrates. Enzymes are a class of proteins that facilitate biochemical reactions within cells. Their activity is frequently dependent on the presence of specific cofactors or substrates [21].

Based on the detection method, the biological network can be classified as follow:

- ***in-vivo:*** The networks of protein complexes that have been generated as a result of biochemical processes or electrostatic forces and which perform a specific biological function as a complex [15].

- ***in-vitro:*** A certain technique is carried out in an environment that is under strict control and is located away from a live creature. Tandem affinity purification, affinity chromatography, coimmunoprecipitation, protein arrays, protein fragment complementation, phage display, X-ray crystallography, and nuclear magnetic resonance spectroscopy are the *in-vitro* approaches for PPI identification [15].

- ***in-silico:*** *In-silico* methods are those that are carried out on a computer or via the use of computer simulation. In the process of PPI discovery, *in-silico* techniques include sequence-based approaches, structure-based approaches, chromosomal proximity approaches, gene fusion approaches, *in-silico* 2 hybrid approaches, mirror tree approaches, phylogenetic tree approaches, and gene expression-based approaches [15].

1.1.5 Post-Translational Modification

In biological systems, proteins perform a wide range of catalytic, regulatory, signaling, and structural tasks. Most eukaryotic proteins undergo post-translational changes after being assembled on ribosomes during their whole lifespan [22]. Post translational modifications (PTMs) are the changes that occur by different biomolecules in the amino-acid side chain. The polypeptide backbone is frequently supposed to be inert, although PTMs have traditionally been conceived of being reversibly attached on amino acid side chains to nucleophilic functional groups. As new chemical and functional variations of the protein backbone are found, this paradigm is changing. Importantly, backbone PTMs act in distinct ways to achieve these properties while being able to affect protein structure and function in a manner similar to side chain changes [23]. Comparing the number of genes and the number of proteins generated from the corresponding genes it has been observed that the number of encoded proteins outnumbers the number of genes due to PTMs [24]. PTM crosstalk indicates the coordinated action of several PTMs on one or more proteins for higher-level regulation [25].

PTMs can happen at different phases of the protein life cycle. Additionally, many proteins have autocatalytic domains that allow them to change themselves. PTMs have been proven to be important participants in the etiology of many illnesses, such as cancer, diabetes etc., and play a crucial role in physiological processes occurring in a healthy cell [24].

PTMs enable cells to control protein activities, transmit signals, and react to disturbances. PTMs enhance the variety and functioning of proteins, which raises the complexity of the proteome [25]. It has been demonstrated to work via altering PPIs. In comparison to non-PTM proteins, proteins undergoing a PTM were shown to participate in more interactions and to occupy more central positions [26].

1.2 Brief Literature Survey

The fundamental building blocks of a living thing are proteins. All living species depend heavily on the interactions between proteins, which carry out specific tasks and are essential for cellular and biological function. Since several PPIs take place throughout a biological process, focusing on the precise set of interactions aids in understanding the molecular mechanism behind the specific biological activity and helps to assign roles to unidentified proteins. Transporters, molecular machinery, and molecular chaperones are just a few examples of complexes that are created when two or more proteins physically bind to sustain the assembly [27]. The experimental techniques can address a limited portion of PPIs for an organism. Proteome scale PPI analysis for any full organism becomes impractical and time-consuming because to the significant computing overhead. To solve these problems, effective algorithms that work with high-throughput parallel architecture must be created. Different computational analyses are presented in biological research with two goals in mind: either to help natural systems looking for biological answers or to offer a computational solution for biological systems. In this section, previous research works are discussed which include different methodologies for *in-silico* analysis and different applications of PPIN

1.2.1 Clustering Large Biological Data

Proteins interact with other proteins to perform a wide range of essential biological and cellular tasks. Proteomic data has been expanding at an exponential rate in recent years. The exponential expansion in meta-genomic sequence accumulation enabled by contemporary high-throughput sequencing methods has the potential to improve large-scale functional annotation significantly. Researchers now have a significant barrier in processing these lengthy and repetitive sequences. One of the primary actions taken to lessen the duplication of these massive resources and the analysis of such big biological sequences is clustering by similarity.

Big Data, as the name implies, is a concept that promises to manage large data volumes quickly by using several processors housed on multiple nodes. This huge dataset necessitates parallel processing across several nodes. To process large datasets, Apache Software Foundation came up with a framework, Apache Hadoop, an open-source framework with a cluster of commodity hardware. Hadoop provides a module called MapReduce which distributes the computing among the clusters and combines the result on completion and also provides a distributed file system known as Hadoop Distributed File System (HDFS) which is a specially designed file system for a huge dataset with a cluster of commodity hardware and streaming access patterns.

Hadoop lacks some computational issues such as *in-memory* computations, real-time computations, etc. To address the issues, Apache Software Foundation introduced Apache Spark which accelerates the computational speed by using *in-memory* computation with a new memory abstraction called Resilient Distributed Dataset (RDD). Both Apache Hadoop and Apache Spark frameworks play a pivotal role in clustering large protein datasets.

Protein sequences may be clustered using a number of different methods [28–31]. Partition-based clustering, hierarchical clustering, and density-based clustering are the three main types of sequence clustering analysis. Decomposing a dataset into a collection of independent parts, each of which stands in for a cluster, is the basis of partitional clustering. On the other hand, hierarchical clustering structures the data into a tree structure with several levels of nesting. A cluster that is based on density refers to a set of objects that are connected through density and is considered maximal concerning density-reachability. Any object that is not part of a cluster is considered to be noise. DBSCAN is one of the most popular density-based classification algorithms. DBSCAN algorithm explores clusters by examining the ϵ-neighborhood of each point within the database. DBSCAN is a density-based spatial clustering approach that explores the neighborhood within a radius and finds out the minimum number of neighborhood points within the radius. From a set of objects, each object is checked with the core object point to determine whether the object is density reachable or not. The concept of density reachability is based on the transitive closure of direct density reachability, which exhibits an asymmetric relationship. Only core objects exhibit mutual density reachability. The relationship of density connectivity exhibits symmetry. A density-based cluster refers to a collection of objects that are connected by density and are considered maximal in terms of density-reachability. Any object that is not assigned to a cluster is considered to be noise [32].

1.2.2 Large-Scale Interaction of Human Proteome

After the Human Genome Project was finished, there was a massive increase in genetic sequence data. Proteomic data, which includes measurements of protein abundance across a variety of biological contexts, has grown at an exponential rate to keep pace with genomic data. Sequence annotation, structural details, ontology links, annotations of function, interactions, PTMs, and illness associations are all included. More than 180 million un-reviewed proteins and around 500 thousand manually annotated and reviewed proteins are accessible via UniProt, a publicly available collection of protein data. Together, these proteins have the capacity to produce between 124 million and 15 billion protein interaction pairings [33]. Other well-known interaction databases, such as DIP, catalog over 80,000 interaction pairs involving 28,850 proteins across 834 species [34]. About 400,000 Human protein-protein interactions are also included in BioGRID's massive database of over a million [35]. About one million interactions involving 113,926 proteins are stored in IntAct [36]. There are around 30,000 human proteins involved in the three billion interactions found in STRING's database of 24.6 million proteins from 5,090 species [37].

The study of proteins, particularly their structures and activities as well as the analysis of their interconnection networks, is known as computational proteomics. By feeding sequence-level data to computers like neural networks and support vector machines (SVM) that have already been trained with information taken from biologically determined structures of protein sequences, it is possible to predict the structures of protein sequences. The same approach may be used to estimate the relative solvent accessibility of individual amino acids in protein sequences. Only 13 protein structures were saved in PDB's database in the 1970s, but because of improvements in structure detecting techniques, the database today has more than 100,000 protein sequence structures [38]. Prediction of protein structures in excess of 90% accuracy is an aim that is yet to be achieved [39]. PPINs have been explored extensively to understand how proteins interact together to form complexes. Researchers have been working on Pairwise PPIN Alignment [40] and Multiple PPIN Alignment [41]. Both these require heavy computational resources due to the sheer volume of data available. Understanding the function of a protein will help us uncover the root of several diseases. The function of a protein can be predicted by several methods such as homology-based [42,43], sequence motif-based [44,45], structure-based [46–48] and network-based methods [49]. All these methods have to rely on a huge set of data in order to provide accurate function prediction. Every biomolecule in the world functions by interacting with other biomolecules. Similarly, DNA or other biomolecules may dock with proteins at certain sites. Accurate prediction of such sites will help eliminate diseases and facilitate the proper functioning of the protein. A lot of research has been dedicated to the prediction of such sites [50,51]. Just like other areas of computational biology, the problem of protein binding site prediction also relies on a huge amount of non-trivial data for accurate prediction of such sites. Currently, several works are trying to draw inferences about species not only from sequence but also from interacting networks [52]. The problem of finding conserved structures between multiple PPINs is computationally very heavy since it requires the matching of networks of proteins. MIT researchers have created a database IsoBase [53], which identifies functionally related proteins across five major eukaryotic model organisms: Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditis elegans, Mus musculus, and Homo Sapiens. They are further trying to extend the implementation of Isobase to Genetic Interaction Network Alignment. Ay et al. have developed an algorithm SubMap [54] where they have aligned two metabolic pathways using a mixture of homology and topological similarity. Further the database BIND [55] archives biomolecular interaction, complex and pathway information.

1.2.3 Host-Pathogen PPIN

Controlling highly contagious viral illnesses requires the identification of possible virus-host interactions. This might aid in the creation of novel medications to treat viral infections [56]. Because of viruses' unidentified harmful processes, infectious illnesses continue to be one of the most common and serious causes of mortality in people [57]. Here, the molecular interactions between the virus and its host are crucial. Thus, for a better understanding of the process of infection and the pathophysiology of infectious disorders, virus-host PPIs are essential [58, 59]. There have been a number of advancements in PPI predictions across different types of animals. These PPI prediction techniques provide crucial data for further examining the transmission of illness between other species. Pathogenic microorganisms can leverage host capabilities and elude host immune responses by manipulating host processes thanks to PPIs between viruses and host proteins. Varied information about proteins such as sequence [60–64], domains [65–68], structure [69, 70], and virus host protein interactions [71] are used for computational prediction of PPIN.

To anticipate new host-virus PPIs, many computational techniques have been developed. In innovative host-virus interaction predictions, a variety of predictive models have been put out depending on the knowledge about interactions that are currently available. Several researchers have proposed different models for the prediction of host-pathogen PPIN. A linear motif-based prediction model has been proposed by

Bacerra *et al.* [72] to predict HIV-Human PPIN. Segura *et al.* [73] suggested employing motif-domain interactions to represent the network of human-pathogen interactions. Human viral proteomes were clustered using structural sequencing by Kharrat *et al.* [74]. A bi-clustering strategy was put out by Mukhopadhyay *et al.* [75] to forecast HIV-1-infected human proteins using interaction-based analysis. By using the bi-clustering approach, a set of rules for association was extracted from the interaction of HIV-1 proteins. Mukhopadhyay *et al.* [76] made additional advancements to their work by adding type and pattern-based biclustering to already existing interactions to anticipate novel host proteins. Hierarchical clustering using the protein sequence of EBOV and Influenza virus was proposed by [77]. Phosphorylation clustering was used to analyze the infection on the protein of bronchitis virus in a work by Spencer *et al.* [78]. On the predictions of Salmonella-human interaction, several methods were put out [79, 80]. Virus-host PPI prediction studies have effectively used SVM-based different techniques with varied input features [81, 82]. Mei et al. [57] suggested a transfer learning-based method with three distinct classifiers, where each classifier was applied to three characteristics based on the gene ontology (GO) database [83]. According to reports, molecular mimicry performs a crucial part in host-pathogen interactions, when a viral protein imitates a host protein's structural binding surface. As a consequence, the viral protein attaches to a different host protein in a competitive manner and spreads throughout the host [84–87]. Assessment through semantic similarity-based method using GO annotations plays a significant role in interacting host-pathogen proteins [88–91].

1.2.4 PTM in Protein Sequences

The chemical alterations that take place after a protein is generated are referred to as PTM of proteins. It can affect the electrophilicity, interactions, and structural properties of proteins. Nearly all proteins undergo PTM, which have a profound impact on the structure and dynamics of proteins and is critical to many biological processes [92]. Depending on the addition of the different chemical groups to the amino acid side chain, a number of computational methods have been developed and studied for different PTMs (*e.g.,* phosphorylation, Glycosylation, Sulfation, Acetylation, Palmitoylation etc).

As many proteins undergo PTMs, the proteome is actually more complicated than it would otherwise be. PTMs include the covalent alteration of the amino acid sequences, and they have the effect of expanding the range of potential protein species. [93]. PTMs are essential in controlling various cellular processes in response to certain physiological demands [94] *e.g.,* sub-cellular localization of proteins [95, 96], stability of proteins [97], and regulates enzymes [98]. It has been identified that about 5% of the human genome codes for enzymes that catalyze the process of PTMs in protein sequence leading to alteration of protein [99]. These chemical changes of proteins, such as phosphorylation, acetylation, methylation, carboxylation, or hydroxylation, are frequently regulated by enzymes [100].

Experimental methods for mapping and measuring PTMs have advanced significantly in the last 30 years. For example, protein-based analysis using liquid chromatography (LC) and mass spectrometry (MS) allows the discovery of hundreds of PTMs across whole proteomes [101]. However, the lack of understanding of the underlying enzymatic processes and their impacts on stable and dynamic proteins continues to restrict the detection and characterization of PTMs. In this context, *in-silico* approaches, which are frequently based on the most recent understanding of PTMs, are a potential method to conduct preliminary analysis and prediction that can supervise *in-vivo* and *in-vitro* research and contribute to the function of PTMs in cellular processes [92].

The experimental techniques to locate PTM sites are expensive and time-consuming. The need for computational approaches is highly significant [102]. Different high-throughput experimental techniques, such as mass spectrometry [103], microarray of protein sequences [104], and phosphor-specific proteolysis [105] has been used to study PTM sites. Numerous studies have shown that sequence-based prediction techniques, such as predicting protein localization, [106], identifying membrane proteins [107,108], identification of functional classes of enzyme [109]. Sequence alignment-based predictions of protein 3D structures can promptly offer an enormous amount of information and understanding for both fundamental research and therapeutic development [110].

The SVM has been successfully used in computational biology and bioinformatics to solve pattern recognition challenges, such as predicting protein subcellular localization [106], PTMs [111]. [112] proposed that SVM also performs well when used to predict PTM sites while predicting phosphorylation sites using the conventional binary encoding strategy, and the results were superior to those of all other techniques. SVM has been used by different researchers for predicting kinase-specific phosphorylation sites [113, 114], methylation sites [115], phosphorylation using k-spaced amino acid [116].

The bioinformatics research community is now particularly interested in the automated prediction of PTM locations. Based on the various types of categorization algorithms in use, the four primary classes of PTM prediction tools that are now accessible may be identified. The first category consists of tools that are generally connected to PTM, such as Eukaryotic Linear Motifs (ELM), which quickly search regular expression patterns to forecast ELMs in protein sequences [117]. An online tool called PROSITE makes a variety of PTM predictions based on consensus sequence patterns [118]. The Scansite tool forecasts kinase-specific and motif-relevant to signal transduction. The conserved sequence motifs serve as traces of significant biological or biochemical tasks carried out by such proteins [119]. Plewczynski *et.al.* [120] proposed a SVM-based automotive server for predicting PTM sites. Basu *et.al.* [121] upgrades the work by proposing AMS 3.0 using a multilayer perceptron classifier. The work was further extended by Plewczynski *et.al.* [122] where the features of 88 different PTMs are clustered based on high-quality index and the performance was measured using a multilayer perceptron classifier.

1.3 Motivation Behind the Current Study and Research Gap Analysis

Computational Strategies for Analysis of Large-Scale Biological Data

In spite of a number of methods [28–31] developed in the last decades, sequence clustering from large sequence data is still a serious challenge. Depending on the clustering structure, there are basically two approaches: hierarchical clustering (HC) and greedy heuristic flat clustering (GHFC). HC-based approach [123] organized the sequence in the form of a clustering tree keeping different levels of operational taxonomic units (OTUs) at various similarity levels that can explore more biological significance. The HC-based approach still has some limitations as it requires pairwise distance computation and storing of the distance matrix. On the other hand, the popular GHFC approach includes Cd-hit [30] and UCLUST [29] which reduces the computational complexity. But GHFC produces lower cluster quality than that of the HC-based clustering approach.

Clustering algorithms should be scalable to vast sequence data employing parallel computation and high-performance computing architecture. However, redesigning current procedures in parallel mode is challenging since each iteration distance calculation fully depends on its earlier step or each division/merging relies on the preceding operations. Thus, in high-dimensional sequence data clustering, parallel density-based clustering may be an efficient alternative.

Large-Scale Interaction of Human Proteome at Fuzzy Semantic Space

Proteins are essential to the structure of any living thing. Proteins are essential for all biological processes because of the role they play in cellular communication and PPIs. The identification of the correct interaction set aids in the decipherment of the molecular mechanism underlying a specific biological event and the assignment of roles to unknown proteins [27]. The PPIN of a whole organism may tell us a lot about how its cells and molecules perform, what signals they send, how diseases develop, and how they are treated.

In-silico analysis of the PPIN can be constructed in two ways: i) prediction-based and ii) assessment-based. The limitation of statistical prediction lies in verifying the PPIs identified by the experiment to handle errors in experimental results such as false positives or false negatives [27]. Most of the *in-silico* prediction methods work in binary decision mode, either interacting or not. The binary mode PPIs are often less efficient in pathway analysis and/or disease-related analyses as there is no indication of interaction affinity between two partners [27,124,125]. These predictions-based approaches suffered from bias-free learning and more appropriately depend on the train-test data. However, the selection of negative data and bias-free cross-validation strategy are the major factors in PPI prediction that are overlooked in most of the prediction models [126, 127]. In addition, these experimental methods can elucidate a smaller subset of the PPI network of an organism. Due to the huge computational overhead, proteome scale PPI investigation for any complete organism becomes infeasible and time-consuming. So, efficient algorithms, compatible with high-throughput parallel architecture need to be designed to overcome these issues. On the other hand, assessment-based studies effectively help to compute the protein interaction affinities between the protein pairs without any prior train-test based predictive evaluation. GO-based semantic assessment strategy is one of the major schemes to evaluate the interaction affinity [128] as it includes different level hierarchical relationships of functional annotation. However, with the increasing number of proteins, the growth of possible PPI relations improves in exponential order.

The above study necessitates the need for a protein interaction network for better and bias-free *in-silico* prediction-based outcomes of cellular and molecular functionalities [129], reconstructing signaling pathways [130, 131], disease associations analyses [56, 132], drug discovery [133]. Till date, there is no evidence of organism-specific protein interactome. Moreover, recent studies show that weighted interactions are proffered in pathway and/or disease analyses [134]. The above study motivates to develop an efficient algorithm to compute the fuzzy binding affinity for large-scale PPIs using a high throughput parallel architecture. The development of such an effective algorithm to compute interaction affinity at the organism level eventually helps to construct a semantic network that could be a key node to assess different cellular and molecular mechanisms, disease analysis, host pathogenic relationships, drug target improvement, etc.

Human-nCoV PPIN

The discovery of host-pathogen PPINs has important implications for the knowledge of infection transmission mechanisms, which in turn underpins the rational medication design that is so crucial to the advancement of therapies. Proteins from both the pathogen and the host work together to promote infection and illness. Infectious disease is actively propagated by the pathogen. The pathogen-host PPIN interaction facilitates pathogen exploitation of host resources and host manipulation of host systems for immune evasion [65,135,136]. The primary focus of studies is the identification of target proteins by comparing pathogen and host PPIN [56,137,138]. Important pharmacological targets are often topologically significant proteins with a large number of interactions. However, some biological pathway relevance suggests that proteins with fewer interactions or topologically insignificant features may be involved in the infection mechanism.

A single-stranded RNA virus known as SARS-CoV-2 has an estimated particle size of 27 to 32 kb and its diameter range is between 65 and 125 nm [139]. The symptoms of COVID-19 in people can range from a cough and cold to potentially lethal signs such as respiratory infections, pyrexia, dyspnea, and multi-organ failure [140]. A comprehensive investigation of the genetic characteristics of SARS-CoV-2 using host-virus PPI might discover potential treatment targets [141, 142]. PPI examines the protein connections between humans and viruses in the host-pathogen interaction study and identifies viral infections and host resistance responses [143]. The use of machine learning algorithms in the prediction of PPI is an interesting and popular area of research where the machine learning classifiers effectively distinguish between protein pairs that interact and those that do not interact using binary classification [144]. Different machine-learning algorithms have been used to identify different host-pathogen PPIN which include zika virus [145,146], HIV [147], SARS-CoV2 [148]. By amplifying viral RNA in real-time reverse transcriptase polymerase chain reaction (rRT-PCR), Brinati *et al.* [149] used machine learning to identify the COVID-19 infection from regular blood test analysis. An ensemble machine-learning technique has been used to determine human-nCoV PPI by Chakraborty *et.al* [150] using five different protein sequence-based features and the result has been validated through web-based tools.

Not only machine learning algorithms but a graph theoretic approach is also useful for determining protein functions [151–153] and spreader proteins [154–157] in a PPIN. Centrality analysis is used to determine the PPIN's transmission capacity and compactness. Betweenness centrality, a novel centrality metric introduced by Anthonisse in 1971 [158]. Sabidussi [159] introduced the concept of closeness centrality, another measure of centrality. Degree centrality [154] and local average centrality [156] are two more crucial centrality measures that are also discovered to be highly useful in this field of study. Thus, identifying target proteins by analyzing PPIN from the pathogen and host, is discovered as the topologically significant proteins with a greater degree of interaction which become major therapeutic targets. However, due to some biological pathway importance, proteins with less interaction or with insignificant topologies could be implicated in the process of infection.

Severe Acute Respiratory Syndrome 2 (SARS-CoV-2) has 89% genetic similarity with Severe Acute Respiratory Syndrome (SARS-CoV). Based on this hypothesis, the above study motivates the development of an *in-silico* human-nCoV PPIN model which can identify the possible spreader proteins of SARS-CoV-2. Possible level-1 and level-2 spreader proteins can further be used to identify the concerned potential Food and Drug Administration (FDA)-approved drugs for COVID-19 treatment.

Drug Repurposing for COVID-19 using in-silico Methods

The respiratory system of animals, including humans, is often affected by coronaviruses, which can cause moderate to severe respiratory tract infections [160]. Two highly pathogenic Human Coronavirus (HCoVs), Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and SARS-CoV, which emerged from animal reservoirs in the past 20 years, have caused widespread epidemics with significant morbidity and fatality rates [161]. In the year 2019, a new coronavirus disease emerged in Wuhan, China, and was abbreviated as COVID-19/nCoV [162]. As nCoV was new to humans, at the time of the research, there were no known drugs that could be used to eradicate COVID-19. To prevent and cure COVID-19, several national and international research teams were developing vaccinations. Treatments with various antiviral medications are taken into consideration and used to end COVID-19 based on prior knowledge of big attacks such as major Ebola, cholera, etc. outbreaks. Thus drug repurposing became one of the fundamental research areas for the treatment of COVID-19.

The fastest approach, drug repurposing, appears to be using an already-approved medication or a substance that is currently through clinical trials for the treatment of COVID-19, even though the discovery and evaluation of a new drug should take even longer. This is because these compounds have either received regulatory approval as drugs or have passed safety studies that indicate a therapeutic potential [163]. FDA-approved medications were discovered to be effective against 66 human proteins or host factors out of 332 SARS-CoV-2 - human PPIs [164]. As reported, human Angiotensin-converting enzyme 2 (hACE2) is abundantly expressed in lung alveolar epithelial cells, endothelial cells that originate from small and big arteries, and other tissues, as well as in the heart, kidney, testis, and gastrointestinal system, recent studies have demonstrated that the SARS-CoV-2 S-protein can bind the hACE2, [165, 166] explaining the symptoms linked to the SARS-CoV-2 infection.

Molecular docking method allows us to characterize how tiny molecules behave in the binding site of target proteins and to better understand basic biological processes by simulating the atomic-level interaction of a chemical molecule with a protein [167]. Prediction of the ligand structure as well as its placement and orientation inside these sites and evaluation of the binding affinity are the two fundamental processes in the docking process. As it can show how a ligand interacts with its biological targets, *in-silico* docking approach is now widely used in drug development studies [168, 169]. Using the main protease of different drugs, molecular docking plays a significant role in drug repurposing studies. [170] proposed a molecular docking simulation using Autodock4 where zafirlukast and simeprevir came out to be the most promising drugs to inhibit spike ACE2 interaction with host protein for combating COVID-19. Virtual screening was performed by [171] on the drug data retrieved from ChEMBL and drug bank on the main protease of nCoV. A generic framework for drug repurposing has been proposed by [172] using drug and disease data collected from public data repositories and different classification models to classify them. The identified drugs have been validated by molecular docking strategies to identify the binding affinities of drug molecules against nCoV host targets. Molecular docking and virtual screening strategies have been applied using a set of FDA-approved drugs on the crystal structure of COVID-19 main protease 6LU7 to identify the repurposed drug [173]. Tipranavir, Lopinavir–Ritonavir, and Raltegravir have been confirmed to have the best interaction position with the main protease of nCoV as proposed by [174].

Most of the research studies proposed different repurposed drugs from the list of FDA-approved drugs that are associated with SARS-CoV-2 proteins. There may be still room for research which might include the repurposing of drugs from symptom-based analysis of COVID-19. The above study motivates the design of an *in-silico* Human-nCoV PPIN model from the spreadability index of SARS-CoV protein. PPIN and symptom-based analysis can be two major focuses in developing repurposed drugs. Different molecular docking strategies can be employed over the available crystal structure of COVID-19 virus to validate the identified drugs.

Analysis of Large-Scale Human-Coronavirus Family Interactome

The family of enclosed, single-stranded RNA viruses known as coronaviruses is taxonomically related to the coronaviridae family [175]. It is not just humans that are impacted by this single-stranded RNA virus; animals and birds are as well. Humans often experience symptoms of the flu or a common cold as a result of coronaviruses, which leads to acute respiratory infections. However, because coronaviruses are contagious, diseases like SARS-CoV and MERS-CoV have the potential to spread globally. Both of these two coronaviruses belong to the family Coronaviridae's genus Betacoronavirus [158]. SARS-CoV, which was initiated in 2003 in South China, had a fatality rate of 15%. Whereas MERS-CoV, which originated in 2012 in the Middle East, had a high fatality rate of 34%. According to Lu et al. (2020), the coronavirus known as SARS-CoV-2 is a member of the same genus as MERS-CoV and SARS-CoV, which is known as Betacoronavirus. There are both structural and non-structural proteins in it. Structural proteins, such as the envelope, membrane, nucleocapsid, and spike proteins, can be identified as prominent examples. Despite the fact that SARS-CoV-2 has just recently been discovered, there is a severe lack of knowledge and data needed to develop immunity against SARS-CoV-2 [176]. Several experimental genomic investigations have shown that SARS-CoV-2 and SARS-CoV are genetically very similar [176–178]. A PPIN is at the center of the proposed method for identifying SARS-CoV spreading nodes. When it comes to identifying protein functions and locating their central/essential spreading nodes, the PPIN is an extremely useful module [152, 154, 155, 179–181].

The discovery of SARS-CoV in caged animals from wild live markets in mainland China raised the possibility that these creatures are the source of the SARS pandemic [182]. Later research hypothesized that the civet may have merely been a SARS-CoV amplifying host and habitat for significant genetic differences allowing effective animal-to-human and human-to-human transfers [183–185]. Lineage C of MERS-CoV, Bat coronaviruses (BatCoVs) HKU4 and HKU5 are two more significant members of the betaCoV family [186]. These viruses continue to circulate in bats [187] despite being originally discovered as genomes in lesser bamboo bats as well as Japanese pipistrelles in 2006, respectively [188]. Human illnesses can be brought on through both alphaCoVs and betaCoVs coronaviruses [175]. It has been challenging to develop models of MERS-CoV disease since MERS-CoV is inherently resistant to small animals, which are often employed to study viral pathogenesis [189]. Different non-human model has been developed for MERS-CoV with different species for preclinical therapeutics study [190–192].

Not only SARS-CoV2, SARS-CoV, and MERS-CoV interacts with human to spread diseases. But other coronavirus species also interact with human proteins. Till date, no study has been done on the entire coronavirus interactome. The above study motivates to design of an *in-silico* model that can identify level-1 spreader proteins of different coronavirus species that interact with human proteins and can efficiently identify suitable repurposed drugs.

Prediction of PTM Sites in Protein Sequences

Among different PTMs, the covalent post-translational alteration of the cysteine thiol side chain by palmitic acid is known as S-palmitoylation. S-palmitoylation is involved in various human disorders and is crucial for a number of biological activities. Protein location, transport, and function are heavily regulated by protein S-palmitoylation [193]. A picture of protein S-palmitoylation as a common yet distinct chemical switch that enables the expansion of protein activities and subcellular localization in minutes to hours emerges from a number of recent experimental findings. Particularly numerous proteins controlled by S-palmitoylation are found in neural tissue [194]. According to a theory, PTMs obtained throughout animal evolution offer proteins new properties that are essential for the development of the nervous system's complexity and capacities [195]. Since learning and memory depend on synaptic plasticity, it plays a crucial function in the brain. Different synaptic proteins express changes in synaptic strength, which are then translated into variations in the structural and functional properties of neurons [196, 197]. Since sex-dependent variations in the brain are common and may be seen even at the level of individual synaptic connections, gender should be regarded as a significant biological variable in neuroscience. Sex variations in neuronal function may be caused by differences in synapse molecular organization [198], signaling pathways [199,200], and plasticity [201,202], as well as differences in memory and learning [203], emotional reactions [204], fear, and anxiety. These differences may also be explained by differences in synaptic plasticity [201, 202]. The occurrence of discrimination based on gender in neuropsychiatric disease [205] is a clear manifestation of the substantial clinical effects of biological sex. Major depressive illness is more likely to affect women than men [206], but autistic spectrum disorder [207] is more likely to affect men. S-palmitoylation, one of several PTMs, stands out as a key process underlying synaptic integrity, and its failure has a connection to neuropsychiatric disease [208]. Due to its S-palmitoylation-mediating activities primarily through sex steroid receptors, one of the palmitoylating enzymes, palmitoyl acyltransferase DHHC7, has a special interest in relation to sex-dependent neuroplasticity [209]. Additionally, DHHC7 modifies other synaptic proteins [26–28], controlling their cell membrane attachment, organizing, and function, all of which are crucial for the efficient operation of synaptic connections [210–212].

Considering the *in-silico* techniques for identifying PTM sites for S-palmitoylation, the CSS-palm [213] and NBA-palm [214] techniques were touted as the forerunners in this sector with reasonable performance. Yang et al. [215] then suggested a regularised biobasis neural network (ANN) technique, which marginally improved performance. Using an encoding system composed of k-spaced amino acid pairs, Wang et al. created the CKSAAP-Palm approach, which performed better in terms of recognition accuracy than earlier techniques [216]. The closest neighbor (NN) technique was then used by Hu et al. [217] to develop the IFS-Palm approach using protein structure and sequence information. Shi et al. [218] additionally provided a predictor via the SVMs approach known as WAP-Palm. However, the identification results were poorer than those of the CKSAAP-Palm approach. Fu et al. [219] suggested a novel predictor for the random forest (RF) approach, and the RF algorithm's out-of-bag assessment showed decent performance in predicting with a reasonably good MCC score. Kumari et al. [220] created a PalmPred model that also used SVMs and sequence profiling data. After that, Pejaver et al. [221] created ModPred, a single PTM sites predictor that includes 23 different PTMs, including palmitoylation.

The need of the hour is to propose a model that can predict the S-palmitoylation sites from data generated using *in-vitro* method. Despite some bottlenecks, deep learning models may be useful in the prediction of sites for S-Palmitoylation. The development of a web server with an extension to other PTM types may be beneficial.

1.4 Objectives of Current Research

In order to address the limitations of current methods in section 1.4, the thesis aims to do the following.

Analysis of Large-Scale Human Protein Sequences Using Big-Data Framework

Clustering is a powerful unsupervised learning technique that may be used to make predictions and eliminate outliers by grouping data into related sets by reducing the redundancies of large biological sequences [222]. For larger values of n, the dimensionality of the input space increases due to the usage of the *n-gram* feature representation, which is often used in sequence clustering and classification. Processing all of the data with the available computer resources has become a difficult task as the amount of data grows daily. The time has come to approach it as a big data issue, requiring cutting-edge technology to store and handle the data in a distributed manner [222]. Big data is a concept that allows to process of different large-scale data using different frameworks (such as Apache Hadoop, and Apache Spark). Thus, an Apache Spark-based DBSCAN algorithm is being proposed in this section of the thesis that efficiently employs the higher dimensional input space and large-scale human sequence data using parallel computing resources [223]. The proposed method was implemented on an

Apache Spark cluster that eventually helps to parallelize the resource and computation efficiently.

Large-Scale Interaction of Human Proteome at Fuzzy Semantic Space

Key insights into an organism's cellular and molecular functions, signaling pathways, and underlying disease causes may be gained by studying its large-scale PPIN. For every given organism, the total amount of all possible positive and negative PPIs far exceeds the number of known interactions. For humans, the total number of interactions included in all available PPI datasets is only around 3.1%. Furthermore, emerging evidence suggests that protein binding affinities may be used to effectively identify protein complexes, conduct illness association research, rebuild signaling networks, etc. Keeping this in mind, the work proposed an Apache Spark-based parallel architecture for designing of a fuzzy semantic score of binding affinity using the GO network to assess the binding affinity between any two proteins at an organism level. In the proposed work, human PPIN of $\sim$ 180 million potential interactions resulting from 18,994 reviewed proteins for which GO annotations are available is being used. This continues in the construction of a fuzzy semantic network at the proteome level for the extraction of meaningful biological insights.

Computational Modeling of Host-Pathogen PPIN

SARS-CoV-2, a novel coronavirus, interacts with host proteins to reproduce the genome of the host cell. Because of this, understanding how viruses spread illness and finding possible COVID-19 drugs may both benefit from the identification of viral and host PPIs. Host-pathogen PPINs are important for comprehending the method of infection transmission, which is necessary for creating novel, more potent therapies, including logical drug design. The PPIs between the pathogen and host cause infection and illness to progress. The pathogen actively contributes to the spread of infection. Pathogen and host PPIN permit pathogenic microorganisms to utilize host capabilities by manipulating the host mechanisms to abscond from the host's immune responses [56, 137, 138]. The goal is to identify target proteins using PPIN analysis of the pathogen and host. In general, it is discovered that topologically significant proteins with a greater degree of interactions become major therapeutic targets. However, due of some biological pathway importance, proteins with fewer interactions or with insignificant topologies could be implicated in the process of infection. As, clinically validated Human-nCoV PPI data is limited in the current literature, thus, this part of the dissertation proposed an *in-silico* model human-nCoV PPIN along with the binding affinity between protein pairs using semantic similarity of the available GO graph information and Fuzzy affinity thresholding is done to detect High-Quality nCoV-Human

PPIN. The chosen human proteins are regarded as nCoV level-1 spreader nodes in humans. The Susceptible-Infected-Susceptible (SIS) model validates the spreadability index for locating level-2 spreader nodes in the human-nCoV PPIN. [158]. The proteins in the network have been subsequently validated with respect to FDA-approved drugs for the treatment of COVID-19.

Drug Repurposing for COVID-19 using a Novel in-silico Method

An efficient medication or vaccination is urgently needed given the steadily rising COVID-19 death rate. Several clinical trials are being conducted to evaluate several drugs and vaccines for the treatment of COVID-19, including Remdesivir, Azithromycin, Favirapir, Ritonavir, and Darunavir. Some of them have already been approved, but others have shown positive outcomes and are currently being evaluated. Drug repurposing research has accelerated due to the prolonged approval process for novel compounds. Analysis of Host-Pathogen PPIN may be used to efficiently gain a good knowledge of disease transmission pathways [224]. Disease progression is aided by the pathogen since it has the capacity to mutate and change. Through the connecting edge of the host and pathogen contact, infection of the pathogen spreads. In order to find new drugs, it is crucial to investigate target proteins and their interactions in networks of host-pathogen PPIs [56]. Though there are different *in-vitro* methods based human-nCoV PPIN [164] there is a need to develop an *in-silico* human-nCoV model and to identify different repurposed drugs based on the host-pathogen PPIN model. Infection of the pathogen gets broadcasted through the connecting edge of interaction between the host and the pathogen. Keeping this in mind, the section of the thesis proposes a two-way analysis which includes, human-nCoV network analysis and COVID-19 symptom [225] based analysis (including *"loss of smell"*) using the *in-silico* model which has been developed to identify potential spreader proteins in a human-nCoV interaction network in the work of Saha *et al.* [158, 226] and which was validated using proteins which are the targets of potential FDA-approved drugs [227] for COVID-19 treatment to detect the potential candidates in the list of FDA-approved drugs for COVID-19. Molecular docking has also been performed on potential FDA-approved drugs with the available major COVID-19 crystal structures.

Assessment of GO-based Protein Interaction Affinities in the Large-Scale Human-Coronavirus Family Interactome

A new coronavirus called SARS-CoV2 replicates by interacting with the host proteins. Therefore, identifying viral and host PPIs might aid researchers in better understanding the manner in which viruses spread illness and help them discover potential COVID-19 medicines. An organism's large-scale PPI network offers useful clues for comprehending cellular and molecular functions, and signaling pathways may illuminate the illness process, among many other advantages. Several similar studies based on GO information have been done on host-pathogen PPINs. There is a real urge to identify a complete PPIN for humans and the coronavirus family. Keeping this in mind, this section of the thesis proposes a novel computational model to develop the host-pathogen PPIN between humans and all other different variants of the coronavirus family. The proposed computational model will be able to assess the interaction affinity between human-coronavirus interactome using GO graph information. The resultant host-pathogen PPIN has been extended further toward the drug-repurposing study by analyzing the FDA-approved COVID-19 drugs.

Prediction of S-Palmitoylation PTM Sites in Protein Sequences

The covalent post-translational alteration of the cysteine thiol side chain by palmitic acid is known as S-palmitoylation. S-palmitoylation is involved in various human disorders and is essential in a number of biological activities. Thus, identifying the precise locations of this modification is essential to comprehending the functional ramifications in physiology and illness. The discovery of target proteins for S-palmitoylation has been accomplished using a variety of techniques. Less research has been done on site-specific S-palmitoylation detection. Using information from massive proteome datasets, tools for predicting specific S-palmitoylation sites in various biological complexes have been created recently [214, 228]. Different machine-learning-based algorithms [121, 228–230] have predicted S-palmitoylation sites with a decent performance score. With the growing number of publicly available large-scale proteome databases of the brain and somatic tissues, there is a need for the development of reliable and accurate computational tools to process them. This part of the thesis proposes a RF classifier-based consensus strategy, which can predict the palmitoylated cysteine sites on synaptic proteins of the male/female mouse dataset generated by the mass spectrometry-based method PANIMoni [231] and a detailed ZDHHC subtype-specific and sex-mouse S-palmitoylome [232, 233] using AAIndex feature database along with position-specific amino acid (AA) propensity information.

1.5 Organization of the thesis

In this dissertation, first, an unsupervised machine learning approach has been employed on a large set of human protein sequences using high-throughput parallel architecture to show the benefit of using big data frameworks on large-scale biological data.

As discussed, protein interaction networks are a valuable clue for finding valuable biological insights. However, it has been identified from different sources that for human proteins, around 96% protein interactions have not yet been explored. Moreover, there is an urgent need to identify pure protein-protein negative interactions (PPNI) for better *in-silico* predictions. Thus, a large-scale organism-specific protein interactome, with respective binding affinity for each pair of proteins, has been developed using fuzzy semantic space.

Understanding the route of virus transmission is crucial for developing innovative, more effective treatments, including rational medication design, which depends on host-pathogen PPINs. Next, the proposed fuzzy semantic network has been used to detect high-quality human-nCoV PPIN and identities for level-1 and level-2 nCoV spreader nodes in humans. The network has been validated with respect to FDA-approved drugs for the treatment of COVID-19.

As vaccines or any specific drug for COVID-19 take a substantial amount of time for approval, to eradicate the disease propagation of COVID-19, efficient drug repurposing studies have gained considerable momentum. Thus, next, a *in-silico* model has been developed to identify the spreader proteins in human-nCoV PPIN which was then validated *w.r.t* different FDA-approved drugs for COVID-19 treatment based on two-way disease analysis.

Next, an *in-silico* model based on the fuzzy semantic network has been proposed to assess the interaction affinity between human-coronavirus interactome for identifying the spreader nodes which eventually gets validated *w.r.t* possible FDA-approved drugs.

PTMs are the changes that occur by different biomolecules in the amino-acid side chain. It has been studied that the proteins which undergo PTMs are more susceptible to interact with other proteins. Thus, an *in-silico* model has been proposed to predict the S-palmitoylation PTM sites in synaptic proteins for male/female mouse. The dataset has been provided by The Nencki Institute of Experimental Biology, Poland. The dataset is private and not available for the private domain.

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

i) **Chapter 2** describes the analysis of large-scale human protein sequences using big data framework

ii) **Chapter 3** elaborates large-scale interaction affinity for human proteome at fuzzy semantic space

iii) **Chapter 4** discusses computational modeling of Host-Pathogen PPIN.

iv) In **Chapter 5** discusses, drug repurposing techniques for COVID-19 using a novel *in-silico* method.

v) **Chapter 6** highlights the assessment of GO-based protein interaction affinities in the large-scale human-coronavirus family interactome,

vi) **Chapter 7** presents S-Palmitoylation Sites Prediction for Synaptic Proteins.

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

Chapter 2

Analysis of Large-Scale Human Protein Sequences Using Big Data Framework

2.1 Background

Proteins work together to carry out a variety of crucial cellular and biological functions. Proteomic information has been growing exponentially in recent years. There are over three billion interactions between 24.6 million proteins across 5,090 different species in the STRING database. It is challenging to manage such a vast, dynamic, heterogeneous network using current computational approaches. After the Human Genome Project was finished, there was a massive increase in genetic sequence data. Proteomic data, which includes measurements of protein abundance across a variety of biological contexts, has grown at an exponential rate to keep pace with genomic data.

The development of modern high-throughput sequencing techniques has resulted in an exponential growth in meta-genomic sequence accumulation that could greatly enhance large-scale functional annotation. Processing of these large and redundant sequences has become a major challenge for researchers. Clustering by similarity is one of the major steps to reduce the redundancy of these enormous resources and analysis of such large biological sequences. The *n–gram* feature representation, generally used in sequence clustering and classification, results in high dimensional input spaces, for larger values of n. However, it becomes intractable to cluster such large-scale sequences by current algorithms due to a large number of dimensions. An efficiently designed, clustering approach can easily scale to handle large-scale sequences by utilizing the power of parallel computing with high-performance computing systems.

Processing these data with limited processing resources is becoming more and more of a challenge as the volume of data continues to grow exponentially. Today, this issue must be tackled as a Big Data one, necessitating sophisticated tools for storing and processing data in a decentralized manner. Apache Hadoop and Apache Spark are the answer to this issue; they employ strategies built around the use of commodity hardware to conduct processes in parallel. Big Data is classified under 5V's. These are Volume, Velocity, Variety, Value, and Veracity. All these classifications are briefly

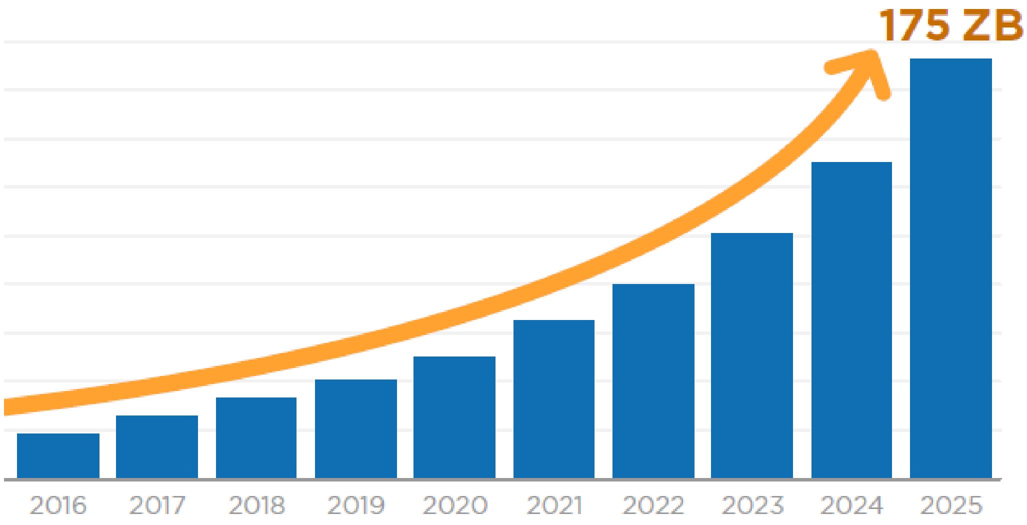


Figure 2.1: Data Growth Forecasted for Big Data [234]

described below:

- **Volume:** It refers to the vast amount of data generated every second.
- **Velocity:**It refers to the speed at which data is generated and the speed at which the data moves around.
- **Variety:** It refers to the different types of data.
- **Value:** It refers to the messiness or trustworthiness of the data. With many forms of Big Data, quality, and accuracy are less controllable, but Big Data and analytics technology now allow us to work with this type of data.
- **Veracity:** Veracity generally refers to the uncertainty of the data, i.e whether the obtained data is correct or consistent. Out of the huge amount of data that is generated in almost every process, only the data that is correct and consistent can be used for further analysis.

Figure 2.1 depicts a forecast of data growth for a range of 10 years. However, just 2% of the 64.1 ZB of data produced in 2020 was maintained or stored until 2021, according to IDC. IDC predicts a 23% CAGR for new data generation from 2020–2025, leading to an estimated 175ZB of data creation in that time frame [234]. Therefore, storing and analyzing the data becomes the primary issue. The time needed to process such massive data grows proportionally with data size. Big Data refers to information volumes and types that exceed traditional data storage and processing methods.

In computational biology, many problems such as protein homology detection, and functional annotation, can be formulated as sequence clustering tasks where amino acid sequence compositions of proteins are considered as the key information. Protein sequences are any combination of 20 constituent elements by maintaining intrinsic dependencies between these amino acids. In a protein sequence, the dependencies between neighboring entities can be explored by generating all possible overlapping sub-patterns of a certain length n, called *n-grams* [28]. *N-gram* based dependencies in data are explored to improve the richness of the representation. However, in *n-gram* representation with higher values of n, the dimension of the input space increases exponentially (20^n). It became intractable to apply a data clustering algorithm with higher dimensional input space while the dataset itself is large. So efficient algorithms need to be designed that can cope with higher dimensional large-scale sequence data.

Plenty of methods have been developed in the last decades for sequence clustering [28–31]. Though the sequence clustering from large sequence data remains a serious challenge. Based on the structural organization of clustering, these approaches can be categorized into two groups: hierarchical clustering (HC) and greedy heuristic flat clustering (GHFC). In HC-based approaches [123], sequences are organized in a hierarchical tree and produce different levels of OTUs at various similarity levels that can explore more biological significance. However, HC-based methods have major drawbacks in that it require pairwise distance computation and requires to store the distance matrix. Cd-hit [30] and UCLUST [29] are widely used GHFC approaches that employ greedy flat clustering to reduce computational complexity. Though GHFCs are faster than HC-based approaches, GHFC produce lower cluster quality than HC-based approaches. It is highly desired that clustering methods can be easily scaled to handle massive sequence data using parallel computation with high-performing computing architecture. However, it is inherently difficult to redesign the existing approaches in parallel mode as each iteration distance computation totally depends on its earlier step or each division/merging relies on the previous operations. In this scenario, parallel density-based clustering could be an efficient alternative in large-scale sequence data clustering with higher dimensional input space.

Big Data, as the name indicates, is a concept that promises to handle massive data sets in a short amount of time by making use of several processors located on separate nodes. In order to handle this massive amount of data, it must be processed in parallel across several nodes. The Apache Software Foundation developed an open-source framework called Hadoop [235] to hasten to compute. Hadoop allows users to store and analyze a massive dataset on commodity computers with extremely high bandwidth between clusters. Hadoop's YARN and MapReduce components facilitate cluster-wide resource management and parallel computation, respectively.

2.1.1 Hadoop and MapReduce

In order to analyze the ever-growing dataset, it is necessary to break down the issues into smaller, potentially independent subproblems for efficient parallel execution. The MapReduce component of Hadoop, an open-source framework provided by the Apache Software Foundation for storing and processing massive datasets with clusters of commodity hardware, distributes the computation among the clusters and merges the output upon completion [236, 237].

Hadoop Distributed File System (HDFS) is a distributed file system provided by Hadoop. Its default block size is 64 MB, however, this can be increased to 128 MB if necessary. HDFS was developed to accommodate large datasets that are stored in a cluster on inexpensive, general-purpose servers and accessed in a streaming fashion. NameNode, Secondary NameNode, TaskTracker, DataNode, and JobTracker all emerged in the HDFS cluster. The first three are master services, while the latter two are slave services, and each master and slave can communicate with each other. The client must keep the data in a cluster of several systems and share the data among those systems in order to store and analyse the massive quantity of data in a reasonable length of time. The workflow of the Map-Reduce paradigm is given in Figure 2.2.

Hadoop technology requires the input data to be partitioned into many blocks, or input splits before it can be processed. All input partitions are either 64MB or 128MB in size or smaller. There is one mapper and one record reader for each input split. Since there are several forks in the input stream, multiple mappers will be executed. Line by line, the record reader pulls data from its associated input split, parses it into key/value pairs, and sends them on to the appropriate mapper and mapper starts. The reducer takes all the key, value pairs as output from the mappers and shuffles and sorts them to limit the number of duplicate keys. The record writer receives output from the reducer as a last resort.

2.1.2 Apache Spark Architecture

Storing and processing a large amount of data requires splitting the problem into some dependent and independent sets for parallel execution. Hadoop with a cluster of commodity hardware can store and process huge datasets in a reasonable time [222]. However, Hadoop lacks some computational issues such as in-memory computation, computing real-time data efficiently, etc. The intermediate data write-back operations to HDFS results in low latency for the Map-Reduce scheme in Hadoop.

On the other hand, Spark by Apache software foundation makes computing faster

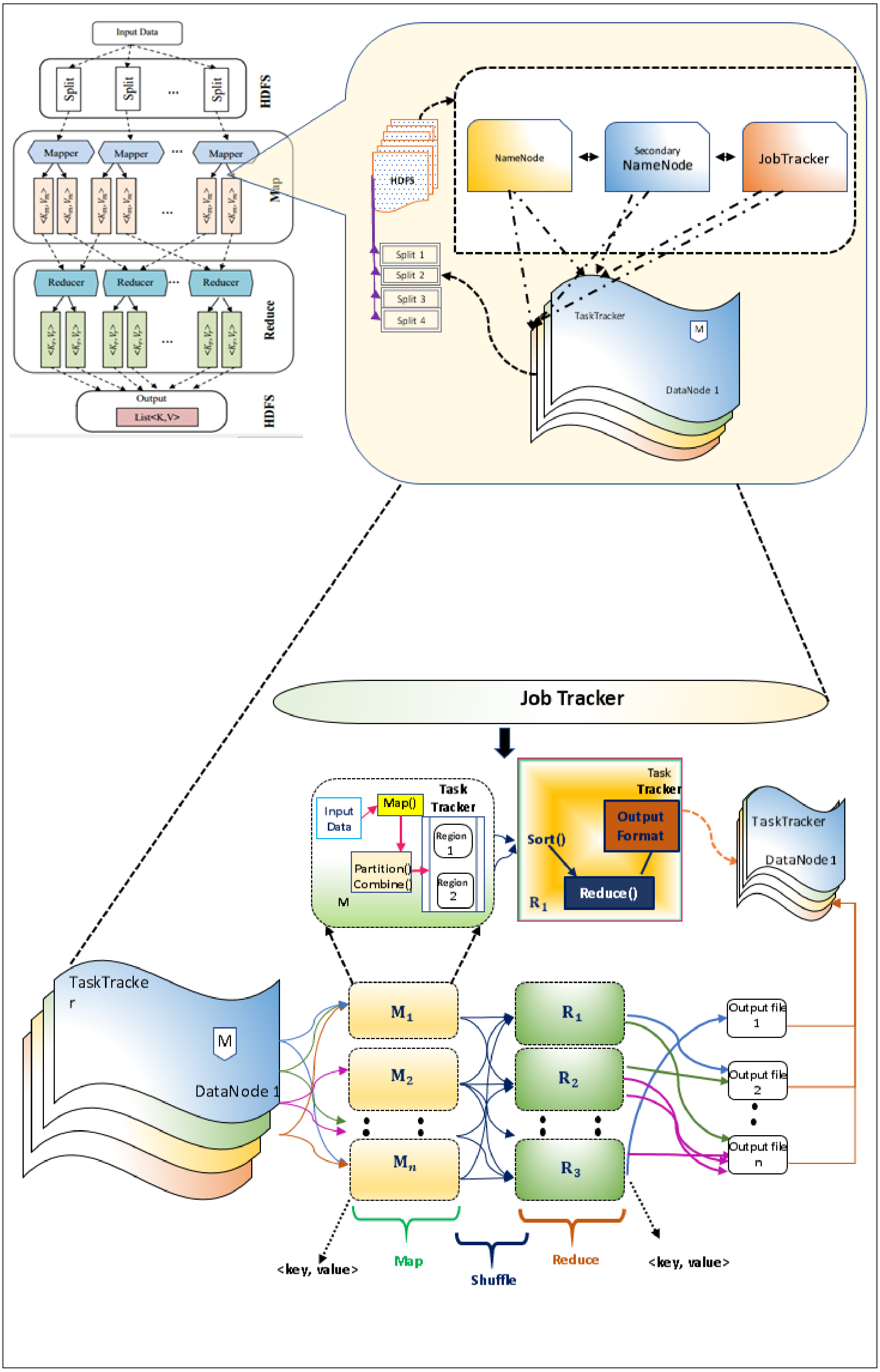


Figure 2.2: The figure shows Hadoop architecture and MapReduce working procedure

by using in-memory computation [238] with a new memory abstraction called Resilient Distributed Dataset (RDD). RDDs are partitioned and distributed across multiple nodes and can rebuild if a partition is lost. Apache Spark architecture is consisting of master and worker nodes. The main module of the master node is the driver program that controls the application by creating Apache Spark context. Apache Spark context works with the cluster manager to manage multiple jobs distributed over the worker nodes. The worker nodes execute the tasks on the partitioned RDD and return the result back to the Apache Spark context. The block diagram of Apache Spark architecture is given in Figure 2.3. A brief detail of Apache Spark architecture are discussed below:

- Apache Spark architecture consists of a master node and several worker nodes. The driver program is the central module of the master node that really does the work of driving the apps.

- Apache Spark context is initially created in the driver program. Spark context is the starting point for all Apache Spark operations. It's analogous to linking to a database. Spark context is applied to everything we do on Apache Spark.

- Apache Spark context then coordinates with the cluster manager to handle the jobs. A job is any task that is executed within a Spark context. This job's execution throughout the cluster is handled by the driver program and Spark context. Jobs are broken down into smaller tasks, which are then assigned to individual worker nodes.

- When a Spark Context generates a RDD, it can be cached across several nodes. As a result, the RDD is split up and sent to different worker nodes.

- The jobs run on the worker node's partitioned RDD and send their output to the Spark Context.

- By adding more worker nodes, we can do the same work much more quickly using parallel processing. Additionally, the more RAM we have, the greater cache we'll have on this worker node. This means that we can store several RDDs in RAM, making in-memory operations far quicker than their disk-based counterparts. Because of this, Spark is much quicker than Hadoop/MapReduce.

Spark stores and processes huge data using a specific type of dataset called RDD. The key characteristics of RDD are discussed below:

- Apache Spark foundation is the RDD. In the Spark program, RDDs may be made, manipulated, analyzed, and stored.

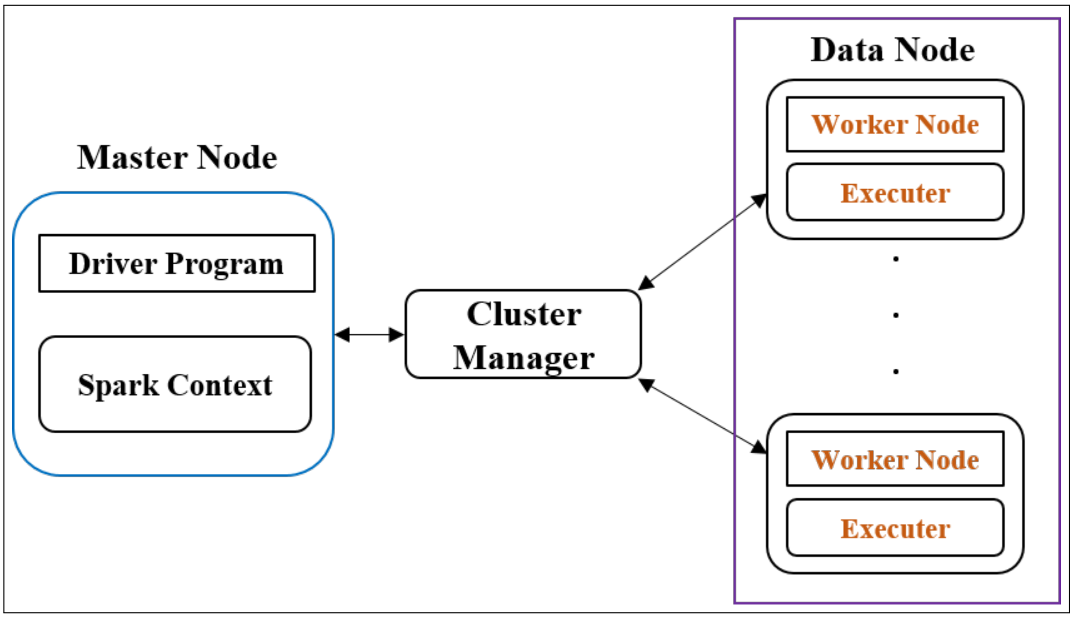


Figure 2.3: Apache Spark Master-node Data-node architecture.

- Any data structure, such as strings, lines, rows, objects, or collections, can be stored in an RDD.

- Multiple nodes might be used to spread the datasets in parts. Spark is responsible for all data partitioning and dissemination.

- No changes may be made to RDDs. The name ”Resilient” was coined for this reason. They are fixed and unmovable. An RDD's context is fixed once it has been constructed.

- Caching and persisting RDDs is possible. Typically, RDDs are kept on discs for storage. The combination of RAM and discs is another viable option for processing. When we require RDD for operations, we obtain it, modify it, and then dispose of it.

2.2 Methodology for Clustering Protein Sequences

In this work, a Spark-based DBSCAN algorithm has been proposed, that efficiently employs on the higher dimensional input space and large-scale human sequence data using parallel computing resources. The proposed method was implemented on an Apache Spark cluster that eventually helped to parallelize the resource and computation efficiently. Experimental result shows efficient speed-up in computation compared with different experimental architectural setups. In addition, the quality of the clusters is assessed with biological significance such as Domain Correspondence Score.

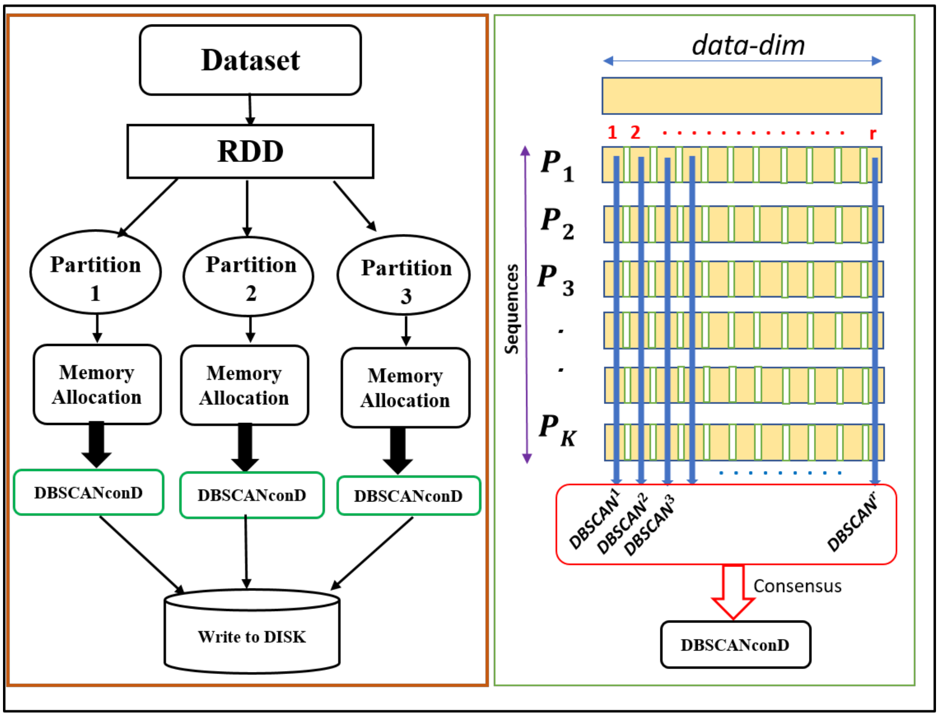


Figure 2.4: Overview of proposed parallel scheme for large scale sequence clustering using DBSCAN(left). DBSCANconD: consensus-based parallel DBSCAN clustering on fixed length sub-dimensional partitioned data (right).

A new computational method for parallel clustering analysis of large-scale human sequence data with higher dimensional input space has been developed. High dimensional input features are generated from each protein sequence by exploring the n-grams feature (at n=3, the total number of features is 20^3=8000). The basic idea is to first partition the data into equal-sized small partitions and process DBSCAN on each partition in parallel mode and at the final step individual results are merged for the final clustering solution. The DBSCAN algorithm within each partition is further paralleled over the higher dimensional input spaces where each data is processed as fixed-length multiple sub-dimensions. The clustering results from each sub-dimension-based processing are merged using a consensus mechanism. Finally, merged results from each partition are considered as the final cluster. The flowchart of the proposed method is presented in Figure 2.4. All parallel computations are processed on Apache Spark architecture.

2.2.1 Spark-based DBSCAN Algorithm

One of the main focuses of unsupervised learning or exploratory data analysis is to group the data based on the characteristics of the data. Density Based Spatial Clustering (DBSCAN), is a well-known data clustering algorithm that is commonly used in data mining and machine learning. The architecture of the DBSCAN algo-

```
Algorithm 2.1: DataPartition
  Input   : data $D$, block_size
  Output: $< key, list(value) >$
  ------------------------------------------------
  Process:
    1. Transform the data D into RDD.
    2. Convert the RDD to an array
    3. noPart ← size(D) / block_size
    4. Parallelize RDD array over noPart partition containing the array and map
       each partition with index as a new_RDD.
    5. Return new_RDD as $< key, value >$ pair
```

rithm is based on two basic parameters: radius(*eps*) and number of minimum points (*minPts*) [239] The radius (*eps*) defines the neighborhood around a data point whereas *minPts* defines a minimum number of neighbors within *'eps'* radius. The neighbor points N ∈ (p) for any point p (∈ D) is defined as:

$$N_{eps}(p) = n \in D | dist(n, p) \leq eps. \tag{2.1}$$

Based on these parameters, data points are classified into 3 categories: core points, border points and noise/outliers. A point is classified as a core point if it has more than *minPts* points within *eps* whereas for a border point, it has fewer than *minPts* within *E*ps, but is in the neighborhood of a core point. An *outlier/noise* point is one that is neither a core point nor a border point. For any two points x and y are said to be connected if x is dense and the distance between x and y is less than *eps*. The distance of two data points x and y is defined as Euclidean distance.

To implement the DBSCAN algorithm in parallel, a method has been proposed by distributing the dataset efficiently can minimize the computation time for large-scale datasets. In this work, two-level parallel processing for large-scale data clustering has been proposed. The first one is to split the dataset over the cluster by transforming it into an RDD and then executing the DBSCAN algorithm to each split in parallel (*see* Algorithm 2.1). Within each partition, DBSCAN is further paralleled over the higher dimensional input spaces where each data is processed as fixed-length multiple sub-dimensions (*see* Algorithm 2.2). Finally, the resultant clusters are collected from different partitions and are written into disks as described in Algorithm 2.3

Algorithm 2.2: *Local_DBSCAN*

```
Input  : D $< key, list(value) >$, $\epsilon$, $MinPts$
Output: $< key, list(value) >$

Process:
  1. The partition generates a list (values).
  2. An arbitrary point $p$ is selected from the partitioned data for searching the
     neighborhood points. If N_$\epsilon \geq MinPts$ then we assign point p as the core point
     else the point is marked as noise.
  3. If $p$ is marked as a core point, then we then make cluster $c$ with point $p$ and all
     the other points belonging to cluster $c$. Each cluster is assigned a cluster
     number.
  4. Repeat steps 2 and 3 until all the points of the datasets are assigned to a
     cluster or are marked as noise.
  5. Returns the data points with their assigned cluster number.
```

Algorithm 2.3: *Clustering*

```
Input  : $q$, $\epsilon$, $MinPts$
Output: Cluster

Process:
  1. D ← Load data of dimension (k,n)
  2. PartitionRDD ← DataPartition(D, block_size)
  3. For each partRDD in PartitionRDD
       • r ← n/q
       • split data dimension over r number of sub-dimensions of length $q$ as
         partRDD_i
       • process Local_DBSCAN in parallel over all sub-dimension set as $CL_i$ ←
         Local_DBSCAN(partRDD_i, $\epsilon$, $MinPts$)
       • CL ← Consensus result overall $CL_i$
  4. Collect results for all partitions of PartitionRDD and merge them as the final
     results.
  5. Write to disk
```

2.2.2 Cluster Validation

The clustering results are analyzed and validated in the context of biological significance. We have incorporated Domain Correspondence Score to quantify the clustering results as described in [28]. Domain Correspondence Score is the measure of purity for any cluster in terms of functional domain annotation for a group of sequences. Domain Correspondence Score is a more effective metric for both single and multiple-domain proteins that belongs to a particular cluster. For detailed Domain Correspondence Score computation please see [28]. Higher Domain Correspondence Score indicates highly conserved domain annotation within a cluster and lower represents the reverse or even corrupted. A cluster with exactly identical domain annotation (single or multiple) for every point (protein) ensures Domain Correspondence Score as 1 which indicates a high-quality pure cluster.

2.3 Experimental Result

In this work, human protein sequence data have been chosen for end-to-end experiments. A total of 20431 reviewed human protein sequences are collected from UniProtKB/Swiss-Prot [240] as fasta formats and then converted into the desired higher n-gram (here n$\leq$3) feature representations. The experimental setup is carried out mostly on bigram and trigram features where the input spaces range from 400(20^2) to 8000 (20^3). The experiment has been devised with two objectives,

1. speed-up efficiency
2. quality of cluster

First, we concentrate on the performance benefit of clustering on large biological data by leveraging the power of parallel computation on the Apache Spark framework. Then, the quality of the clustering result is analyzed and validated with respect to biological relevance.

2.3.1 Speed-Up Efficiency

To compare the performance of the proposed method on different parameters such as data size, data points, and data dimension, three different experimental environments (EE) have been set up. They are 1 Master and 2 Slave multi-node based physical Spark cluster referred to as *EE1*, Spark community edition distribution over cloud hosted by databricks as *EE2* and a single standalone node as *EE3*. The performance speed-up has been reported in Table 2.1. The parallel execution of the DBSCAN clustering algorithm on higher dimension (bigram or trigram) protein sequence data with *EE1* setup is faster than the other two (*EE2*, *EE3*).

In all three *EE*-setups, the algorithm has experimented with different data-size and dimensions. The results depict that increasing the size of the dataset, the speed-up ratio of *EE1* compared to *EE2* and *EE3* also increases. As the size of the datasets ex-

Table 2.1: Performance speed up of different execution setup.

Data Size	Data Points	Data Dimension	EE1#	EE2#	EE3#	Speedup (EE2/EE1)	Speedup (EE3/EE1)
16	20413	400	0.3	0.5	0.4	1.67	1.33
326	20413	8000	6.9	32	16	4.63	2.32
653	20413x2	8000	20	71	48	3.55	2.4
981	20413x3	8000	35	$	$	-	-

#- units-time in minutes, $-unable to process

Table 2.2: Performance speed up of different execution setup.

Clustering methods	Total Clusters	Non-Singleton	%non-Singleton at (Domain Correspondence Score=1)	corrupt	Seq/clust	%Redundancy Reduction
BiGram_con	10452	421	83.1	12	1.954	48.7
BiGram_all	20369	49	100	0	1.003	0.21
TriGram_con	14355	362	84.2	2	1.423	31.9
TriGram_all	20126	54	100	0	1.006	0.14
Uniref90 [31]	19537	293	81.0	2	1.046	0.42
Cd-hit90 [30]	19544	108	84.1	1	1.045	0.425

ceeds the block size of the Spark framework (i.e 128MB), the parallel execution becomes faster compared to sequential and pseudo cluster execution. To analyze the dataset, first, the cluster consensus by partitioning the data into equal-length sub-dimensions has been obtained. From the consensus result, singleton clusters are extracted as noise in all clustering and the remaining are referred to as non-singleton clusters. The results were obtained from the complete dimension-based approach and sub-dimension-based consensus approach (*see* Table 2.2) in both bigram and trigram features.

2.3.2 Quality of Cluster

In complete dimension base approach in both features produces a maximum number of clusters 20369 and 20126 which suggests that the redundancy removal power is very low 0.21% and 0.14%. In contrast, the Consensus-based approach results in a higher redundancy reduction rate of 48.7% and 31.9% for bigram and trigram respectively although the bigram-based approach produces a maximum number of corrupted clusters. Interestingly, all the clusters resulting from complete dimension are single-domain protein suggests the approach is not suitable for multi-domain proteins. The above results suggest that consensus-based sequence clustering with higher dimensional representation is sensitive to both single-domain and multi-domain proteins and highly powerful in removing redundant sequences.

In the second phase, domain-based analysis is incorporated to quantify the cluster quality in the context of biological importance. Domain annotation data for protein sequences are collected from Pfam database [241]. The quality of the clusters is analyzed through Domain Correspondence Score for each non-singleton cluster. If all the proteins in a cluster agree with the same Pfam domain, the cluster is considered as pure or otherwise corrupted (*see* Table 2.2). Among the non-singleton clusters, TriGram_con surpasses other*state-of-the-art* methods in creating high-quality compact clusters as 84.2% having Domain Correspondence Score 1 whereas in Cd-hit is 84.1% but the number of non-singleton clusters is more than 1/3 of TriGram_con.

2.4 Discussion

Here, in this work, a two-level parallel DBSCAN clustering for human protein sequences that address the computational issues of large-scale biological data processing and analysis is presented. The experimental result showed efficient speed-up in the proposed method and effectively reduces the redundant from sequences as the method has achieved a higher sequence/cluster score (1.423) considering only two corrupted clusters. In the proposed method with trigram feature (consensus), the percentage of non-singleton clusters having Domain Correspondence Score=1 is higher than in other *state-of-the-art* approaches. The quantitative evaluation shows that the clustering results improved with higher values of n and the speed-up ratio improves with increasing data size.

In this study, it has been discussed that the development of modern high-throughput sequencing techniques has resulted in exponential growth of proteomic data in recent years. Analysis of this large amount of protein sequence data with existing resources is extremely challenging with existing computational resources. The computational issue can be resolved by analyzing the problem using a Big Data framework. Not only protein sequences, but the variety and veracity issue plays a big role in such a large-scale dynamic PPI network. Handling large, dynamic, heterogeneous protein interaction networks using existing computational methods is really a tedious job. This became the basis of developing an organism-specific computational model for large-scale protein interactions. In the following chapter, an *in-silico* model has been proposed for developing fuzzy semantic scape for human proteome using GO graph-based approach.

Chapter 3

Large-Scale Interaction of Human Proteome at Fuzzy Semantic Space

3.1 Background

Through their interactions with other proteins, proteins play a variety of biological and vital roles in cellular processes. Exploring the contact affinity of protein pairs is a very significant study topic in computational biology to reveal cellular and molecular capabilities, signaling pathways, and critical insights into disease mechanisms. This chapter presents a GO graph-based approach for computing the fuzzy semantic score (FSS) of the human proteome. Large-scale PPI network of an organism provides key insights into its cellular and molecular functionalities, signaling pathways, and underlying disease mechanisms. For any organism, the total number of unexplored protein interactions significantly outnumbers all known positive and negative interactions. Proteins are involved in various biological functions in the cell through interactions with other proteins. Such interactions are often modeled as graphs with proteins as nodes and interactions as binary edges. These graphs are widely called protein-PPINs [242, 243]

A large-scale PPI network of an organism provides valuable clues for understanding cellular and molecular functionalities [129], reconstructing signaling pathways [130, 131], multi-molecular complex detection [243], disease associations analyses [56, 132], drug discovery [133,244,245], etc. Tremendous effort has been invested into developing in-vitro experimental methods to extract positive PPIs [246–249]. However, these experimental methods are able to produce only a fraction of the PPI network of an organism while being costly and time-consuming.

Several computational approaches have been introduced to support PPI prediction [66, 126, 250–256]. The effectiveness of many of the *in-silico* PPI prediction methods depends heavily on the selection of the positive and negative datasets during the training process. While biologically validated positive PPI datasets are available, negative databases are scarce [257, 258]. Therefore, *in silico* curation of negative datasets is an interesting research problem for the bioinformatics community. If an interactome is being considered with respect to any given organism, the total number of unexplored protein interactions significantly outnumbers the known positive/negative interactions, even if the ones obtained from *in-silico* methods are considered. For example, in the human proteome dataset from UniProt database [259], 20 350 reviewed human proteins result in $\sim$ 207 million potential interaction possibilities. However, so far the information available is about $\sim$ 5.61 million positive and $\sim$ 0.76 million negative interactions, including both computational and experimental data. Together this is $\sim$ 3.1% of the interaction possibilities in the Human interactome. Therefore, there is a large number of unexplored interactions that need to be assessed. One of the major challenges here is the cost of experimental PPI detection in the *in-vitro* case and the sheer size of the complete interactome and lack of supporting biological data in the *in-silico* case. Conventional PPI prediction methods generate binary interactions, whereas recent studies show that weighted interactions are proffered in pathway and/or disease analyses [134]. As has been shown in [260] protein-protein binding affinities may be effectively assessed with the help of the associated ontological networks.

3.1.1 GO-based Model

Here, the probability of interaction between proteins based on the knowledge from GO [261] knowledge base is being assessed which is the world's largest source of information on the functions of genes. GO is a controlled and structured vocabulary of ontological terms that describe information about protein's localization within cellular components (CC), participation in biological processes (BP), and association in molecular function (MF) (*see* Figure 3.1). GO terms are grouped into three independent direct acyclic graphs (DAGs) where nodes represent specific GO terms and the links among nodes represent different hierarchical relationships.

Each of the three sub-graphs has a single most generic GO term, that can be considered the root in the sense that all the other nodes of that sub-graph are reachable from it. The root nodes are GO:0005575, GO:0003674 and GO:0008150 for CC, MF and BP, respectively. Because of that, tree oriented notions are often generalized and used with GO sub-graphs. For example a specific GO term is considered a descendant of the more generic GO terms from which it has incoming edges [262–264].

The UniProt-Gene Ontology Annotation (UniProt-GOA) database (see `http://geneontology.org/docs/go-consortium/`) includes information about GO annotation of proteins in the UniProtKB dataset. For any protein, GO annotations are broadly classified into two groups based on inference evidence: inferred from electronic annotation (IEA) and manually reviewed (non-IEA). The non-IEA evidence includes data inferred from Experiment (EXP), Direct Assay (IDA), Expression Pattern (IEP), High Throughput Direct Assay (HDA), etc. (for details see `http://geneontology.`

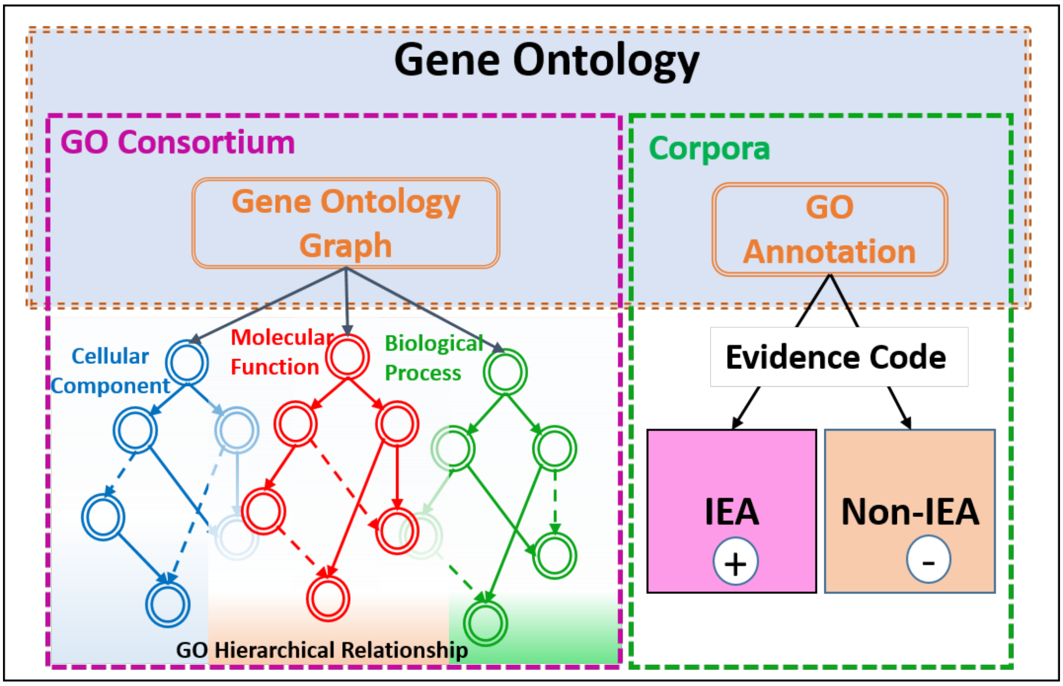


Figure 3.1: Schematic diagram of Gene Ontology details. Ontologies are organized in 3 hierarchical relationship graphs.

`org/docs/guide-go-evidence-codes/`). The inclusion of IEA which provides $\sim 25\%$ of all annotations enhances coverage. In practice most of the semantic assessments with GO are determined in two orientations, one being 'with-IEA' and the other 'without-IEA'. This results in a total of six semantic assessment scores for each pair of proteins, *i.e.*, IEA, and non-IEA scores for CC, BP, and MF, respectively. Here, such scores are computed and combined into a single semantics assessment score, ranging from 0 to 1, which is referred to as FSS of binding affinity.

While writing this thesis, the GO dataset contains 44,579 annotations (release-version:*02-2020*) of which 29,267 are BP, 11,126 MF, and 4,186 CC. The GO terms annotated to proteins can be used to infer functional relationships between them. To compute the binding affinity between two proteins it is necessary to estimate the semantic similarity scores of all the GO term pairs, where the first term is annotated to one of the proteins and the second to the other. Semantic similarity computation for any GO pair by exploring three different GO relationship graphs is a time-consuming task. Thus there is a strong need for a method that can handle large data and is easy to distribute.

Here, an efficient algorithm has been proposed to compute the fuzzy binding affinity for large-scale PPIs. A high throughput parallel architecture based on Apache Spark is being used to implement the proposed algorithm. Spark is a popular platform for large-scale data analysis [265, 266]. Its main advantages over previous frameworks like Hadoop are expressive and high-level API, and the ability to keep intermediate results in memory between computation phases. The latter saves disk I/O and results in a huge efficiency gain. Furthermore, Spark has sub-modules for data analytics, graph processing, machine learning and streaming so combining such applications in one project does not introduce any integration overhead at the same time Spark core optimizations, of which there are many, are applied on the end-to-end data processing pipeline.

Several methods have been developed to measure the semantic similarities between protein/gene pairs. The existing approaches can be classified into three broad categories based on the information that has been used from the GO relationship graphs. These are node-based, link or edge-based, and hybrid methods. Edge-based methods rely on the distance between two GO terms in the GO graph [267–269]. Usually, the distance is computed as the number of edges in the shortest graph path between two terms or as an average over all paths. Such distance can easily be normalized and converted into a similarity measure. Another way to compute similarity is to check the depth of the first common ancestor of the nodes. The deeper from the root of the sub-graph the common ancestor is, the more the nodes have in common. Unfortunately, the nodes and edges in the GO sub-graphs are not distributed uniformly, nor edges at the same level in the ontology correspond to the same semantic distance between terms [270]. In [271] such issues were addressed by weighting edges differently according to their hierarchical depth or taking into account node density and link type. Yet, this does not fully solve the problem and edge-based strategies and edge-based approaches are considered ineffective in terms of semantic assessment [260].

Node-based approaches utilize the properties of the pair of GO terms themselves and their ancestor or descendant nodes [272–275]. Sometimes the concept of information content (IC) is used [276], which is a measure of how specific and informative a GO term is. Resnik [272] has defined semantic similarity as the IC of the most informative common ancestor (MICA) of two GO terms. The MICA and the lowest common ancestor node refer to the same ancestor of two GO terms where the MICA is presented in the context of searching common path between GO terms, and the latter is presented in the context of IC of GO terms. In semantic similarity computation, several methods have used the IC values of query proteins [273, 274]. Another IC-based approach has been proposed by Schlicker *et al.* [275] where the relevance similarity measure has been defined using the location of the query GO terms in the DAG by considering the properties of MICA [277]. Mazandu and Mulder [278] have proposed a method that normalizes the IC-based semantic similarity to 1 when measuring the similarity between the same GO terms. A new approach, *GraSM* has been introduced in [279] to avoid the over-reliance on MICA. It is designed in such a way that it can be applied to any IC-based method where the semantic similarity is calculated by the average IC of the disjunctive common ancestors (DCAs). The DCAs are identified by the number of distinct paths from the query GO terms to MICA [279].

The IC-based methods have an obvious advantage, as they use IC to indicate the specificity of a GO term and are because of that free from the problems of nonuniform semantic distance and edge density. However, corpora-dependent IC calculation can cause problems. In general, IC-based approaches suffer from two major issues. First, the IC calculation depends on the annotation corpora set, as the same GO term may have different IC values when different corpora are used [280]. Second, the IC is biased by the research trend [270], as the GO terms related to popular fields tend to be annotated more frequently than the ones related to less popular fields and the annotation of some terms may still be missing in the corpus [277]. These issues largely degrade the overall performance and effectiveness of methods that only use ICs.

To overcome the limitations of the IC-based approaches, many hybrid methods have been developed that consider both edges and nodes in the DAG. Wang *et al.* [253] have proposed a hybrid GO-universal method that calculates the semantic similarities based on the topology of GO DAG. It takes into account the topological position characteristics in the GO sub-graphs and considers the number of children's terms instead of the frequency of terms from the annotation corpus. GO-universal defines the topological position characteristic of the root to be 1 and calculates the topological position characteristic of a non-root GO term by using a ratio based on the number of children of all ancestor GO terms [281]. A hybrid structural similarity-based method has been proposed by Nagar and Al-Mubaid [282] using the shortest path plus either IC generated from corpora or structure-based IC generated from DAG. In recent work, Dutta *et al.* [128] have proposed a hybrid semantic similarity measure between two GO terms based on a combination of topological properties of the GO graph and average IC of the DCAs of the GO terms.

Although GO-based methods are popular in assessing PPI binding affinity, one of the major challenges lies in managing the computational overhead. For example, in a human PPI network, the total number of protein interactions to be explored is in the order of hundreds of millions. This has been the primary concern for most of the prior studies [128, 277, 279, 281], which concentrated on a smaller subset of the complete dataset. Due to the proposed distributed architecture and optimized algorithm, the problem has been eliminated. In the proposed work, the complete human proteome has been used, and the underlying GO annotations for efficient assessment of the PPI binding affinity and propose a novel fuzzy semantic score. The outcome of the approach is not limited to quantitative analysis. The fuzzy semantic network helps to understand different biological mechanisms such as protein complex module identification and functional analysis, high-quality data identification for the PPI model, PPI interaction analysis, etc. The work includes,

- Design of a fuzzy semantic score of binding affinity using the GO network to assess the binding affinity between any two proteins at an organism level. Here, the work has been done using human proteins.

- Implementation of the underlying algorithm using a Spark-based parallel architecture and using it to process a human PPI network of $\sim$ 180 million potential interactions resulting from 18,994 reviewed proteins for which GO annotations are available.

- Construction of fuzzy semantic network at proteome level from the above designed binding affinity function and extraction of meaningful biological insights.

- Validation of the developed method with respect to the available *state-of-the-art* methods on benchmark datasets.

- Development of a *Fuzzy*PPI web-server with precomputed PPI affinity scores, available freely for non-commercial use, at: `http://fuzzyppi.mimuw.edu.pl/`.

3.2 Dataset

Here two datasets are used. The first one, obtained from UniProt, contained information about human proteins and their GO annotation. It has been used as the base for predictions. The second one has been obtained by combining several popular datasets and is composed of experimentally validated information about positive and negative interactions between human proteins. It has been used as the benchmark for analysis.

3.2.1 Proteome Data

The UniProt knowledgebase (UniProtKB) [259] is the combined and uniform collection of proteins from Swiss-Prot [283], TrEMBL [284], and PIR-PSD [285]. This database maintains a bi-directional cross references between *"PIR-PSD"* and *"Swiss-Prot + TrEMBL"* protein entries. The main objective of this database is to maintain a single entry and to merge all reports for any particular protein [283]. Thus, it reflects a complete structure of protein repository for different organisms where many of the entries are derived from genome sequencing projects.

As UniProt is highly popular, with over 0.4 million unique visitors to its website per month [286] and over 20,000 total citations of its main publications, a thorough analysis and annotation by biologists ensure the quality of this database.

In UniProt, the human proteome (*id: UP000005640*) contains 20,350 manually annotated human proteins. Among these, 18,994 proteins are annotated with GO annotations which is $\sim$ 93% of reviewed proteins. In the proposed experiment, the fuzzy binding affinity is assessed for the resulting $\binom{18{,}994}{2} \sim 180$ million pairs of unique proteins.

UniProt allows to export protein data together with annotations in a text format presented in first part of Figure 3.2. For each protein with UniProt id (e.g. P29274) and annotated to it GO term (e.g. GO:0030673) there is information about GO type (C, F or P) and evidence type (IEA, non-IEA). This is encoded in a hierarchical format as shown in Figure 3.2 and store as a Parquet file which is a binary and columnar format integrated with Spark with optimizations that allow significant performance improvement.

3.2.2 Protein Interaction Data for Benchmarking

To prepare the benchmark dataset experimentally validated human PPIs from several popular datasets have been used such as HIPPIE [287], STRING [37], BioGRID [35], DIP [34], HuRI [288] for positive data and Negatome 2.0 [257], Trabuco *et al.* [258] for negative data. The statistics of the interaction data are presented in Table 3.1. All the databases are combined into a single dataset with ternary information regarding the interaction status of the protein *i.e.* whether the protein pairs interact, do not interact, or are unknown. It has been considered that there is an interaction between two proteins if there was evidence for that in any of the positive datasets and no evidence for the lack of interaction in all of the negative datasets. The other way around, it has been considered that there is no interaction between two proteins if there was evidence for that in any of the negative datasets and no evidence for interaction in all of the positive datasets. Gold samples of interactions are also being distinguished with high confidence. The resulting combined dataset has information about 5,107,321 positive interactions, 730,122 negative interactions and 18,295 out of the total 18,994 proteins are somehow accounted. For the gold sample, the statistics are 361,076 positive interactions, 182,667 negative interactions, and 15,261 proteins. The statistics of interactions for individual dataset are presented in Table 3.1

The databases have different scoring schemes to quantify the interaction quality. For example for BioGRID the values are in the range $[-\infty, +\infty]$, while for HIPPIE in [0,1]. This is summarized in Table 3.2 where the threshold is also specified for considering two proteins to interact and for including the pair in the gold sample. For other databases binary information is present but the data is grouped in high and low-confidence groups. For example, for DIP there are two groups core and non-core and it is considered that both of them as the source of evidence for interactions, but only core as the source of evidence for high-quality interactions. Sometimes, as is the case for HuRI, one group is a superset of the other, a high-quality group. Finally, *GoldPos* and *AllPos* interaction datasets are constructed by considering the union of individual gold data and all data respectively from the above-described five positive interaction databases. The degree-range specific density plot of all proteins is shown in Figure 3.3.

Recall, that for negative interactions it is required that no positive interaction

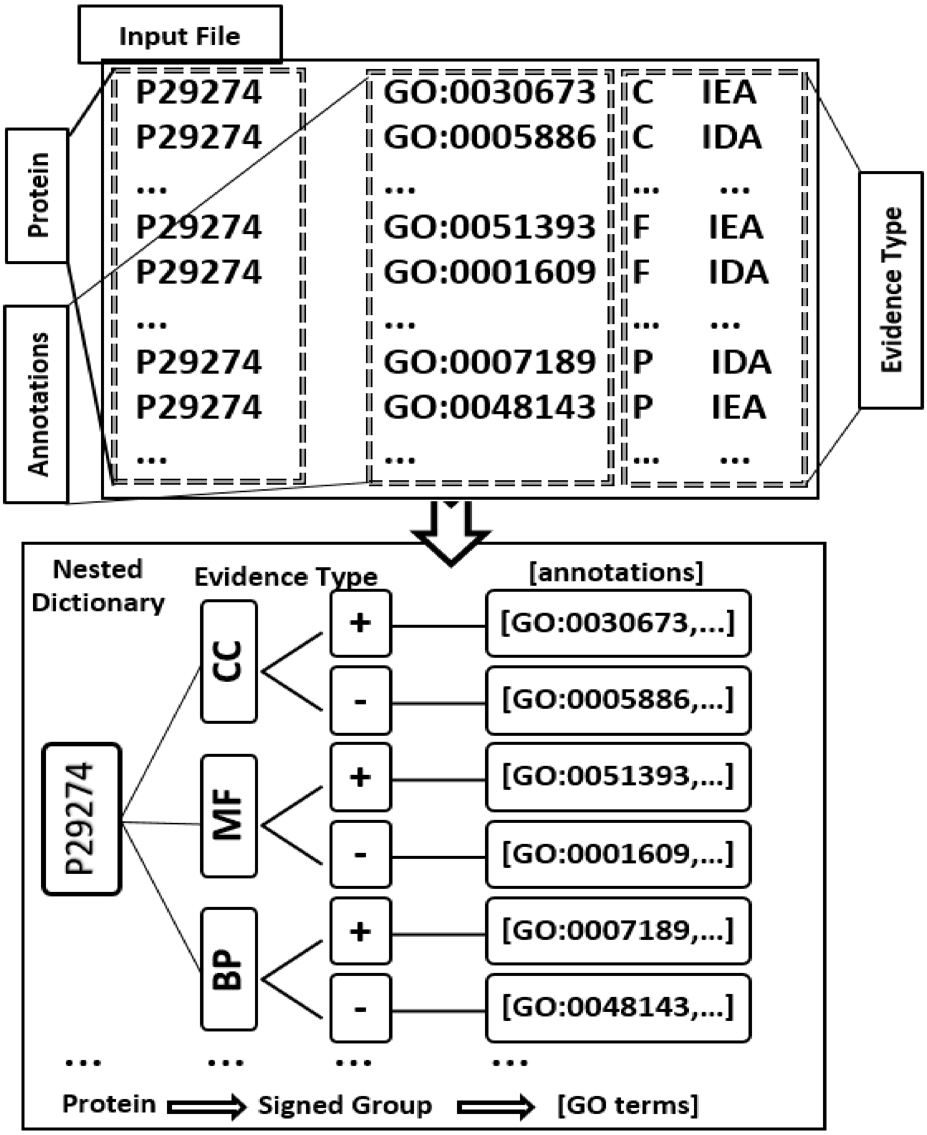


Figure 3.2: Transformation of UniProt text format to hierarchical representation of Protein-GO annotations.

Table 3.1: Database wise proteome level statistics of interactions of UniProt reviewed proteins.

Database	Existing Intr.		Unexplored Intr.	
	Individual	Total	Total	(in %)
DIP [34]	4,652	5,611,787	**200,680 877**	**96.9%**
HuRI [288]	64,711			
BioGRID [35]	104,509			
HIPPIE [287]	342,996			
STRING [37]	5,415,071			
Negatome *2.0* [257]	1,482	758,411		
Trabuco *et al.* [258]	756,994			

exists in any of the datasets. In Negatome *2.0*, [257] the negative data are categorized as gold based on the selection strategy as of Manual stringent (MS). For the dataset by Trabuco *et al.* [258], high-quality negative set is selected by removing the positive interactions that have any type (*in-vivo*, *in-vitro* and *in-silico*) of evidence of positive interaction (*AllPosR*) while all negative set is constructed by removing the interactions that belong to the gold quality positive (*GoldPos*) interactions. Note that the *AllPosR*

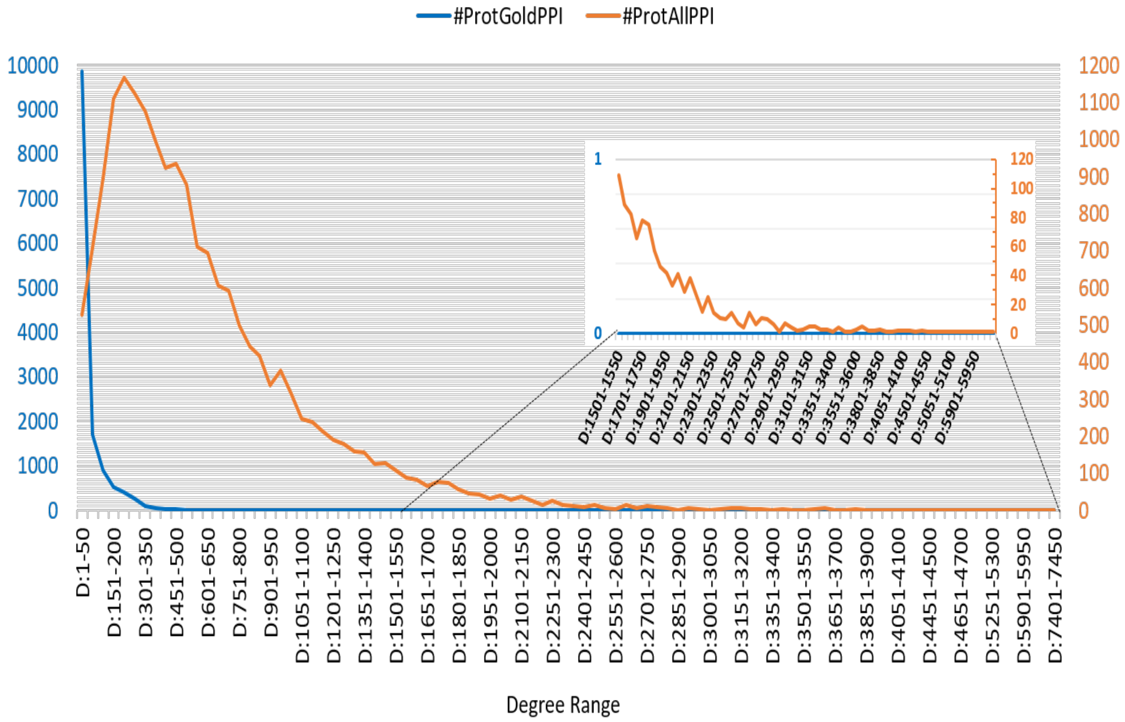


Figure 3.3: Degree-range wise plot of positive interactions.

Table 3.2: Details of benchmark interaction data selection

Database	Score range/groups	All-Interactions	Gold Quality
BioGRID	$[-\infty, +\infty]$	≥ 0	≥ 10
HIPPIE	$[0, 1]$	≥ 0.75	≥ 0.9
STRING	$[0, 1]$	≥ 0.15	≥ 0.9
DIP	*core(C), non-core(NC)*	*C+NC*	*C*
HuRI	*HuRI(H), HuRI-Union(HU)*	HU	H
Negatome *2.0*	*Manual(M) , Manual-stringent(MS)*	*M+MS*	*MS*
Trabuco *et al.*	Negative (N^T)	$N^T - GoldPos$	$N^T - AllPosR$

set includes *in-silico* interactions where *AllPos* excludes *in-silico* interactions.

3.3 Methodology

Building on the model from work [128] in this section it is formally defined how the fuzzy semantic network can be used to assess the interaction affinity of any two proteins with GO annotations. The semantic similarity between any two proteins is estimated based on the similarities of pairs of GO terms such that one GO term annotates the first protein and the other the second. The pairs are considered independently for each of the GO sub-graphs, *i.e.*, it is considered that only pairs of terms from the same GO sub-graph. For each GO sub-graph, semantic similarity is computed in two ways based on the evidence type of annotations, *i.e.*, for IEA and nonIEA which are denoted as + and -, respectively. A Spark-based parallel implementation of this scheme has been designed by leveraging its high throughput architecture for large-scale proteome-level interaction analysis. The implementation details of the fuzzy function and parallel processing are presented in the following sub-sections.

3.3.1 GO Annotation

It has been assumed that the annotation of proteins with GO terms (*see* Figure 3.4 (A)) is available as $terms_{SG}^{iea}(p)$, such that for each protein p, a GO sub-graph, $SG \in \{CC, MF, BP\}$, and the evidence type annotation $iea \in \{+, -\}$ *i.e.* IEA/nonIEA it returns the set of all GO terms from V_{SG} that are annotated with specified IEA/nonIEA annotation type iea. For example, as shown in Figure 3.4(A), $terms_{CC}^{+}(P29274) =$

$\{GO : 0030673, \ldots\}$ while $terms^{-}_{MF}(P29274) = \{GO : 0001609, \ldots\}$. Similarly, $prot^{iea}(t)$ has been defined, such that for each protein t and evidence type annotation $iea \in \{+, -\}$ *i.e.* IEA/nonIEA it returns the set of all proteins that are annotated with term t, *i.e.*, $prot^{iea}(t) = \bigcup_{GO \in \{CC,BP,MF\}} \{p \in V_{SG} || t \in terms^{iea}_{SG}(p)\}$.

3.3.2 GO Sub-graphs

Here, the standard notation is being used to denote a directed graph $G = \langle V, E\rangle$, where the first element of the tuple V is the set of nodes and the second E the set of edges between the nodes, i.e., $E \in V \times V$.

Let $CC = \langle V_{CC}, E_{CC}\rangle$, $MF = \langle V_{MF}, E_{MF}\rangle$ and $BP = \langle V_{BP}, E_{BP}\rangle$ be the GO sub-graphs for CC, MF, BP, respectively. By assumption $V_{CC} \cap V_{MF} = V_{CC} \cap V_{BP} = V_{MF} \cap V_{BP} = \emptyset$ from which it follows that $E_{CC} \cap E_{MF} = E_{CC} \cap E_{BP} = E_{MF} \cap E_{BP} = \emptyset$. It is also assumed that *anc-or-self* and *desc-or-self* functions are defined for each graph node with standard meaning

3.3.3 Fuzzy Semantic Scoring for Proteins

The prevailing theory holds that proteins with similar roles are most likely to interact with one another. Also the more functions they share and the more specific the common functions are the higher the interaction chance is. The functions of proteins are concluded from GO annotations that these proteins have. Different GO sub-graphs represent different types of functionality and are examined separately and then combined. This way numerous annotations from one GO sub-graph do not dilute possibly less numerous but highly shared annotations in other sub-graphs.

At first, the semantic score (SS) of binding affinity for pairs of proteins is defined and then normalize it into fuzzy semantic score. The scoring of proteins is based on the scoring of the pairs of GO terms annotated to them that is defined in Subsection 3.3.4. For any pair of proteins p_a and p_b SS of binding affinity is defined as:

$$SS(p_a, p_b) = \sum_{\substack{SG \in \{CC,MF,BP\} \\ iea \in \{+,-\}}} \frac{S^{iea}_{SG}(p_a, p_b) + S^{iea}_{SG}(p_b, p_a)}{|terms^{iea}_{SG}(p_a)| + |terms^{iea}_{SG}(p_b)|} \tag{3.1}$$

where $S^{iea}_{SG}(p, q)$ is the similarity of proteins p and q within the GO sub-graph, SG, based on the IEA/nonIEA annotation type iea, and is defined as:

$$S^{iea}_{SG}(p, q) = \sum_{t_p \in terms^{iea}_{SG}(p)} max_{t_q \in terms^{iea}_{SG}(q)} sim_{SG}(t_p, t_q) \tag{3.2}$$

and $sim_{SG}(t_p, t_q)$ is the semantic similarity for a pair of GO terms that is defined in Subsection 3.3.4. Semantic score is normalized with the max-min normalization. It linearly transforms SS into fuzzy semantic score, FSS:

$$FSS(p_a, p_b) = \frac{SS(p_a, p_b) - min_{p,q}(SS(p,q))}{max_{p,q}(SS(p,q)) - min_{p,q}(SS(p,q))}. \tag{3.3}$$

Note that fuzzy semantic score results are limited to [0, 1] and can be viewed as the probability of interaction between proteins p and q.

3.3.4 Semantic Similarity for GO terms

In the following Subsections, $MDMS_{SG}(t_i, t_j)$ is defined to be the maximal difference in membership strengths (see Subsection 3.3.4) of the GO terms t_i, t_j to natural clusters covering of sub-graph and $SIC(t_i, t_j)$ to be the shared information content (see Subsection 3.3.4) between the target GO-terms. With those, it defines the *semantic similarity* of a pair of GO terms that come from the same GO sub-graph.

Definition 3.1 Semantic Similarity for a GO sub-graph, and any two vertices t_1 and t_2 where $t_1, t_2 \in V_{GO}$ the semantic similarity of t_1 and t_2 is defined as:

$$sim_{SG}(t_1, t_2) = (1 - MDMS_{SG}(t_1, t_2)) * SIC(t_1, t_2) \tag{3.4}$$

Note that the semantic similarity of two terms is highest if at the same time their maximal difference in membership strengths is small and shared information content is high.

Maximal Difference in Membership Strengths (MDMS)

The $MDMS_{SG}(t_i, t_j)$ is computed with respect to the set CC_{SG} of cluster centers of sub-graph, which is defined in the next subsection. Each cluster center computes how close other nodes are to it, *i.e.*, how strong their membership to this cluster. Large differences in membership mean that nodes are not similar. Following [128] Gaussian is being used to convert the distance to the cluster center into similarity and compute the maximum difference of such similarities overall cluster centers:

$$MDMS_{SG}(t_i, t_j) = max_{c \in CC_{SG}} \left| e^{-\frac{d(c,t_i))^2}{2k^2}} - e^{-\frac{d(c,t_j))^2}{2k^2}} \right| \tag{3.5}$$

where $d(t_1, t_2)$ is the smallest distances between nodes t_1 and t_2 taken in V_{SG} and in V_{SG}^{-1}. That is $d(t, t) = 0$ and for $t_1 \neq t_2$ $d(t_1, t_2)$ has been defined as the smallest n such that there exists a sequence of nodes $w_1, w_2, \ldots, w_{n-1}$ for which there exists a path $(t_1, w_1), (w_1, w_2), \ldots, (w_{n-1}, t_2) \in V_{SG}$ or a path $(t_1, w_1), (w_1, w_2), \ldots, (w_n, t_2) \in V_{SG}^{-1}$.

Natural Cluster Covering of a GO Sub-graph

In [128] the cluster centers CC_{SG} for each GO sub-graph were chosen to be the nodes with high values of ratio of nodes reachable from them together with nodes from which they are reachable to the total number of nodes in the graph. The ratio is briefly defined first and then explained how it has been improved upon the selection of cluster centers.

To formally define the ratio, transitive closure of the edge relation E^* is used to define as the set of all node pairs $\langle u, v \rangle \in V \times V$ such that there exists a (u, v)-path in the graph, *i.e.*, $(u, w_1), (w_1, w_2), \ldots, (w_n, v) \in E$. The node v is reachable from u if $\langle u, v \rangle \in E^*$ and the set of all nodes reachable from u is denoted as $E^*(u)$. With this, the ration used to determine the cluster centers for any $v \in V$ is defined as:

$$PrpMes_{SG}(v) = \frac{|\{u \in V_{SG} | u \in E^*_{SG}(v) \vee v \in E^*_{SG}(u)\}|}{|V_{SG}|} \tag{3.6}$$

The ratio-based method proposed in [128] and other works leads to selecting 49 cluster centers for CC, 32 for MF, and 68 for BP. Here, centroid-based clustering methods are adapted to work on the respective graphs that consider the graph topology by minimizing the distance between cluster centers and cluster members. The best results with the K-Medoids Clustering method [289] are obtained. Starting from the centers obtained from [128] and adding more centers using the k-medoids++ initialization method until each term was reachable from some center. This resulted in 149, 728, and 577 GO nodes as cluster centers from CC, MF, and BP sub-graphs respectively. The new sets of centers share 23 (CC), 19 (MF), and 28 (BP) common GO terms with Dutta *et al.* with 0.131 (CC), 0.126 (MF), and 0.045 (BP) Jaccard similarity. In the gold standard positive and negative PPI dataset, the newly developed cluster center selection algorithm has improved the overall performances (best Area Under Curve *AUC* score: 0.780 at $\mu = 0.15$) compared to Dutta *et al.* [128] having best AUC score of 0.764 at $\mu = 0.15$) and in *ALL-PPI* dataset, the improvement is around 1.5%.

Shared Information Content (SIC)

Now it is taken into account how informative the common ancestors of a pair of terms are. The intuition is similar to the notion of term frequency–inverse document frequency (TFIDF) [290] where rare words tend to more correctly summarize a document than very common words. Similarly, some terms that are commonly annotated to proteins are less useful in comparing them as opposed to rarely occurring terms. For that IC of a term in GO sub-graph has been defined with respect to the evidence type annotation iea:

$$IC_{SG}^{iea}(t) = -\log \frac{|\sum_{t' \in anc\text{-}or\text{-}self(t)} prot_{SG}^{iea}(t')|}{\sum_{t' \in V_{SG}} prot_{SG}^{iea}(t')} \tag{3.7}$$

To obtain the SIC between a pair of GO terms, the average IC of all DCAs of the GO terms is computed [128]:

$$SIC_{SG}^{iea}(t_i, t_j) = \frac{\sum_{a \in DCA(t_i, t_j)} IC_{SG}^{iea}(a)}{|DCA(t_i, t_j)|} \tag{3.8}$$

The DCA of any GO terms t_i, t_j are those common ancestors a such that for a given difference of number of paths to a from t_i, t_j they have the highest information content. Let $CA(t_i, t_j)$ be the set of all common ancestors of terms t_i and t_j defined as $CA(t_i, t_j) = anc\text{-}or\text{-}self(t_i) \cap anc\text{-}or\text{-}self(t_j)$. Let $PD(a, t_i, t_j)$ be the difference in number of (a, t_i) and (a, t_j) paths. The DCA is defined as follows:

$$\begin{aligned} DCA(t_i, t_j) = \{a \in CA(t_i, t_j)| \\ \forall_{a' \in CA(t_i, t_j)} PD(a, t_i, t_j) = PD(a', t_i, t_j) \\ \implies IC(a) \geq IC(a')\} \end{aligned} \tag{3.9}$$

3.3.5 Parallel Implementation with Spark

The distributed algorithm and its Spark implementation of the fuzzy semantic scoring function is presented in this section. Given a set of proteins, GO sub-graph, and annotation type *iea i.e.* IEA/nonIEA that defines GO terms assigned to the proteins, the algorithm computes $S_{SG}^{iea}(p, q)$ as defined in section 3.3.3. The algorithm is divided into four phases that include: one phase of pre-processing of the input data, two phases of pre-computation of values that otherwise would be computed many times, and the final computation using the pre-computed values. The overall parallel implementation is detailed in the following Figure 3.4.

Preprocessing: Hierarchical representation of Protein-GO annotation data

In the first phase, the annotation data is parsed and pre-processed into a hierarchical data structure representing *iea* annotations. The information about 18,994 proteins and their annotation with $\sim$ 266 million GO terms is stored as a text file (for sample representation *see* in Figure 3.4(A)) and is considered as an input to the pre-processing step. The information is organized into a hierarchical structure in which each protein ID is mapped to a list of its GO terms grouped by GO sub-graphs and annotation types. This structure is used to extract the dataset of proteins with their respective GO terms for selected sub-graphs and *iea*, which is needed in subsequent algorithm phases.

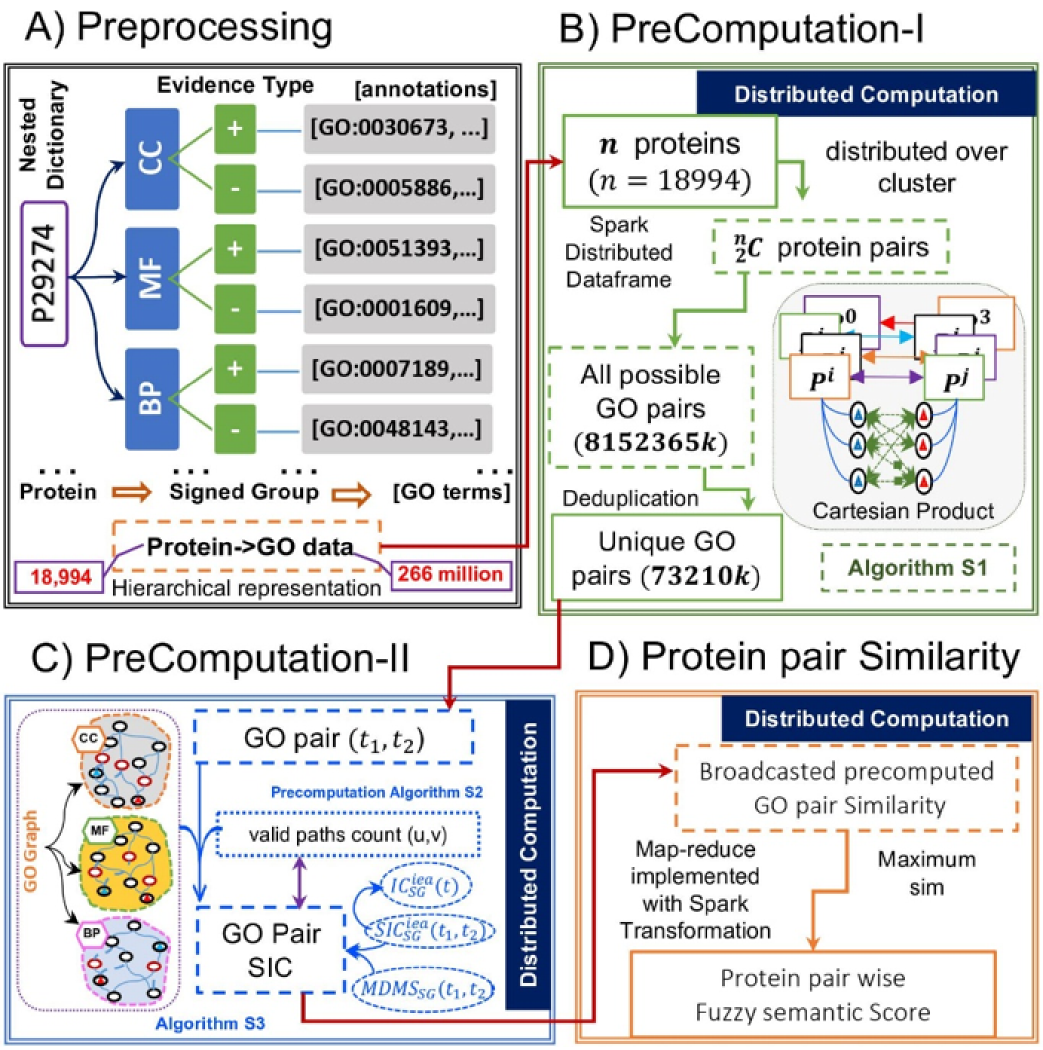


Figure 3.4: Parallel implementation details with Spark. **A)** Preprocessing: Representation of Protein-GO annotation data into a hierarchical structure in which each protein ID is mapped to a list of its GO terms grouped by GO sub-graphs and annotation types. **B)** PreComputation-I: Distributed computation of GO pair generation from all possible protein pairs which is quadratic with respect to the number of proteins. **C)** PreComputation-II: The semantic similarity computation for independent unique GO pairs by exploring GO sub-graphs. **D)** Similarity Computation at the protein pair level using the broadcasted semantic similarity from the previous step. Finally, the maximal similarity is computed using map-reduce implemented with Spark transformations followed by the normalization phase.

As the size of the input data is small, this phase does not have to be distributed.

PreComputation-I: Finding unique GO term pairs from all protein pairs

The main goal of the work is to compute Fuzzy Semantic Score for each protein pair as defined in section 3.3.3. For that, it is needed to calculate and combine semantic similarities of all GO term pairs for the protein pair. As the number of protein pairs is quadratic with respect to the number of proteins, this computation is substantial

Algorithm 3.1: UniqueGOPairComputation

```
Input   : Protein with GO annotations
Output: A set of non-redundant GO pairs from all possible protein pairs
                                    /*  Cartesian product of proteins   */
cartesianDF = proteinsDF.join(proteinsDF).where
 (($p1, p2$) ⇒ $p1.protein < p2.protein$)
                                    /* GO pairs for all protein pairs */
pairsDF = cartesianDF.flatMap(($p1, p2$) ⇒
  for go1 in p1.goterms do
    for go2 in p2.goterms do
      yield (min(go1, go2), max(go1, go2))
    end
end
)
                                    /*    Removing duplicate pairs     */
result = pairsDF.distinct.collect
```

and should be distributed. The pseudocode is presented in Algorithm 3.1. Naive implementation would repeat computation of semantic similarities for GO term pairs for different protein pairs. In order to avoid this, semantic similarities are being precomputed for all pairs that appear at least once for any protein pair. Using Spark, the set of proteins along with their GO terms is distributed on the cluster in the form of a Spark distributed collection DataFrame. Then, the Spark built-in mechanism is leveraged to obtain the Cartesian product of the DataFrame with itself. Those rows of the product are kept in which the first protein is lexicographically smaller than the second one to get rid of duplicate combinations. Then, using Spark's flatMap transformation, DataFrame is obtained for all GO term combinations for all proteins. Also, it has been taken care of not to distinguish the pairs where the same elements are in a different order. Finally, duplicate pairs are removed with Spark distinct transformation, which is implemented efficiently with grouping and combining on the map side of the shuffle. The resulting set of unique GO term pairs for human proteins is small enough which is 8,152,365k pairs before de-duplication and canonical ordering, and 73,210k after de-duplication to be collected to a single machine.

PreComputation-II: Computation of the semantic similarity for unique GO term pairs

Next, the semantic similarity score is computed, as defined in subsection 3.3.4, for all unique GO term pairs obtained in the previous step. The computation for each pair is independent so it can be easily distributed if the number of GO term pairs grows large.

Algorithm 3.2: numberOfPaths

```
Input  : GO Graph $G(V, E)$
Output: numberOfPaths$[u][v]$ – number of valid paths from any node $u$ to its
        successors node $v$. $u, v \in V$

numberOfPaths ={}
for $u$ in $V$ do
    numberOfPathsFromNode(u)
end
```

Thanks to limiting of the number of pairs in the previous step, for human proteins the score needs to be computed for only 73,210k pairs which can be completed in half an hour by a sequential, centralized algorithm and with further linear speedup due to distributing. The resulting similarities take little more than 1GB and can be easily broadcast to every node in the cluster for the next step.

To compute SS for all the GO term pairs, at first SIC scores are computed (see 3.3.4) for each pair, then $MDMS$ (see subsection 3.3.4) for each pair, and finally both the scores are combined to get the semantic similarity.

It starts by explaining how to compute the SIC score for each GO term pair. Based on the GO sub-graph, for each pair (t_1, t_2) of GO terms present in the sub-graph the number of different paths from t_1 to t_2 are precomputed. This can be done in time quadratic to the graph size using depth-first search. The result can be stored in a sparse form as a map that for each node stores a map from another node into the number of paths between them. The lack of entry in the map for a given node pair denotes that there are no paths between the pair. The relevant pseudocode is presented in Algorithm 3.2

The information content IC_{SG}^{iea} is also precomputed (see subsection 3.3.4) for all the GO terms. For that, once all the paths are computed for each GO term as well as IC scores, SIC score can also be computed for each GO term pair. The pre-computed structure with the number of paths is used to iterate over all of the descendants of a term, along with the number of paths to it. The iteration is performed over all the combinations of two descendants excluding the pairs that are not present in unique GO term pairs set refereed in subsection 3.3.5 for reducing the redundancy.

The relevant maximum IC_{SG}^{iea} value is being updated by calculating the difference between the number of paths to each descendant in the pair and aggregating the value across term pairs. In the case of human proteins, $\sim 38.5\%$ of the values are discarded as only term pairs with non-zero values are allowed. The pseudocode in Algorithm 3.3 demonstrates the computation of SIC scores for all GO term pairs.

Now the computation of semantic similarity for GO term pairs can be explained. For each of the cluster centers c of sub-graph SG, the smallest distance to all other GO terms t is being computed using breadth-first search on the sub-graph and on its transposition and the minimum of both values is used as $d(c, t)$: the smallest distance between c and t. Then, for all the GO term pairs for which the SIC is calculated in the previous point, the $MDMS$ and multiply SIC and $MDMS$ are also computed to obtain semantic similarity. Note that if the SIC score for a term pair is 0, it is not present in the SIC structure, so the corresponding computation of *sim* for this pair is also absent. Additionally, as with SIC, any semantic similarity scores equal to 0 are excluded. In the case of human proteins, $\sim 0.3\%$ of values are discarded this way. As SIC, $MDMS$ and *sim* functions are symmetric, their values for ordered pairs are stored only.

```
Function numberOfPathsFromNode(u)
  Input  : GO Graph $G(V, E)$, node $u \in V$
  Output: dictionary with number of valid paths from node u to each of its
          successors nodes

  if numberOfPaths[u] is not defined then
      uPaths = {}
      uPaths[u] = 1
      for v in successors(u) do
          vPaths = numberOfPathsFromNode(v)
          for w in keys(vPaths) do
              uPaths[w] += vPaths[w] numberOfPaths[u] = uPaths
          end
      end
  end
  return numberOfPaths[u]
```

Computation of the semantic similarity for protein pairs

Finally, the computation of the similarity for each protein pair can be explained(*see* Algorithm 3.4) which follows the description in Subsection 3.3.3. First, for each protein pair and all its combinations of GO terms, the similarity is being computed using the broadcasted semantic similarity from the previous step. Then, the maximal similarity is found using map-reduce implemented with Spark transformations and used to normalize the results.

Algorithm 3.3: GOpairSICcomp

```
Input  : GO Graph $G(V, E)$, required GO pair list, numberOfPaths, IC
         scores
Output: SIC scores for GO term pairs

D ={}
for anc in V do
    paths = numberOfPaths[anc]
    score = IC(anc)
    for go1 in keys(paths) do
        for go2 in keys(paths) do
            if (go1, go2) in required GO pairs then
                diff = $|paths[go1] - paths[go2]|$
                D[go1][go2][diff] = max(D[go1][go2][diff], score)
            end
        end
    end
end
SIC ={}
for go1 in keys(D) do
    for go2 in keys(D[go1) do
        sic=mean(values(D[go1][g2o]))
        if sic ≠ 0 then
            SIC[go1][go2] = sic
        end
    end
end
```

Optimizations

In this section, the optimizations applied in the proposed implementation is described that significantly influence the performance of the computation.

Multiple variants at once: In order to compute the fuzzy semantic score for protein pair (p, q), similarities $S_{SG}^{iea}(p, q)$ across all GO term sub-graphs and annotations need to be combined. Instead of running the algorithm 6 times, once for each sub-graph and annotation variant, and going through all its stages, everything has been calculated at once and keep track of the variants.

Data format: Throughout the whole computation, all the GO terms and variant identifiers are encoded as integers instead of strings. This decreases the amount of memory needed along with the intra-cluster communication, which saves time.

Spark optimizations: The program has been implemented using Spark's Dataset API, rather than lower-level RDD API. Dataset API is less expressive, as all computations have to be translatable to SQL-like operations on structured data. Yet, this

Algorithm 3.4: ProteinPairSimComp

```
Input  : Protein list
Output: Normalized semantic similarity scores for protein pairs

                                  /*    Semantic score broadcast      */
semanticSimilaritiesBC = broadcast(semanticSimilarities)

                                  /*  Cartesian product of proteins   */
cartesianDF = proteinsDF.join(proteinsDF).where
 ((p1, p2) ⇒ p1.protein < p2.protein)

                                  /*     Similarity computation       */
simDF = cartesian.map((p1, p2) ⇒
 sem = semanticSimilaritiesBC.value
 (p1, p2, calculateSim(p1, p2, sem)
).filter((p1, p2, sim) ⇒ sim≠ 0)

                                  /*       Score normalization        */
maxSim = simDF.map((p1, p2, sim) ⇒ sim).reduce(max)
normalizedSimDF = simDF.map((p1, p2, sim)⇒      (p1, p2, sim / maxSim))
```

Function calculateSim(p1, p2, semanticSimilarities)

```
Input  : Proteins p1, p2, semanticSimilarities for GO terms
Output: Semantic similarity score for protein pair

local = {}, localT = {}
len1 = p1.goterms.length
len2 = p2.goterms.length
for go1 in p1.goterms do
    for go2 in p2.goterms do
        goMin = min(go1, go2)
        goMax = max(go1, go2)
        s = semanticSimilarities[goMin][goMax]
        local[go1][go2] = s
        localT[go2][go1] = s
    end
end
sim = (sum(row-wise-max(local)) + sum(row-wise-max(localT))) / (len1
 + len2)
return sim
```

enables several significant optimizations within Spark internal framework [291, 292], such as:

- *Memory Management and Binary Processing:* leveraging application semantics to manage memory explicitly and eliminate the overhead of JVM object model and garbage collection.

- *Cache-aware computation:* Algorithms and data structures to exploit memory hierarchy.
- *Code generation:* Using code generation to exploit modern compilers and CPUs.

3.4 Experimental Results

To compute fuzzy semantic scores of binding affinity for all possible protein pairs complete human proteome dataset is used. The proposed method is able to quantify the binding affinity of any two proteins within a range of [0,1]. To interpret the numerical result, a threshold cut-off μ is used. If the affinity score is greater than μ, it is considered as positive otherwise negative. A comparison of different values of μ is shown in Table 3.3 where for a given value of μ the numbers of positive and negative interactions can be seen as well as how well the classifier matches the gold sample and whole dataset. The performance evaluation is highlighted in terms of low false positive rate (FPR) and false negative rate (FNR). *precision* (*a.k.a* positive predictive value) which is the fraction of true positives to the total of true-positive (TP) and false positives (FP); *recall* (*a.k.a* sensitivity) which is the fraction of true positives to the total of TP and false negatives (FN); and AUC of receiver operating characteristic, which is obtained by plotting the true positive rate (TPR) against the FPR is also computed.

3.4.1 High Quality Interactions

Positive High-Quality Interactions (PHQI) is being identified with low FPR for $\mu \geq 0.6$. The low FPR score *i.e. type-1* error indicates that the interactions determined to be positive with $\mu \geq 0.6$ are very reliable, *i.e.*, have a very-low possibility of being negative. Thus the interactions with increasing threshold from 0.6 onward identifies PHQI. The significant performance scores are highlighted (green) in the Table 3.3.

Negative High-Quality Interaction (NHQI) is also identified with low FNR for $\mu \leq 10^{-3}$. The low FNR score *i.e. type-2* error that the interactions determined to be negative with $\mu = 10^{-3}$ are very reliable, *i.e*, have a very low probability of being positive. The proposed approach has achieved low FNR (< 0.01) on the *Gold-PPI* dataset at $\mu = 10^{-3}$ threshold. However, in *All-PPI* dataset the FNR score is 0.053 at the mentioned threshold. The thresholds with significant FNR scores are highlighted (orange) in the Table 3.3.

Table 3.3: Comparison of effects of different values of the threshold cut-off

μ	FuzzyPPI		Gold-PPI					All-PPI				
	Pos	**Neg**	**FPR**	**FNR**	**Pre**	**Rec**	**AUC**	**FPR**	**FNR**	**Pre**	**Rec**	**AUC**
10^{-5}	155101018	25275503	0.963	0.009	0.673	0.991	0.514	0.763	0.041	0.889	0.959	0.598
10^{-4}	154560195	25816326	0.962	0.01	0.674	0.99	0.514	0.762	0.042	0.889	0.958	0.598
10^{-3}	148640177	31736344	0.95	0.012	0.676	0.988	0.519	0.752	0.053	0.889	0.947	0.598
0.01	120473678	59902843	0.857	0.025	0.695	0.975	0.559	0.674	0.118	0.893	0.882	0.604
0.03	84384625	95991896	0.673	0.05	0.739	0.95	0.638	0.528	0.232	0.903	0.768	0.62
0.05	59805296	120571225	0.502	0.09	0.784	0.91	0.704	0.39	0.341	0.915	0.659	0.634
0.07	43432944	136943577	0.364	0.139	0.826	0.861	0.749	0.281	0.44	0.927	0.56	0.637
0.09	32219592	148156929	0.261	0.191	0.861	0.809	0.774	0.2	0.526	0.938	0.474	0.638
0.1	27810210	152566311	0.222	0.218	0.876	0.782	**0.78**	0.168	0.565	0.943	0.435	**0.64**
0.2	6774257	173602264	0.044	0.511	0.957	0.489	0.723	0.032	0.82	0.973	0.18	0.574
0.3	2236133	178140388	0.01	0.718	0.983	0.282	0.636	0.007	0.92	0.987	0.08	0.537
0.4	872529	179503992	0.002	0.837	0.993	0.163	0.58	0.002	0.963	0.993	0.037	0.518
0.5	429050	179947471	0.001	0.905	0.996	0.095	0.547	0.0008	0.982	0.996	0.018	0.509
0.6	182139	180194382	0	0.943	0.998	0.057	0.528	0	0.991	0.998	0.009	0.505
0.7	80483	180296038	0	0.967	0.999	0.033	0.517	0	0.995	0.999	0.005	0.503
0.8	24458	180352063	0	0.982	1	0.018	0.509	0	0.997	1	0.003	0.501
0.9	10184	180366337	0	0.993	1	0.007	0.503	0	0.999	1	0.001	0.501
0.95	2654	180373867	0	0.997	1	0.003	0.501	0	1	1	0	0.5
1	0	180376521	0	1	1	0	0.5	0	1	1	0	0.5

Pre: represents precesion.
Gold-PPI: represents Gold-PPI(Positive and Negative).
All-PPI: represents All-PPI(Positive and Negative).
Rec: represents Recall.

Table 3.4: Performance analysis of *JUPPI* and *Ding et al.* with three different datasets using Fuzzy semantic scoring threshold.

Method	**Data Type**	**AUC**	**AUPRC**	**ACCU**	**F1**	**MCC**
JUPPI [138]	PHQI-NHQI	0.996	0.996	0.973	0.973	0.946
	AP-AN	0.972	0.974	0.910	0.908	0.823
	UP-UN	0.939	0.940	0.868	0.865	0.738
Ding *et al.* [293]	PHQI-NHQI	0.924	0.927	0.836	0.832	0.672
	AP-AN	0.856	0.857	0.767	0.763	0.535
	UP-UN	0.836	0.826	0.747	0.737	0.496

3.4.2 Critical Threshold

The experimental results from Table 3.3, suggest that the PHQI and NHQI interactions can be successfully identified using proposed scoring scheme ($\mu \leq 10^{-3}$ for NHQI and $\mu \geq 0.6$ for PHQI). With the increasing values of threshold the quality of negative interactions gradually decreases. Similarly, the quality of the positive interactions drops as the threshold is decreased. To impose a critical cut-off on the scoring, different evaluation metrics such as precision, recall, and AUC have been used. The precision scores show an improving trend with the increase of threshold whereas the recall score shows a better performance at a lower threshold. AUC score became crucial to fix the threshold point, where both the precision and recall are on balance. Based on the AUC scores on both datasets (*ALL-PPI* and *Gold-PPI)*, 0.1 is considered as the critical threshold. With the threshold cut-off of 0.1, for both datasets, the highest AUC value is obtained (0.64 in *All-PPI* and 0.78 in Gold-PPI).

3.4.3 Validation *w.r.t State-of-the-art*

The proposed approach is able to provide a fuzzy semantic score between any protein pair that signifies interaction affinity between them. The biological significance of these scoring schemes is to categorize the interaction space into different significant levels such as high confidence, both positive and negative interaction, and low confidence which is uncertain.

This categorization has a significant effect on machine learning-based PPI prediction models. To establish the importance of the proposed scheme, three different datasets are selected from the interaction data pool using above described scoring cutoffs and PPI prediction performances have been evaluated using two independent

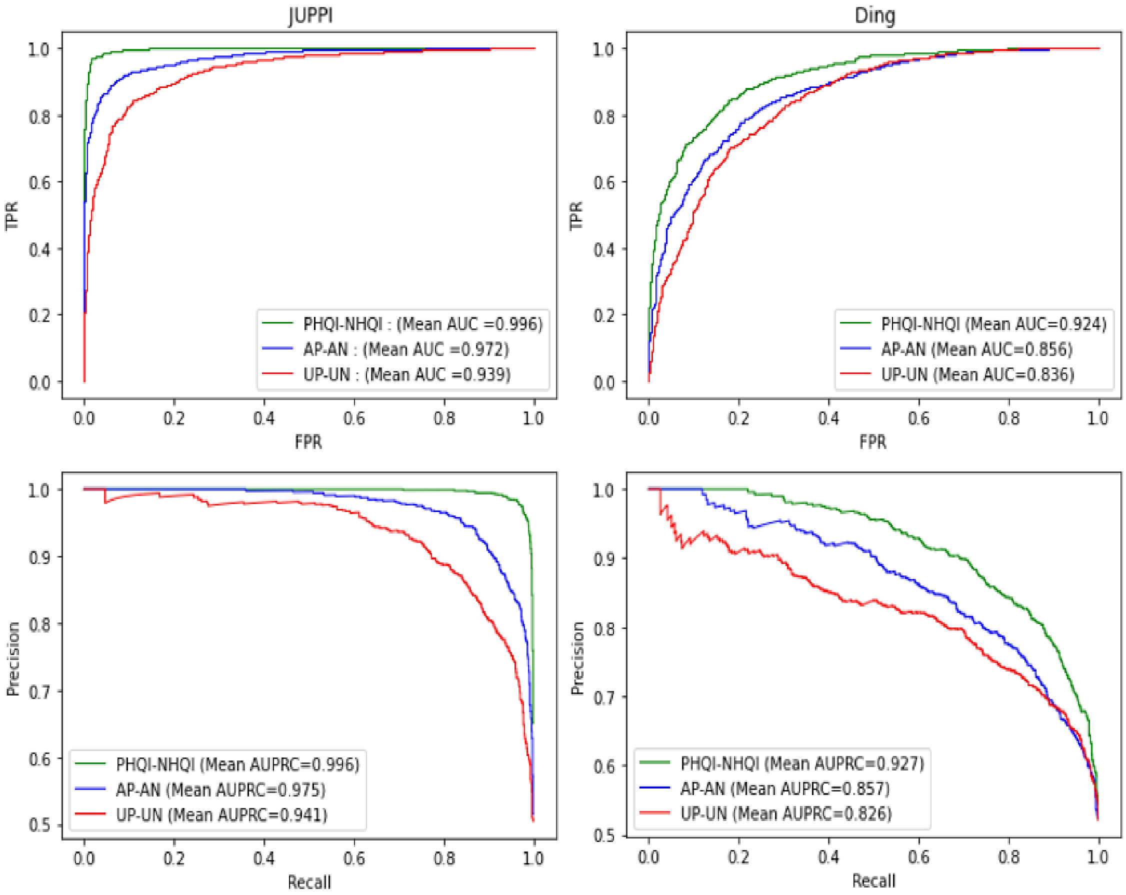


Figure 3.5: Performance changes with respect to AUC and AUPRC on three set of PPI data extracted from PHQI-NHQI, AAP-AN and UP-UN. Plots in the 1^{st} row represents AUC, and the 2^{nd} row represent AUPRC, respectively. The 1^{st} and 2^{nd}, column-wise plots show the curves from JUPPI [138] and Ding *et al.* [293].

PPI prediction methods, *JUPPI* [138], and *Ding et al.* [293]. First, the PHQI-NHQI dataset is selected from high-quality positive (with $\mu \geq 0.6$) and high-quality negative (with $\mu \leq 10^{-3}$). Second, the AP-AN dataset is selected randomly from the pool of all known positive and negative interaction data. Finally, a UP-UN dataset is selected from the uncertain region of interaction space ($\mu < 0.6$ and $\mu > 10^{-3}$) as shown in Table 3.3. In this final dataset, all ambiguous interactions are removed from both positive and negative interaction sets for clarity of the test. The performances of both methods are evaluated with five statistical metrics (AUC, AUPRC, ACCU, F1 and MCC) and scores are reported in Table 3.4. Both methods have shown a significant performance improvement on the PHQI-NHQI dataset compared to AP-AN and UP-UN. The performances on the UP-UN dataset is worse than the other two due to the uncertain positive and negative selection in training data. The AUC and AUPRC curves from three sets of interaction datasets are shown in Figure 3.5.

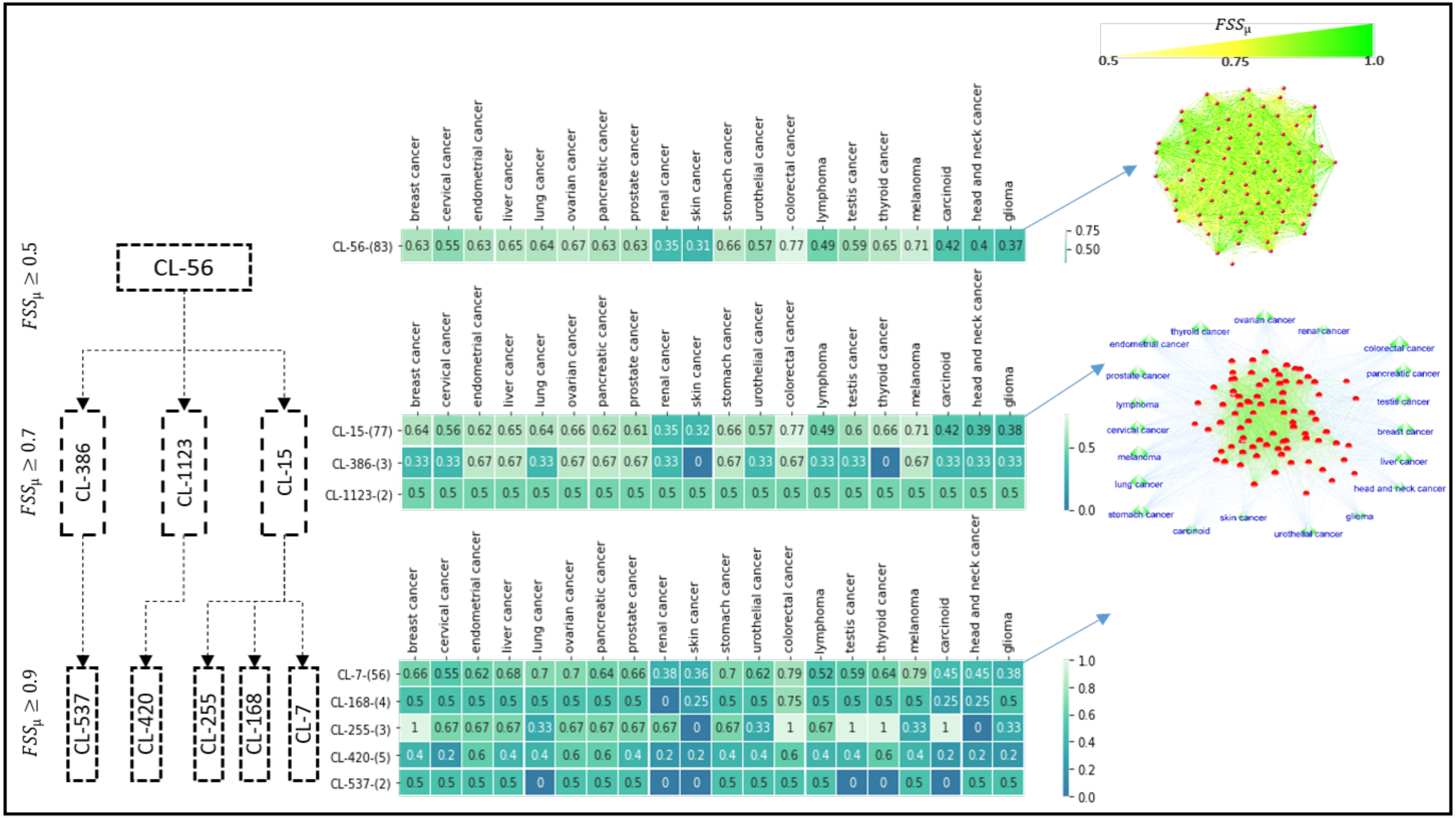


Figure 3.6: Cluster hierarchy and heatmap representation of clusters extracted at thee μ thresholds (≥ 0.5, ≥ 0.7 and ≥ 0.9) of FSS using MCL algorithm [294]. The significance of the clusters are highlighted with different cancer diseases association. Scores within the heatmap cell represents the percentage of protein association with respect to cancer disease type. Protein to cancer association (high confidence) are retrieved from Human Protein Atlas [295].

3.4.4 Proteome Level Disease Analysis

In this section, the significant sub-clusters have been extracted from a complete fuzzy semantic network of the human proteome. For graph sub-cluster retrieval, MCL clustering algorithm [294, 296] has been employed on the complete human interactome with developed fuzzy semantic scores. To evaluate the significance of the developed *Fuzzy*PPI model, these sub-clusters are being characterized with respect to the disease associated with twenty groups of human cancer disease from the Pathology Atlas of Human Protein Atlas [295]. The cluster-disease association is presented with an example cluster on three cut-off thresholds (μ) of FSS (≥ 0.5, ≥ 0.7 and ≥ 0.9). The cluster hierarchy and their association are shown in Figure 3.6.

3.5 Discussion

In this chapter, a new fuzzy affinity-based scoring scheme has been presented for the prediction of interaction affinity pairs of human proteins. The work is built upon the GO associations of respective proteins and the underlying GO graphs for MF, CC, and BP sub-graphs. A graph clustering approach has been used to identify representative cluster centers in a graph, and the ancestor-descendent relationships between two nodes have been utilized to design the fuzzy affinity functions. One of the major limitations of the work is that it depends heavily on the GO annotations of the respective proteins. In cases, where GO annotations are not available, the interaction affinity cannot be predicted. However, among $20,350$ reviewed human proteins, only $1,356$ proteins did not have a matching GO annotation and had to be excluded from the study, which allowed to successfully estimate interaction affinity between ~ 180 million of protein pairs. Also, the GO annotation and underlying GO graphs get updated periodically. Minor perturbations in the GO network may lead to changes in the PPI affinity scores. Therefore, it is also proposed to update the web server periodically with new releases of the GO annotations.

Computation of such a large number of interaction estimations would not have been possible without the use of Spark-based parallelization. As for comparison, the work by Dutta *et al.* [128] used only $4,726$ interactions from ~ 2000 human proteins. With a standard desktop computer and sequential implementation, the graph clustering and affinity assessment score estimation algorithms take days/weeks to complete. In contrast, ~ 180 million interaction estimations took less than a week in a Spark-based parallel cluster setup. It has been planned to extend our work with both reviewed and unreviewed proteins and also for multi-organism PPI prediction problems, leading to more than a trillion interactions.

Finally, one of the major objectives of the work is to qualitatively categorize the protein interactions, based on the affinity scores. It may be observed that the *Fuzzy*PPI may be useful for identifying high-quality positive interactions with very low FPR, as well as high-quality negative data selection with very low FPR. The work has also been compared with the *state-of-the-art* in the domain and its effectiveness has been validated.

In this chapter, a new method has been depicted that can compute the interaction fuzzy affinity between any two protein pairs. The development of such an effective scheme to compute fuzzy interaction affinity at the organism level eventually helps to construct a fuzzy semantic network. This fuzzy semantic network could be a key node to assess different cellular and molecular mechanisms, disease analysis, host pathogenic relationships, drug target improvement, etc. In the following chapter, an *in-silico* host-pathogen network model has been developed for human-nCoV PIN which would help in understanding the disease transmission mechanism by identifying the potential spreader proteins.

Chapter 4

Computational Modeling of Host-Pathogen PPIN

4.1 Background

Host-pathogen PPINs are significant for understanding the mechanism of transmission of infection, which is essential for developing new and more effective therapeutics, including rational drug design. Progression of infection and disease results due to the interaction of proteins in between pathogen and host. Pathogen plays an active role in spreading infection. Pathogen and host PPIN permit pathogenic microorganisms to utilize host capabilities by manipulating the host mechanisms to abscond from the host's immune responses [65, 135, 136]. Detection of target proteins through the analysis of pathogen and host PPIN is the central point of research [56, 137, 138]. Topologically significant proteins having a higher degree of interaction are generally found to be important drug targets. However, proteins with fewer interactions or topologically not substantial may be involved in the mechanism of infection because of some biological pathway relevance. Clinically validated human-nCoV PPI in the current literature gives the motivation to develop a new computational model for the human-nCoV PPI network. The proteins are subsequently validated which involves the host-pathogen interactions with respect to potential Food and Drug Administration (FDA)-approved drugs for COVID-19 treatment.

Coronavirus belongs to the family of Coronavirae. Besides affecting human beings, it also infects birds and mammals. SARS-CoV-2 replicates the host cell's genome by interacting with the host proteins. Due to this fact, the identification of virus and host PPIs could be beneficial in understanding the disease transmission behavior of the virus as well as in potential COVID-19 drug identification. The International Committee on Taxonomy of Viruses (ICTV) has declared that nCoV is highly genetically similar to the SARS-CoV epidemic in 2003 having $\sim 89\%$ genetic similarity. Though the common symptoms of the coronavirus are common cold, cough, etc., it is accompanied by severe acute, chronic respiratory disease and multiple organ failure leading to human death. SARS-CoV and MERS-CoV were the two major outbreaks in 2003 and 2012 before SARS-CoV-2. The source of origin of SARS-CoV was located in Southern China. Its fatality rate was within 14%-15% [297], due to which 774 people lost their lives among 8804 affected cases. Saudi Arabia was marked as the base for the commencement of MERS-CoV. 858 persons among 2494 infected cases were defeated in their battle against the MERS-CoV virus. Therefore, in comparison to SARS-CoV, it resulted in a mortality rate that was significantly greater, at 34.4% [298].

COVID-19 evolved in the Chinese city of Wuhan (Hubei province) [299]. The first case of human species affected by nCoV was observed on 31st December 2019 [300]. Soon it expands its adverse effect on almost all nations within a brief period [301]. The World Health Organisation (WHO) recognises that the widespread catastrophic outbreak of nCoV is primarily the result of widespread community transmission, and on January 30, 2020, they announce the declaration of a global health emergency. After proper assessment, WHO presumes its fatality rate to be 4% which urges global researchers to work together to discover an appropriate treatment for this pandemic [302,303].

All three epidemic creators SARS, MERS, and SARS-CoV-2, are biologically included in the genus beta coronavirus under the Coronaviridae. Both structural and non-structural proteins are involved in the formation of SARS-CoV-2. After entering the human body, the structural proteins, such as the envelope protein, membrane protein, nucleocapsid protein, and spike protein, connect with the receptors and play a vital part in the transmission of the disease [304]. So, there is an urgent need to understand and analyze the mechanism of disease transmission of this new virus.

In this research work, PPIN are the most significant attribute in studying the disease propagation mechanism from SARS-CoV-2 to humans. It plays a crucial role in identifying essential proteins [137,157,158,305,306] responsible for various diseases. They are also significant in identifying protein functions [152,153,180,307]. According to Lotem *et al.* [308], though human PPIN is constantly expanding, very little information is available about the human PPIN, which gets generated in disease conditions. With the enhancement in the availability of human PPIN data, the primary focus of research has been shifted from the basic understanding of the PPIN to the study of the PPIN underlying various kinds of human disease [309]. According to the work of Ideker *et al.* [310], PPIN study mainly falls under four categories: 1) Identification of human disease genes based on network analysis, 2) Implication of additional genes involved in the disease by using protein networks, 3) Identification of protein subnetworks involved in diseases and 4) Classification of case-control studies based on protein PPIN.

It has been reported that SARS-CoV has $\sim$ 89% [311,312] genetic similarities with nCoV. SARS-CoV-Human protein-PPIN has also been studied widely and is available in the literature [313–315]. Recently, a computational model has been developed to identify potential spreader proteins in a Human-SARS-CoV interaction network using the SIS model [158]. In addition, sequence information of 29 nCoV proteins has been released [33] GO information of 14 of the nCoV proteins are available [33, 316]. A method to predict interaction affinity between proteins from the available GO graph has been recently developed [128]. Assessment of interaction affinity between nCoV proteins with potential Human target/bait proteins, which are susceptible to SARS-CoV infection, has been done. Fuzzy affinity thresholding is done to detect high-quality human-nCoV PPIN. The selected human proteins are considered level-1 human spreader nodes of nCoV. level-2 spreader nodes in human-nCoV PPIN are detected using the spreadability index and validated by SIS [158, 317] model. The developed model is validated for the target proteins of the potential FDA-approved drugs for COVID-19 treatment [227]

4.2 Dataset Description

Human-SARS-CoV PPIN serves as a baseline for the proposed model. The potential level-1 and level-2 human spreaders of SARS-CoV become the possible candidate set for selecting level-1 human spreaders of SARS-CoV-2. Various datasets have been curated for this purpose which have been outlined below:

4.2.1 Human PPIN

The dataset [318, 319] consists of all possible interactions between human proteins experimentally documented in humans. Human proteins are represented as nodes, while edges represent the physical interactions between proteins. It is a collection of 21557 nodes and includes 342353 edges/interactions.

4.2.2 SARS-CoV PPIN

PPIs caused by SARS-CoV are included in the dataset. There are 7 distinct proteins and 17 interacting helices present. Since densely coupled proteins play a more direct effect in infection spread than solitary ones, only these are taken into account [313].

4.2.3 SARS-CoV - Human Protein-PPIN

The dataset consists of 118 instances of interactions between SARS-CoV and human hosts. The purpose of this method is to retrieve the primary human interactions associated with SARS-CoV [313].

4.2.4 SARS-CoV-2 Proteins

This data is collected from the pre-released dataset of available SARS-CoV-2 proteins from UniProtKB [33, 320], which includes 14 reviewed SARS-CoV-2 proteins.

4.2.5 GO Graph and Protein-GO annotations

GO graph types *i.e.* CC, MF, and BP are collected from GO Consortium [316]. In addition, the protein to GO-annotation map is retrieved from the UniProtKB database.

4.2.6 Potential COVID-19 FDA-approved drugs

Six potential FDA-approved: Lopinavir [321], Ritonavir [322], Azithromycin [323], Remdesivir [324–326], Favipiravir [327, 328], and Darunavir [329] have been identified from the DrugBank [330] published white paper [227] which have been used for validating the proposed model.

4.3 Methodology

The developed computational model for human-nCoV PPIN that has been proposed here consists of two crucial methodologies 1) validation of the SIS model in addition to the identification of spreader nodes based on the spreadability index and 2) *fuzzy*PPI model. First, the SIS model identifies spreader nodes of SARS-CoV proteins which are the candidate set of nCoV interactors. Then, the *fuzzy*PPI model is applied to extract the human-nCoV interactions, and finally, nCoV spreaders are identified using the SIS model.

4.3.1 Identification of Spreader nodes by Spreadability Index Along With the Validation by SIS Model

In human-nCoV PPIN, the former acts as a pathogen/bait while the host, the human, acts as 'Prey'. The transmission of infection starts when a pathogen enters a host body and infects its protein, affecting its directly or indirectly connected neighbor proteins. Considering this method of transmission, PPIN of humans and SARS-CoV are considered to detect spreader nodes. Spreader nodes are those nodes/proteins that transmit the disease fast among their neighbors. However, not all the nodes in a PPIN are spreaders. So, proper detection of spreader nodes is crucial. Spreader nodes are found by the utilization of the spreadability index, a metric that quantifies the transmission capacity of a given node or protein. Moreover, the compactness of PPINs and their capacity for transferability are assessed by centrality analysis. Nodes exhibiting high centrality values are commonly regarded as spreader nodes or the most pivotal nodes within a network.

The spreadability index [158] is one of the centrality-based measures that combines three major topological neighborhood-based features of a network. They are

- Node weight (N_w) [331]

- Edge ratio (E_r) [332]
- Neighborhood density (N_D) [332]

Nodes having a high spreadability index are considered spreader nodes. The spreader nodes thus identified are also validated by the SIS model [317]. The SIS model is implemented to design the SARS-CoV and SARS-CoV-2 outbreak into a disease model consisting of proteins based on their present infection status. A protein can be in either of the three states:

- S: Susceptible, which means that every protein is initially susceptible though not yet infected but at risk of getting infected by the disease.
- I: Infected, which means that the disease already infects the protein.
- S: Susceptible, which means proteins again become susceptible after getting recovered from the infected state.

This model is implemented to generate the overall infection capability of a node after a certain range of iterations. Thus the sum of the infection capability of the top selected spreader nodes is computed by this model, which is compared against the sum obtained for the selected top critical nodes by other existing centrality measures like:

Betweenness Centrality (BC) [181] is one of the ways of measuring a node's impact on the transmission of information between every pair of nodes in a graph, considering that this transmission is always executed over the shortest path between them. Mathematically, it is defined as:

$$C_B^{(u)} = \sum_{s \neq u \neq t} \frac{\rho(s,u,t)}{\rho(s,t)} \tag{4.1}$$

where $\rho(s,t)$ is the total number of shortest paths from node s to node t, and $\rho(s,u,t)$ is the number of those paths that pass through u.

Closeness Centrality (CC) [159] is a procedure for detecting nodes that transmit information within a network efficiently. Nodes with high closeness centrality values are considered to have the shortest distance to all available nodes in the network. It can be mathematically expressed as:

$$C_C^{(u)} = \frac{|N_u| - 1}{\sum_{v \in V} dist(u,v)} \tag{4.2}$$

Where, $|N_u|$ denotes the number of neighbors of node u and dist (u,v) is the distance of the shortest path from node u to node v.

Degree Centrality (DC) [333] is considered the simplest among the available centrality measures that only count the degree of a node, i.e., the number of directly connected neighbors. Nodes having a high degree are said to be the highly connected module of the network. It is defined as:

$$C_D^{(u)} = |N_u| \tag{4.3}$$

Where, $|N_u|$ denotes the number of neighbors of node u.

Local Average Centrality (LAC) [302] of a node represents how close its neighborhood proteins are. It is defined to be the local metric to compute the essentiality of the node for transmission ability by considering its modular nature, the mathematical model of which is highlighted as:

$$LAC(u) = \frac{\sum_{w \in N_u} deg_{C_u}^{w}}{|N_u|} \tag{4.4}$$

Table 4.1: Computation of spreadability index of Figure 4.1 and validation of selected top 5 spreader nodes by the SIS

Rank	Proteins	$E_{out}^{S_i}$	$E_{in}^{S_i}$	E_r	N_D	N_w	S_i	Sum of SIS infection rate of top 5 nodes
1	Node 3	6	3	1.75	6.94	2.83	14.99	1.19
2	Node 9	5	4	1.20	7.07	3.00	11.48	
3	Node 6	5	2	2.00	3.93	2.60	10.46	
4	Node 8	6	2	2.33	2.27	3.25	8.55	
5	Node 1	5	4	1.20	4.21	3.40	8.45	

N_D: Neighbourhood Diversity

The proposed method for selecting spreader nodes in SARS-CoV PPIN [158] has performed better than the other existing *state-of-the-art* like BC [181], CC [159], DC [154] and LAC [156]. The detailed comparison and results are given below in Table 4.2, Table 4.4 to Table 4.7 with reference to the Figure 4.3 . A synthetic PPIN is considered in Figure 4.1 to demonstrate the entire methodology of the spreadability index (*see* Table 4.1). In addition, computational analysis of the spreadability index of the proposed model with one of the other methodologies LAC [334] has been highlighted in Table 4.3. $E_{out}^{S_i}$ is the total number of edges that are outgoing from the ego network S_i [158]. Whereas, $E_{in}^{S_i}$ is denoted as the total number of interconnections in

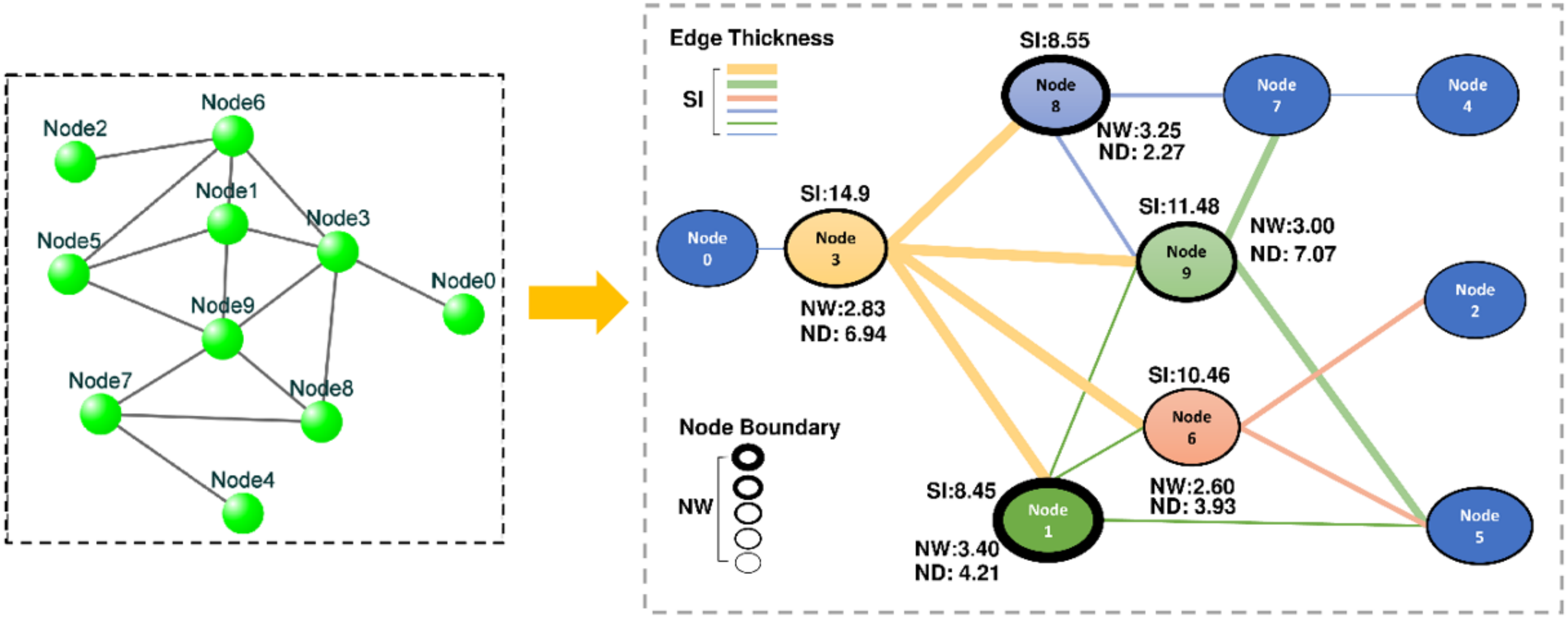


Figure 4.1: Synthetic PPIN: The network consists of 10 nodes and 25 edges. N_D, N_w, and Spreadability Index of the top 5 nodes have been highlighted. While the thickness of the edges highlights the rank according to SI, the thickness of the boundary of the nodes highlights the rank according to N_w.

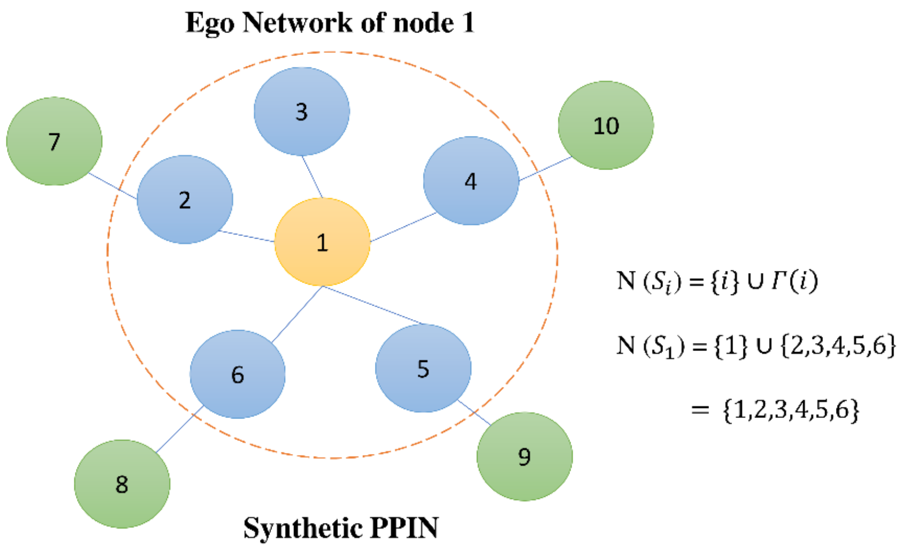


Figure 4.2: Ego Network of node 1 in a synthetic PPIN is highlighted by the dotted circle.

the neighbor of node i [332]. $E_{out}^{S_3}$ of node 3 is 6 while $E_{in}^{S_3}$ of node 3 is 3, which highlights that node 3 has the highest transmission ability from its ego network to outside when compared to other nodes. Node 3 also has the highest spreadability index. But LAC failed to rank node 3 in the first position. The same scenario can be observed for some other nodes in the synthetic network too. Besides SIS validation result shows that the selected top-ranked spreader nodes in this proposed model have the highest infection capability compared to the other ranked nodes. Ego network can be defined as:

Ego network [332] of node i (S_i) is defined as the grouping of node i itself along with its corresponding level-1 neighbors and interconnections. N(S_i) consists of the set of nodes that belong to the ego network, S_i i.e., $i \cup \Gamma(i)$ where $i \cup \Gamma(i)$ denotes node i*'s* neighbors. It is used in the identification of spreader proteins in SARS-CoV-Human PPIN in [158]. The figure of Ego network is given in Figure 4.2.

4.3.2 *Fuzzy*PPI Model for Potential SARS-CoV-2 - Human Interaction Identification

The binding affinity between any two interacting proteins can be estimated by combining the semantic similarity scores of the GO terms associated with the proteins [56, 128, 335–337]. A greater number of semantically similar GO annotations between any protein pair indicates higher interaction affinity. The *fuzzy*PPI model is a hybrid approach [128] that utilizes both the topological [338] features of the GO graph and information contents [273, 274, 337]of the GO terms.

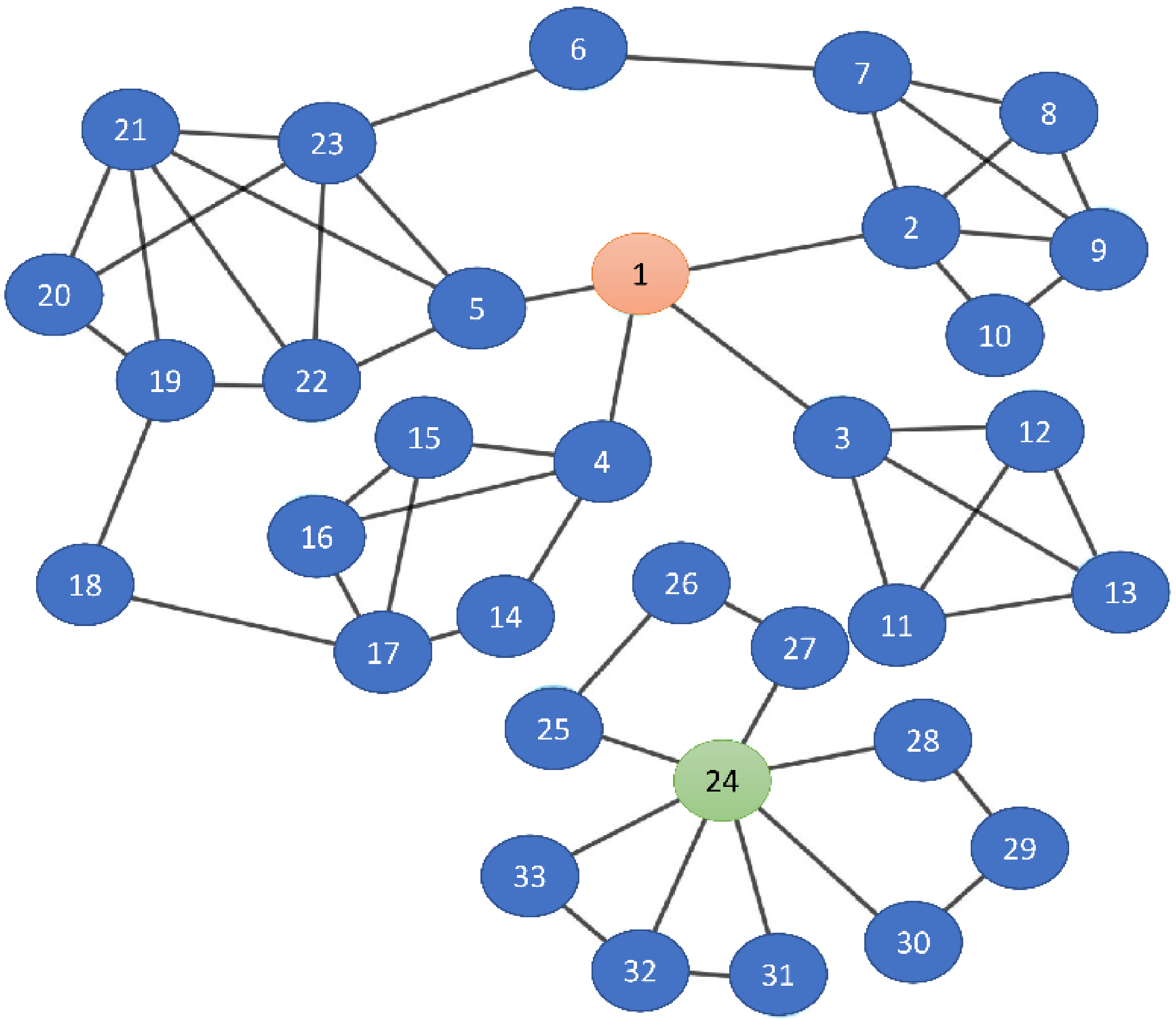


Figure 4.3: Synthetic protein-PPIN: The PPIN consists of 33 nodes and 53 edges. Nodes 1, 24 are the essential spreaders. Node 1 connects the four densely connected modules of the PPIN, which turns this node to stand in the first position having the highest spreadability index. Node 24 holds the second position for the spreadability index. Node 24 is one of the most densely connected modules itself despite getting isolated from the main PPIN module of node 1

GO is organized in three independent DAGs are MF, BP, CC [316]. The nodes in each GO graph represent GO terms, and the edges represent different hierarchical relationships. In the proposed work, the two most essential relations, *'is_a'* and *'part_of'* have been used for GO relations [339].

The semantic similarity between any two proteins is estimated by considering the similarities between all pairs of annotating GO terms belonging to a particular ontological graph. The similarity of a GO term pair is determined by considering specific topological properties of the GO graph and the average IC [340] of the DCAs [335,336] of the GO terms as proposed in [128]. *Fuzzy*PPI first relies on a fuzzy clustering of the GO graph where the selection of GO terms as cluster center is based on the level of association of that GO term in the GO graph. Then, the cluster centers are selected based on the proportion measure of GO terms. The proportion measure for any GO term t is computed as:

$$PrM\,(t) = \frac{|An\,(t)| + |Dn\,(t)|}{|N_O|} \tag{4.5}$$

where, *(t)*, *Dn(t)* represents the ascendant and descendant of term t and N_0 is the total number of GO terms in ontology. A higher value of the proportion measure *(PrM(t))* signifies higher coverage of ascendants and descendants associated with the specific node. Finally, the GO terms for which this *(PrM(t))* is above a predefined threshold are selected as cluster centers. In this work, the cluster centers are chosen based on the threshold values as suggested in [56, 128]

After selecting the cluster centers, the degree of membership of a GO term to each of the selected cluster centers is calculated using its respective shortest path lengths to the corresponding cluster centers. The membership of the GO term to a cluster decreases with an increase in its shortest path length to the cluster center. The membership function is defined as:

$$MmFc(t) = e^{-\ \frac{-(x-c_i)^2}{2\ k^2}} \tag{4.6}$$

where, c_i is i *-th* center and k is the width of membership function, and x is the shortest path length from t to c_i. The difference *D(*t_i*,*t_j*)* in membership values between the GO pair t_i and t_j for each cluster center, is computed to find the weight parameter. The weight parameter is defined as:

$$Wt(t_i, t_j) = 1 - maxD(t_i, t_j) \tag{4.7}$$

This weight value determines how different two GO terms can be with respect to the cluster centers. Next, the SIC is computed using the average IC [340] of the *DCA* of the GO term pair (t_i,t_j) for all three GO graphs. The SIC is defined as:

$$SIC(t_i, t_j) = \frac{\sum_{a \subseteq DCA(t_i,t_j)} IC(a)}{|DCA(t_i, t_j)|} \tag{4.8}$$

where *DCA(t_i,t_j)* represents the disjunctive common ancestors of GO-term t_i and t_j. The semantic similarity,*sim*, between the GO term pair t_i and t_j is defined as:

$$sim(t_i, t_j) = Wt(t_i, t_j) \times SIC(t_i, t_j) \tag{4.9}$$

The semantic similarity of protein pair P_i,P_j for each GO-type *i.e.* CC, MF, and BP, is estimated by utilizing the maximum similarity of all possible GO pairs from the annotations of proteins P_i and P_j for each type of GO. The interaction affinity of protein pair (P_i,P_j) is defined as the average of CC, MF, and BP-based semantic similarity.

This work uses the available ontological information to calculate the fuzzy interaction affinity score between the protein pairs of SARS-CoV-2 and spreader human proteins (*see* Figure 4.4). Here, the SARS-COV's level-1 and level-2 spreader proteins are employed as the primary target for the proposed *fuzzy*PPI model for interaction affinity computation. A bipartite relation of GO pairs is primarily generated from each pair of proteins for each type of GO annotations *i.e.* CC, MF, and BP independently (Figure 4.4A) To reduce the computational overhead and time, semantic similarity scores are previously computed between all GO pairs belonging to a particular GO type using equation 4.9 [128]. The semantic similarity is computed by exploring the topological properties of the GO sub-graph. For each type of GO sub-graph, a different set of cluster center nodes, which are GO terms, are identified based on proportion measure (*see* Equation 4.5) that rely on the annotation score and GO relationship graph hierarchy. The GO semantic similarity is estimated with a distance-based measure between the target GO pair by exploring the membership score (*see* Equation 4.6 and 4.7) and values (*see* Equation 4.8) compared to respective cluster centers of each GO sub-graphs (Figure 4.4B). For each GO type, the max of all possible scores of the bipartite links in a particular GO sub-graph is considered the final SS of that type of GO.

Similarly, all three different scores are evaluated and averaged to find the interaction affinity for the annotated protein pair. Then, the fuzzy score of interaction affinity is computed by normalizing the interaction affinity using max-min normalization. Finally, with high specificity threshold (*see* Figure 4.5), high-quality interactions are extracted for human-SARS-CoV-2 PPIN which involves 78 interactions involving

Table 4.2: Computation of spreadability index of Figure 4.3 and validation of selected top 5 spreader nodes by the SIS

Rank	Proteins	$E_{out}^{S_i}$	$E_{in}^{S_i}$	E_r	N_D	N_w	S_i	Sum of SIS infection rate of top 10 nodes
1	1	13	0	14.0	5.19	3.40	76.15	2.46
2	24	4	2	1.66	12.5	1.87	22.70	
3	4	6	1	3.50	3.63	2.40	15.11	
4	5	8	3	2.25	4.39	3.60	13.48	
5	19	6	2	2.33	3.8	2.80	11.66	
6	23	5	4	1.20	6.58	3.00	10.89	
7	17	4	1	2.50	3.33	2.00	10.33	
8	6	7	0	8.00	0.87	3.00	10.00	
9	2	4	4	1.00	6.84	2.83	9.68	
10	22	6	4	1.40	3.88	3.60	9.03	
11	25	7	0	8.00	0.71	3.00	8.71	−−
12	27	7	0	8.00	0.71	3.00	8.71	
13	28	7	0	8.00	0.71	3.00	8.71	
14	30	7	0	8.00	0.71	3.00	8.71	
15	18	6	0	7.00	0.85	2.66	8.66	
16	20	7	2	2.66	1.78	3.50	8.26	
17	7	4	3	1.25	4.15	2.80	7.98	
18	21	3	6	0.57	6.66	3.33	7.13	
19	3	3	3	1.00	4.00	2.60	6.60	
20	16	4	2	1.66	2.06	2.75	6.19	
21	15	4	2	1.66	2.06	2.75	6.19	
22	31	6	1	3.50	0.75	3.33	5.95	
23	33	6	1	3.50	0.75	3.33	5.95	
24	32	4	2	1.66	1.75	2.75	5.66	
25	8	4	3	1.25	1.88	3.25	5.60	
26	14	6	0	7.00	0.40	2.66	5.46	
27	9	2	4	0.60	3.64	2.80	4.98	
28	10	5	1	3.00	0.50	3.00	4.50	
29	13	1	3	0.50	1.70	2.50	3.35	
30	11	1	3	0.50	1.70	2.50	3.35	
31	12	1	3	0.50	1.70	2.50	3.35	
32	29	2	0	3.00	0.00	1.33	1.33	
33	26	2	0	3.00	0.00	1.33	1.33	

Table 4.3: Computation of the LAC of the synthetic network of Figure 4.1 and validation of selected top 5 spreader nodes by the SIS model.

Rank	Proteins	LAC	Sum of SIS infection rate of top 5 nodes
1	Node 1	2	0.86
2	Node 9	1.6	
3	Node 5	1.33	
4	Node 8	1.33	
5	Node 3	1.2	

37 human level-1 spreaders proteins.

4.4 Experimental Results

The proposed computational model of human-nCoV PPIN contains high-quality interactions and proteins identified by Fuzzy affinity thresholding and spreadability index validated by the SIS model. The sources of input and the generated results always play a crucial role in any computational model, which is also true for the proposed model.

4.4.1 Spreader Nodes Selection in Human-SARS CoV Interaction Network Using Spreadability Index

SARS-CoV - human PPIN, up to level-2, is formed by the combination of SARS-CoV - human and Human-Human PPIN datasets. SARS-CoV - human dataset generates the direct level-1 human interactions of SARS-CoV, while the human-human PPIN dataset is used to fetch the corresponding level-2 human interactions. Potential spreader nodes are identified using the spreadability index validated by the SIS model [158]. The entire process of the detection of spreader nodes in SARS-CoV - human PPIN is depicted in four steps in Figure 4.6

- 6 Spreader nodes in SARS-CoV PPIN are detected by the spreadability index.
- Corresponding level-1 human proteins of the spreader nodes in SARS-CoV PPIN are identified.
- 24 Spreader nodes in level-1 human proteins of the spreader nodes in SARS-CoV PPIN are detected.
- The same process is repeated, and 9 spreader nodes in level-2 human proteins of the spreader nodes in SARS-CoV PPIN are identified.

Table 4.4: Computation of CC of Figure 4.3 and computation of spreadability rate of selected top 10 spreader nodes by SIS model [158]

Rank	**Proteins**	**Closeness Centrality**	**Sum of SIS spreadability rate of top 10 nodes**
1	1	0.085	1.94
2	5	0.083	
3	2	0.082	
4	4	0.082	
5	23	0.081	
6	3	0.081	
7	21	0.081	
8	22	0.081	
9	7	0.08	
10	15	0.08	
11	16	0.08	– –
12	19	0.079	
13	14	0.079801	
14	9	0.079602	
15	20	0.079602	
16	6	0.079602	
17	8	0.079404	
18	17	0.078818	
19	10	0.078624	
20	11	0.078049	
21	12	0.078049	
22	13	0.078049	
23	18	0.07767	
24	24	0.041558	
25	32	0.041237	
26	28	0.041237	
27	30	0.041237	
28	25	0.041237	
29	27	0.041237	
30	31	0.041184	
31	33	0.041184	
32	29	0.040921	
33	26	0.040921	

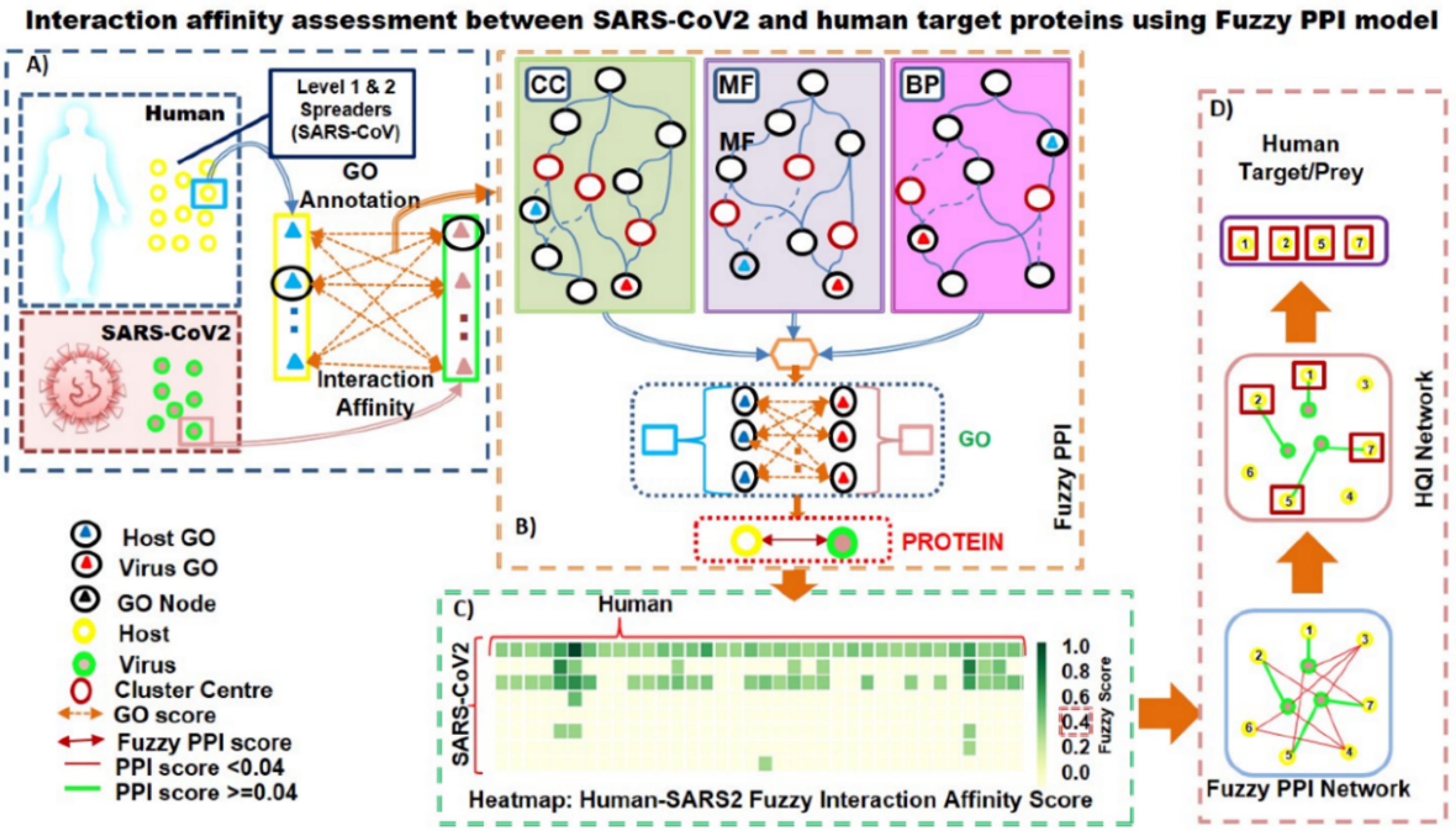


Figure 4.4: Schematic diagram of *Fuzzy*PPI model. **A)** The *fuzzy*PPI model finds the interaction affinity between the SARS-CoV-2 and Human proteins (Level-1 and Level-2 spreader of SARS-CoV) using ontological gene information. **B)** All GO pair-wise interaction affinities are assessed from three independent GO-relationship graphs CC, MF, and BP. The fuzzy interaction affinity of a protein pair is computed from all three pair-wise scores of all GO-pair affinities. **C)** Heatmap representation of *fuzzy*PPI score. **D)** Network representation of Human and SARS-CoV-2 proteins with 0.2 onward thresholds of *fuzzy*PPI score at high specificity. Finally, high-quality interactions are extracted to retrieve the potential human prey for SARS-CoV-2 at the 0.4 threshold.

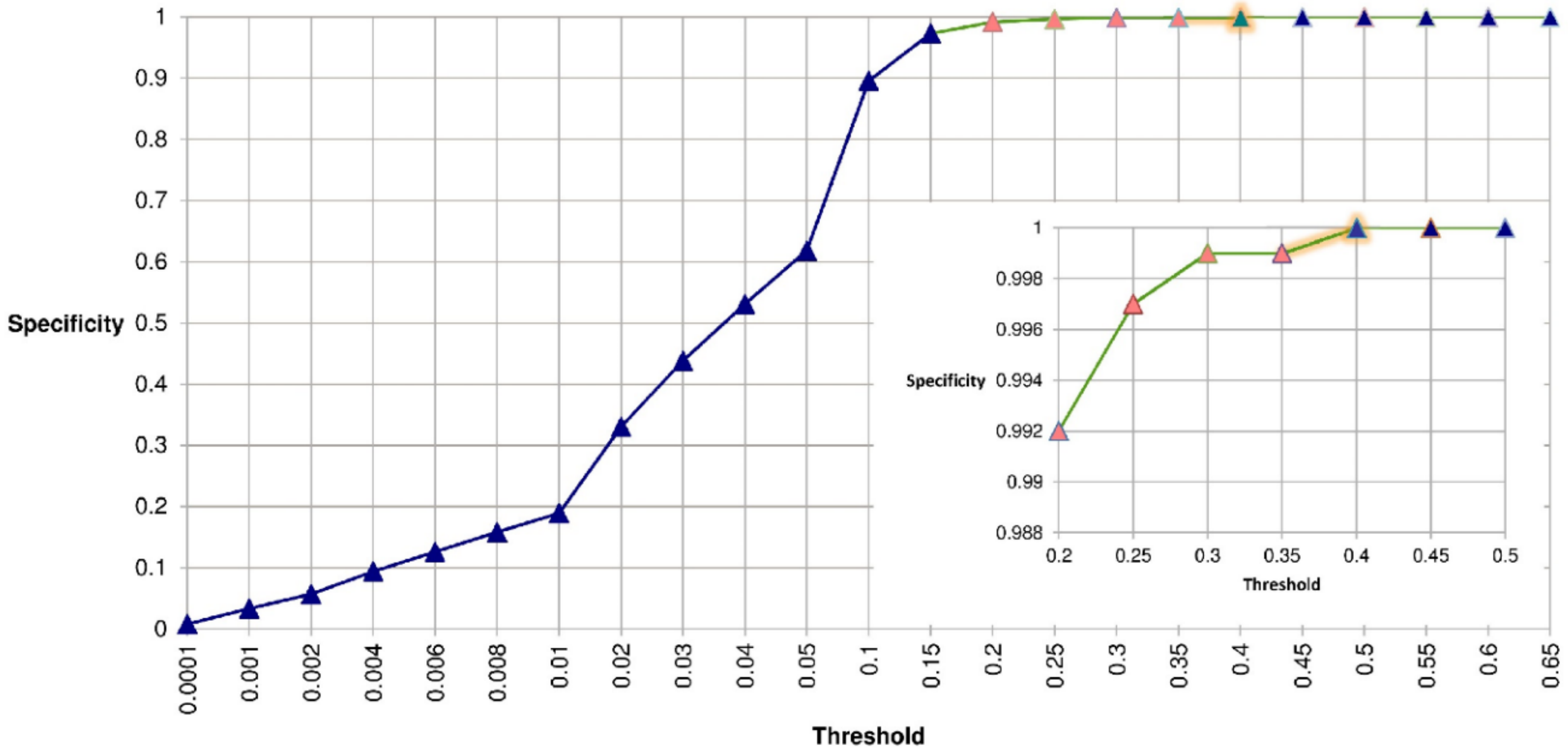


Figure 4.5: Specificity at different threshold (x-axis) of binding affinity obtained from *fuzzy*PPI model for complete human proteome interaction network. At 0.2 onward threshold, it produces high specificity with respect to benchmark positive and negative interaction data. High-quality interactions are extracted at a 0.4 threshold with $\sim$ 99.9% specificity.

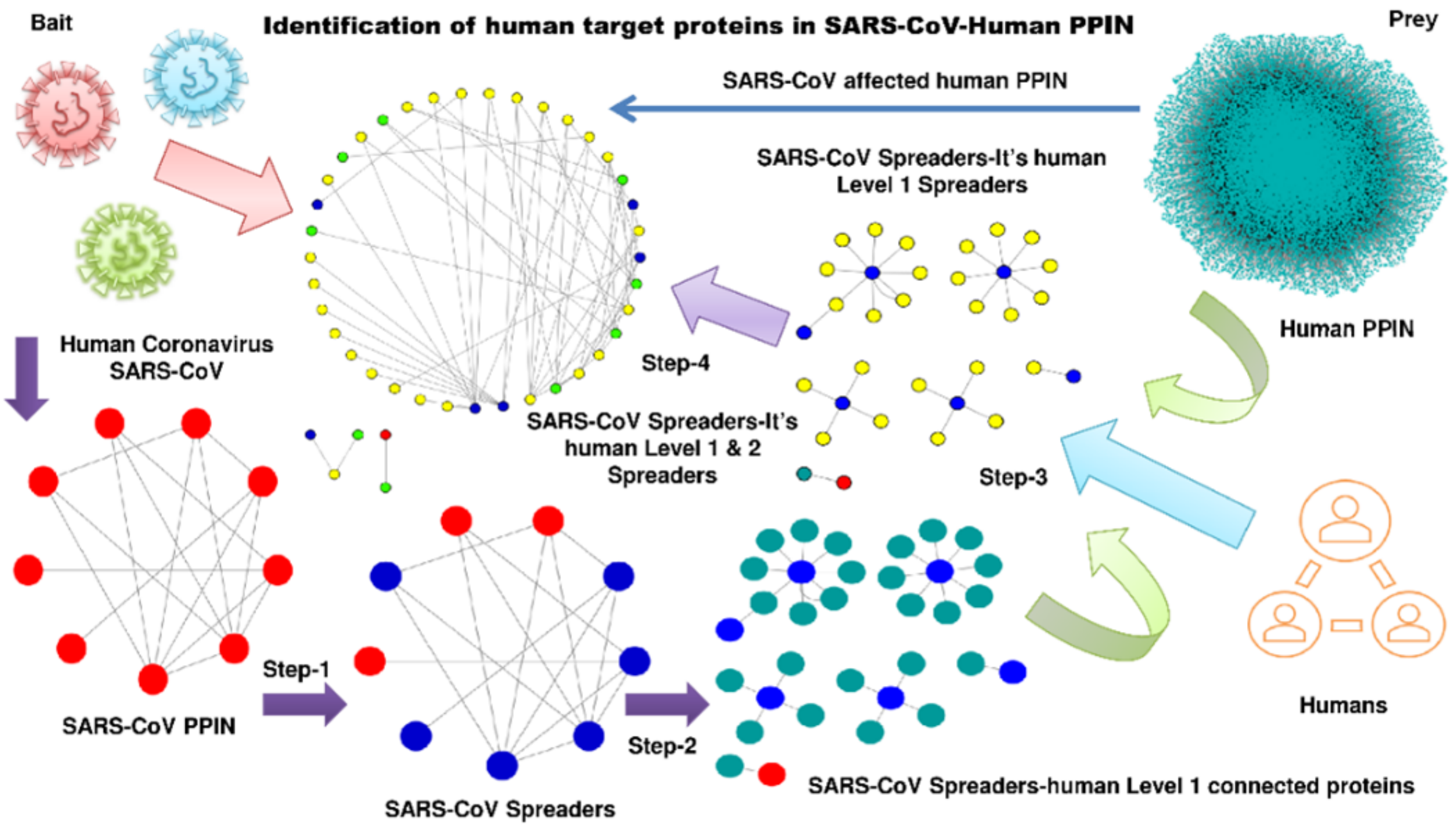

Bait
Identification of human target proteins in SARS-CoV-Human PPIN
Prey
SARS-CoV affected human PPIN
SARS-CoV Spreaders-It's human Level 1 Spreaders
Human PPIN
Human Coronavirus SARS-CoV
Step-4
SARS-CoV Spreaders-It's human Level 1 & 2 Spreaders
Step-3
Humans
Step-1
SARS-CoV PPIN
Step-2
SARS-CoV Spreaders
SARS-CoV Spreaders-human Level 1 connected proteins

Figure 4.6: A computational model for the selection of spreader nodes in Human-SARS CoV PPIN by spreadability index. Red-coloured nodes represent SARS-CoV proteins, while blue-colored nodes are the selected spreader nodes in it. Deep green colored nodes represent level-1 human-connected proteins with SARS-CoV proteins, while yellow-coloured nodes represent the selected human spreaders. Light green colored nodes represent level-2 human spreaders of SARS-CoV

Table 4.5: Computation of BC of Figure 4.3 and computation of spreadability rate of selected top 10 spreader nodes by SIS model [158]

Rank	Proteins	Betweeness Centrality	Sum of SIS spreadability rate of top 10 nodes
1	1	269.1	2.2
2	2	117.93	
3	4	117.1	
4	3	114	
5	5	108	
6	24	57	
7	23	56.4	
8	19	45.56	
9	17	39.1	
10	7	36.9	
11	6	32.9	- -
12	18	32	
13	21	29.36	
14	22	20.53	
15	16	12.1	
16	15	12.1	
17	14	12.1	
18	28	7	
19	30	7	
20	25	7	
21	27	7	
22	20	6.63	
23	9	4.16	
24	32	1	
25	29	1	
26	26	1	
27	8	0	
28	11	0	
29	12	0	
30	13	0	
31	10	0	
32	31	0	
33	33	0	

4.4.2 Identification of the human-nCoV proteins interactions using *fuzzy*PPI model

The GO information can be helpful in inferring the binding affinity of any pair of interacting proteins using three different types of GO hierarchical relationship graphs,

Table 4.6: Computation of LAC of Figure 4.3 and computation of spreadability rate of selected top 10 spreader nodes by SIS model [158]

Rank	**Proteins**	**Local Average Centrality**	**Sum of SIS spreadability rate of top 10 nodes**
1	21	2.4	2.19
2	9	2	
3	22	2	
4	8	2	
5	11	2	
6	12	2	
7	13	2	
8	2	1.6	
9	23	1.6	
10	7	1.5	
11	3	1.5	– –
12	5	1.5	
13	16	1.33	
14	15	1.33	
15	20	1.33	
16	32	1.33	
17	19	1	
18	10	1	
19	31	1	
20	33	1	
21	24	0.57	
22	4	0.5	
23	17	0.5	
24	1	0	
25	14	0	
26	18	0	
27	6	0	
28	28	0	
29	29	0	
30	30	0	
31	25	0	
32	26	0	
33	27	0	

namely CC, MF, BP [316]. The *fuzzy*PPI model has been applied to find the interaction affinity between the SARS-CoV-2 and Human proteins using GO-based information (*see* Figure 4.4 and section 4.3.2 for details). To identify the interactors of SARS-CoV-2 on humans using the *fuzzy*PPI model, a set of candidate proteins are selected, which are identified as the level-1 and level-2 spreader nodes of SARS-CoV using the SIS model as depicted in Figure 4.6. The *fuzzy*PPI model is constructed

Table 4.7: Computation of DC of Figure 4.3 and computation of spreadability rate of selected top 10 spreader nodes by SIS model [158]

Rank	**Proteins**	**Degree Centrality**	**Sum of SIS spreadability rate of top 10 nodes**
1	24	7	2.3
2	2	5	
3	23	5	
4	21	5	
5	1	4	
6	7	4	
7	9	4	
8	3	4	
9	4	4	
10	17	4	
11	5	4	– –
12	22	4	
13	19	4	
14	8	3	
15	11	3	
16	12	3	
17	13	3	
18	16	3	
19	15	3	
20	20	3	
21	32	3	
22	10	2	
23	14	2	
24	18	2	
25	6	2	
26	28	2	
27	29	2	
28	30	2	
29	25	2	
30	26	2	
31	27	2	
32	31	2	
33	33	2	

from the ontological relationship graphs by evaluating the affinity between all possible GO pairs annotated from any target protein pair. Finally, the fuzzy score of the interaction affinity of the protein pair is computed from these GO pair-wise interaction affinity into a range of [0,1].

In the proposed work, experimentally validated human PPIs from publicly available interaction databases have been used, such as HIPPIE [287], STRING [37], BioGRID

Table 4.8: Benchmarking details of gold-standard positive and negative interactions

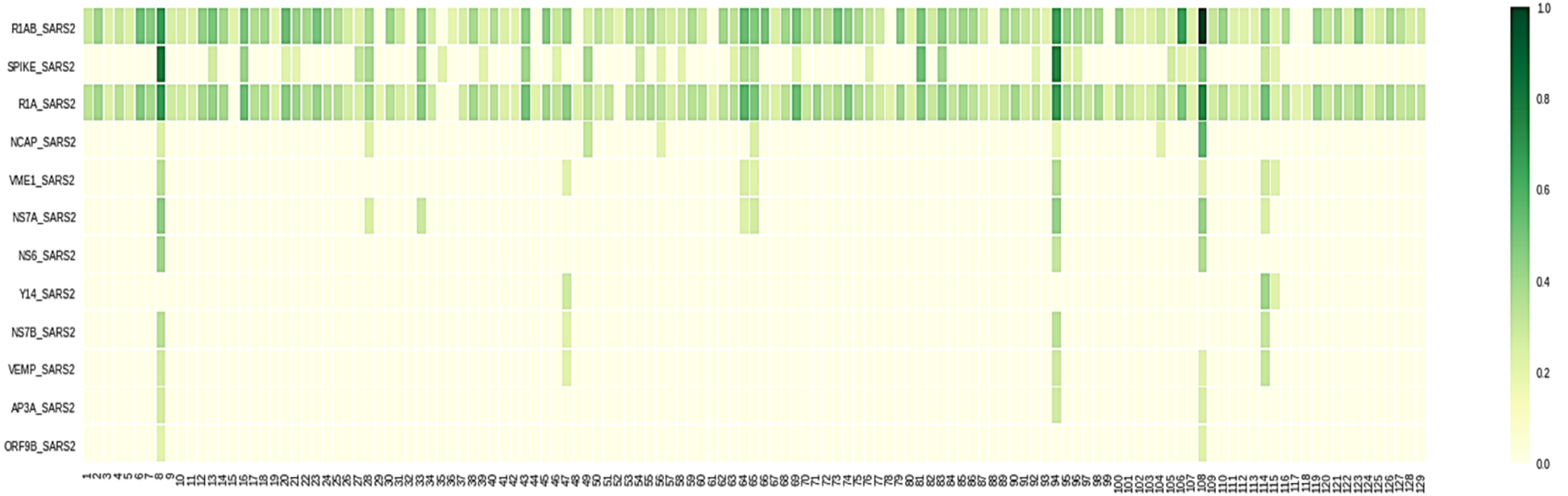
Type	Database	Range of score/Group	Cut-off for Gold standard
Positive	HIPPIE	[0,1]	≥ 0.9
	STRING	[1,1000]	$\geq$900
	BioGRID	[-∞,+∞]	$\geq$10
	DIP	core , non-Core	Core
	HuRI	HuRI(core), HuRI Union	HuRI(core)
Negative	Negatome	Manual, Manual Stringent	Manual Stringent
	Trabuco et al.	Negative (N^{Tr})	N^{Tr}-All Positive^{any}

[35], DIP [34], HuRI [288] for positive data and Negatome 2.0 [257], Trabuco *et al.* [258] for negative data. The positive interactions are also filtered by removing the edges that are common in both positive and negative interaction sets. In each database, Gold standard data is curated by using the scoring scheme provided by the respective databases. The selection criteria are described in Table 4.8.

With this benchmarking data set, the *fuzzy*PPI Model has been assessed with different fuzzy scoring cut-off values. The performance of this assessment is depicted in Table 4.9. In any classification task, specificity signifies the ability to identify a positive sample correctly. In order to identify high-quality positive interactions, here specificity metric has been used. With the increasing value of specificity, the number of FP interactions has shown a sharp fall as depicted in the following Table 4.9. At threshold ≥ 0.2 and ≥ 0.4, the FP is 0.0048% and 0.0001% of total negative interactions respectively. Thus, the Specificity threshold is set at ≥ 0.4. The heatmap, depicted in Figure 4.7, representation of fuzzy interaction affinities with a score $\geq$ of 0.2 for very high specificity $\sim$ 99%. The high-quality interaction is retrieved at threshold 0.4 having almost $\sim$ 99:98% Specificity, which results in a total of 78 interactions between SARS-CoV-2 and humans which includes 37 human level-1 spreaders proteins. The interaction networks predicted from the *Fuzzy*PPI model are shown in Figure 4.8.

4.4.3 Identification of Human Spreader Proteins for nCoV

Human proteins present in the high-quality interactions of human-nCoV PPIN fetched by applying fuzzy affinity threshold are considered level-1 spreaders. From these 37 level-1 spreaders, corresponding level-2 human interactions are obtained using the human-human PPIN dataset. The spreadability index is thus computed for these level-2 human proteins for the identification of level-2 human spreader nodes. The



Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Name	MCL1	SUZ12	ANDR	FUS	SDCB1	AAPK2	CSK21	CLC4M	CDC37	STAT1	CALM1	CDK2	IKKE	BRCA1	PCH2	ACE2	CSK2B	SIAH1	B2CL2	IKKB	TIF1B	ABL1	KPCA	GLYM	HDAC5	JUN	SGTA	XPO1	APC	AKT1	P53	ATX1	TS101	SUMO1	CUL7	FBX6	ESR1	EF1A1	ITAL	SIR1

Index	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
Name	BAG3	MYC	EIF3F	BAG6	KAPCA	CEAM1	VKOR1	1433Z	NXF1	CFTR	NTAQ1	DCTN2	HSP7C	HLAB	TRAF2	ROA1	ATF2	HSPB1	KDM1A	UNK	CUL3	EGFR	B2CL1	TMPS2	UBC	MCM7	SMAD3	NTRK1	TRI27	PTEN	FBW1B	HDAC4	MCM2	CSN5	KAT5	GRB2	A4	CCDC8	UB2D2	CUL1

Index	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
Name	CUL1	PPIA	PHB	TRI23	EP300	TRAF6	SRC	CAV1	CLH1	WNK1	LNX1	HS90A	TANK	ITF2	CD209	UBC9	SFTPD	TERA	SLX4	FOS	CDK1	HGS	TF7L2	TERF2	ELAV1	1433B	DDX1	FINC	MOV10	MYH9	AURKA	BCL2	BMI1	DAXX	CLC4G	TMM17	MDM2	1433E	B2LA1	ZDH17	HDAC1

Index	121	122	123	124	125	126	127	128	129
Name	FBW1A	TRIP6	PRKDC	1433G	EWS	TSNAX	CBP	SUMO2	NUCL

Figure 4.7: Heatmap representation of Fuzzy score of binding affinity between SARS-CoV-2 and human proteins with score≥0.2. Rows represent the SARS-CoV-2 proteins and columns represent potential human target proteins.

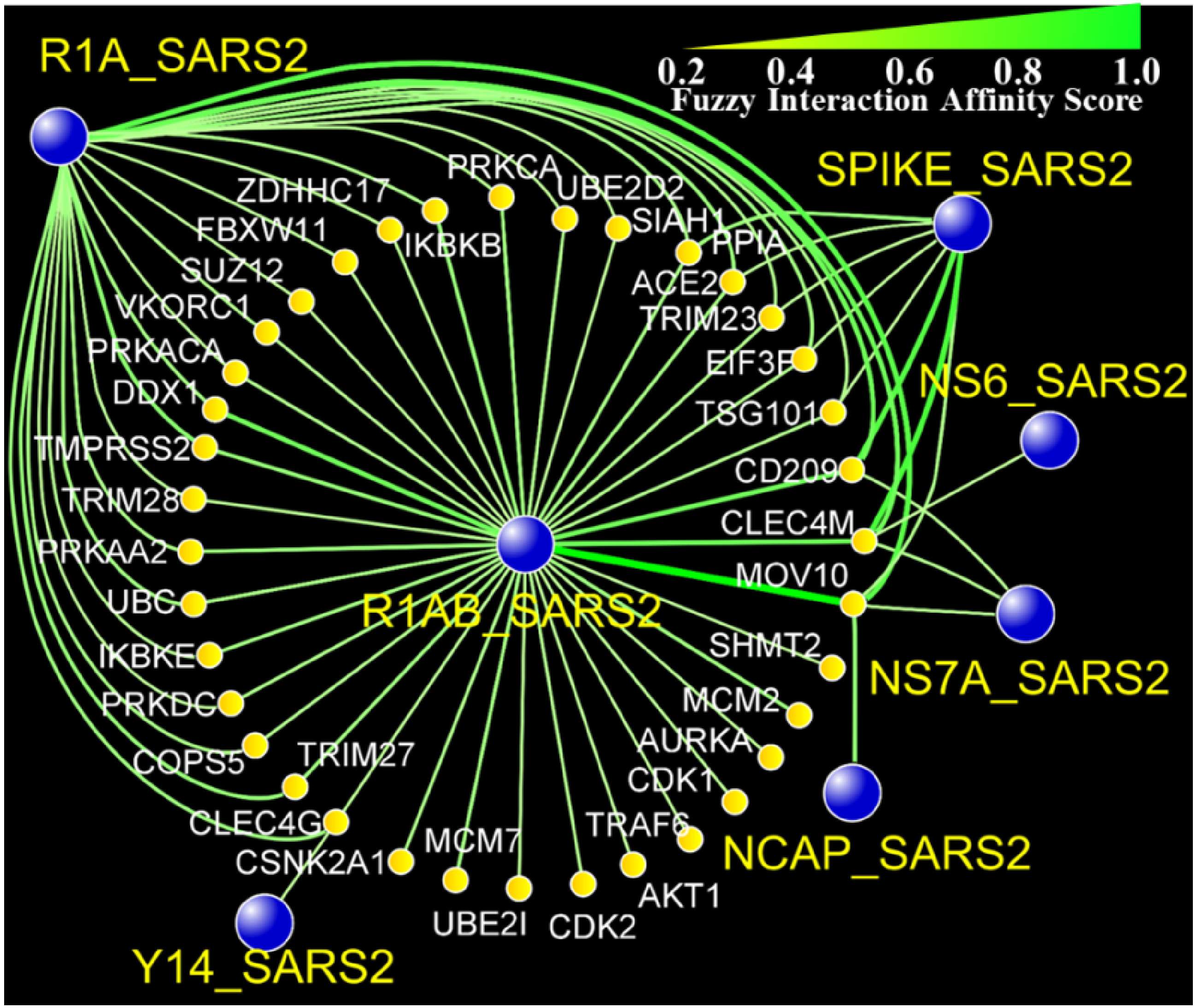


Figure 4.8: Network representation of high-quality interactions (score ≥ 0.4) between SARS-CoV-2 and human proteins. Blue and yellow spherical nodes represent the SARS-CoV-2 and human proteins, respectively. The edge width reflects the fuzzy score of binding affinity.

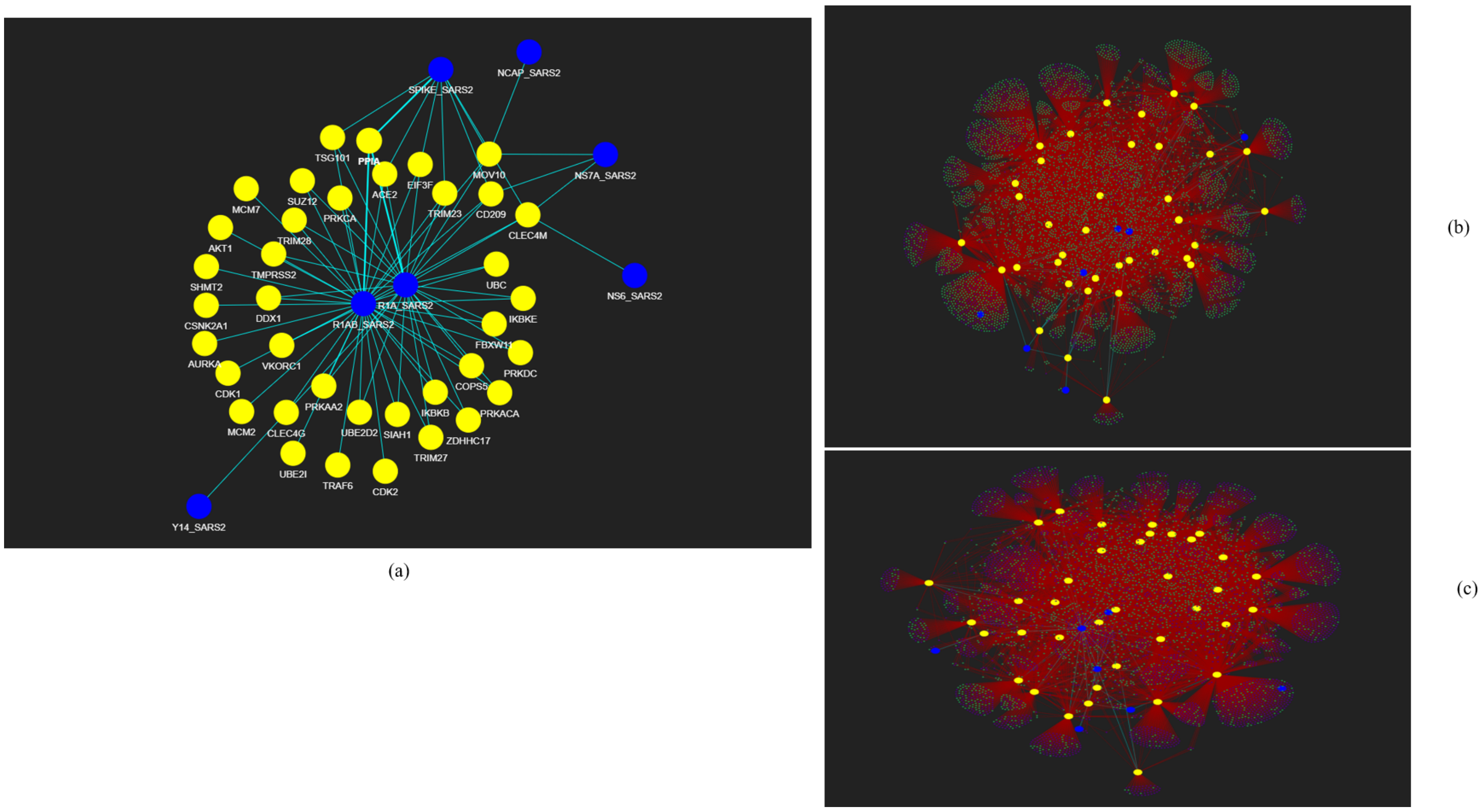


Figure 4.9: Highlights the human level-1 (marked in yellow) and level-2 spreader nodes (marked in green) **a)** SARS-CoV-2 Level-1 human spreaders. **b)** Level-1 & Level-2:low human spreaders at a low threshold of spreadability index. **c)** Level-1 & Level-2:high spreaders at the high threshold of spreadability index

Table 4.9: Fuzzy Interaction affinity-based assessment of benchmark data at different thresholds

Threshold	TP	FN	TN	FP	Precision	Recall	Specificity
0.0001	334179	2721	5144	162368	0.673	0.9919	3.071
0.001	333311	3589	6877	160635	0.6748	0.9893	4.105
0.01	326522	10378	29753	137759	0.7033	0.9692	17.762
0.5	286460	50440	101715	65797	0.8132	0.8503	60.721
0.1	188584	148316	154310	13202	0.9346	0.5598	92.119
0.2	84784	252116	166712	800	0.9907	0.2517	99.522
0.3	48641	288259	167194	318	0.9935	0.1444	99.81
0.4	21187	315713	167502	10	0.9995	0.0629	99.994
0.5	16036	320864	167503	9	0.9994	0.0476	99.995
0.6	12109	324791	167505	7	0.9994	0.0359	99.996
0.7	8879	328021	167509	3	0.9997	0.0264	99.998
0.8	6355	330545	167511	1	0.9998	0.0189	99.999
0.9	2460	334440	167512	0	1	0.0073	100
1	973	335927	167512	0	1	0.0029	100

SIS model also verifies the selection. The computational model of human-nCoV PPIN along with human level-1 (marked in yellow) and level-2 spreader nodes (marked in green) for both low and high thresholds is depicted in Figure 4.9

4.4.4 Validation Using FDA-approved Drugs for COVID-19

After proper assessment of all potential drugs as mentioned in the DrugBank [330], white paper [227], six drugs: Lopinavir [321], Ritonavir [322], Remdesivir [324, 326, 341], Favipiravir [327, 328], and Darunavir [329] are identified which are showing expected results to some extent in the clinical trials done for SARS-CoV-2 vaccine. All approved human protein targets for each of the five approved drugs are fetched from the advanced search section [342] of the drug bank [330, 343]. When searched, these targets are found to play an active role of spreader nodes, in the proposed model of human-nCoV PPIN. This reveals that the selected spreader nodes are of biological importance in transmitting infection in a network makes them the protein drug targets of the potential FDA-approved drugs for COVID-19. The target protein hits in the propoed human-nCoV PPIN for each of the 7 potential FDA-approved drugs are highlighted in Figure 4.10. It can be observed that 3 target proteins for Ritonavir, 2 target protein hit for each of Lopinavir, Darunavir, and Azithromycin, and 1 target protein hit for Remdesivir and Favipiravir. Out of these protein targets, ACE2 is the most important one since it is considered one of the crucial receptors of humans for nCoV to transmit infection deep inside the human cell [344, 345].

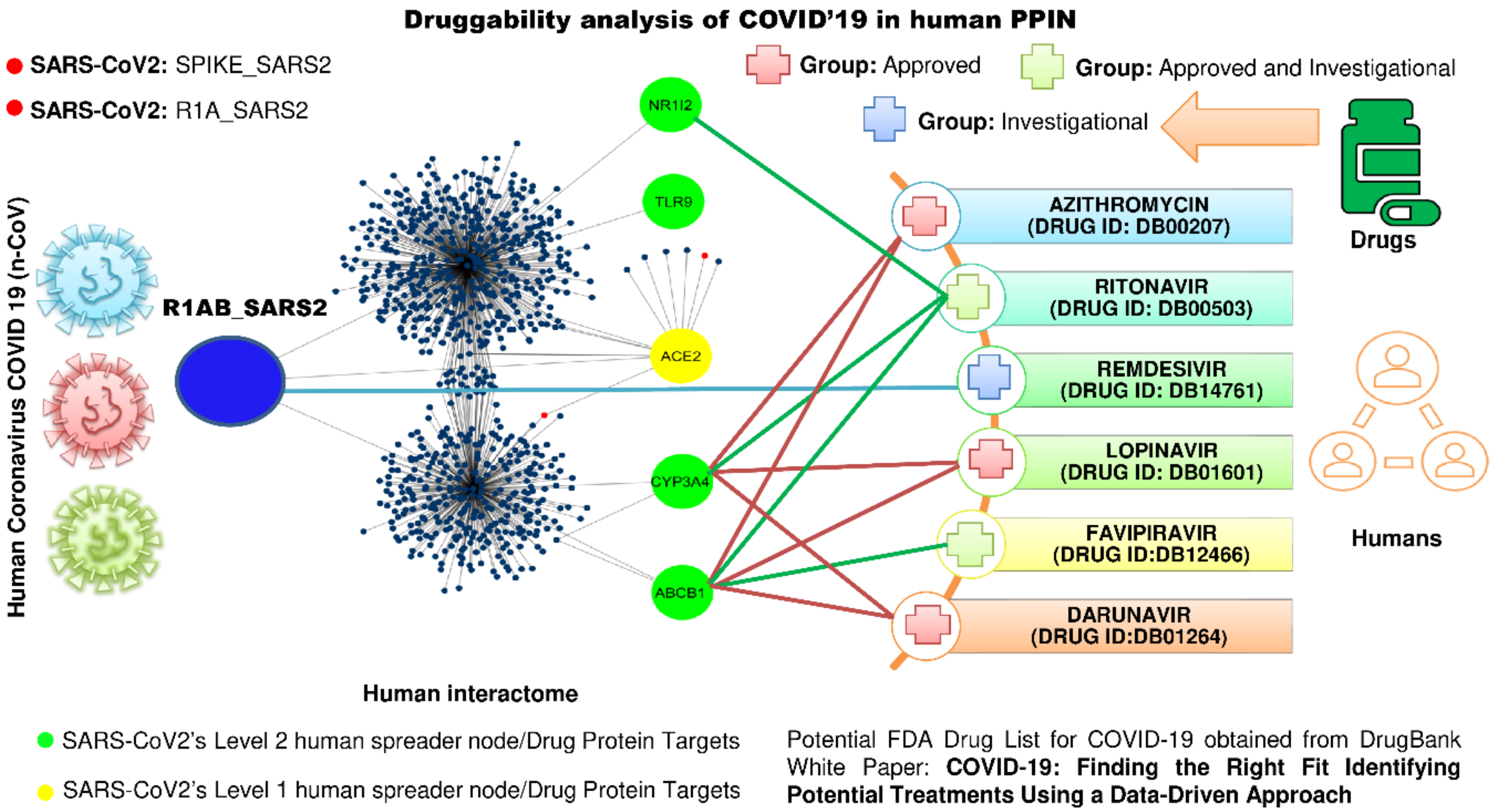

Druggability analysis of COVID'19 in human PPIN
● SARS-CoV2: SPIKE_SARS2
● SARS-CoV2: R1A_SARS2
Human Coronavirus COVID 19 (n-CoV)
R1AB_SARS2
NR1I2
TLR9
ACE2
CYP3A4
ABCB1
Human interactome
Group: Approved
Group: Approved and Investigational
Group: Investigational
Drugs
AZITHROMYCIN (DRUG ID: DB00207)
RITONAVIR (DRUG ID: DB00503)
REMDESIVIR (DRUG ID: DB14761)
LOPINAVIR (DRUG ID: DB01601)
FAVIPIRAVIR (DRUG ID:DB12466)
DARUNAVIR (DRUG ID:DB01264)
Humans
● SARS-CoV2's Level 2 human spreader node/Drug Protein Targets
● SARS-CoV2's Level 1 human spreader node/Drug Protein Targets
Potential FDA Drug List for COVID-19 obtained from DrugBank White Paper: COVID-19: Finding the Right Fit Identifying Potential Treatments Using a Data-Driven Approach

Figure 4.10: Validation of the developed computational model with respect to the target proteins of the FDA-approved drugs for COVID-19 treatment. Yellow- and green-colored nodes denote level-1 and level-2 human spreaders of nCoV, which act as the drug-protein targets.

4.5 Discussion

In any host-pathogen interaction network, the identification of spreader nodes is crucial for disease prognosis. However, not every protein in an interaction network has an intense disease-spreading capability. In the proposed work, SARS-CoV - human PPIN network and the spreader nodes at both level-1 and level-2 using the SIS model have been used. These spreader nodes are considered for computing the PPI affinity score to unmask the level-1 human spreaders of nCoV. In addition, GO annotations have also been considered along with PPIN properties to make this model more effective and significant. With the gradual progress of the work, it has been observed that the selected human spreader nodes, identified by the proposed model, emerge as the potential protein targets of the FDA-approved drugs for COVID-19.

The primary hypotheses of the work may be listed as follows:

- There is a genetic overlap of $\sim 89\%$ [346] between SARS-CoV and SARS-CoV-2, which also leads to a significant overlap in spreader proteins between human-SARS-COV and human-SARSCOV-2 protein-interaction networks.

- *Fuzzy*PPI approach can assess PPI affinities at very high specificity with respect to benchmark datasets, as shown in Figure 4.5

High specificity signifies a meager false-positive rate at a given threshold. Thus, at a 0.4 threshold ($\sim 99{:}9\%$ specificity), the proposed model evaluates high-quality positive interactions in human-nCoV PPIN. Finally, it has been proposed that the developed computational model effectively identifies human-nCoV PPIs with high specificity. The human-nCoV interactions are inferred from another pandemic initiator SARS-CoV, which is highly genetically similar to nCoV. It has also been recognized that the spreadability index of the human spreader proteins, up to level-2, was validated through the SIS model. Due to high network density in human interaction networks, the number of proteins increases with the transition from one level to another. Thus, the proposed model can also identify human spreader proteins in level-2 by using the spreadability index validated by the SIS model.

The proposed method has identified the ACE2 and TMPRSS2 as an interactor of SARS-CoV-2 proteins, which is essential for entry into the human host. SARS-CoV-2 interacts with the SARS-CoV entry receptor ACE2 as SARS-CoV-2 preserves those amino acid residues of SARS-CoV that are essential for ACE2 binding [347]. However, the binding strength of SARS-CoV-2 with ACE2 is 10 to 20 times more than the SARS-CoV-2 - ACE2 attachment [348]. This is because several changes occur in the receptor-binding domains (RBDs) of SARS-CoV-2 spike protein [166]. In addition, the cellular serine protease TMPRSS2 primes SARS-CoV-2 for host entry, and a Serine protease inhibitor blocks SARS-CoV-2 infection of lung cells [166, 347–349]. Thus, TMPRSS2 activity is essential for viral spread and pathogenesis in the infected host [166,347–350]

In a recent study [164], Gordon *et al.* have identified 332 high-confidence SARS-CoV-2-human PPIs where they have worked on the sequence analysis of SARS-CoV-2 isolates. They cloned, tagged, and expressed 26 of the 29 SARS-CoV-2 proteins in human cells and identified the human proteins that were physically associated with each using affinity-purification mass spectrometry (AP-MS). However, while comparing their seminal work with the proposed method, UniProt accession IDs are not directly mapping to the SARS-CoV-2 protein sequences implemented by Gordon *et al.*. In the proposed method, the work is performed only on the UniProt listed SARS-CoV-2 proteins and applies a mathematical model of binding affinity assessment on a subset of UniProt listed reviewed Human proteins. Therefore, direct comparison and validation could not be possible with respect to Gordon *et al.*, primarily because of the unavailability of direct mapping of SARS-CoV-2 proteins into corresponding UniProt accession ids. However, an attempt has been made to map UniProt ids of SARS-CoV-2 proteins of *et al.*, from COVID-19 UniProtKB reference data [300].

One of the key highlights of the study may be underlined by the fact that the target proteins of the potential FDA-approved drugs for COVID-19 overlap with the spreader nodes of the proposed human-nCoV PPIN. Target proteins of six potential FDA-approved drugs: Lopinavir [321], Ritonavir [322], Azithromycin [323], Remdesivir [324, 326, 341], Favipiravir [327, 328], and Darunavir [329] for COVID-19 as mentioned in the DrugBank white paper [227] overlap with the spreader nodes of the proposed *in-silico* human-nCoV PPI model (*see* Figure 4.10). Though clinical trials for the COVID-19 vaccine are on their way to date, three out of the six repurposed drugs, i.e., Remdesivir [341] and Favipiravir [351] are found to be the most promising as well as effective ones. The proposed model successfully identified their protein targets R1AB SARS2, TLR9, ACE2, CYP3A4, and ABCB1 as spreader nodes. This assessment reveals the fact that these spreader nodes indeed have biological relevance relative to disease propagation. It also motivates us to further do a drug repurposing study on the generated SARS-CoV-2 - human PPIN in the subsequent research work [352], which highlights that the drug Fostamatinib/R406 might be one of the potential drugs to be used for SARS-CoV-2.

In this chapter, a computational model has been developed by computing a fuzzy interaction affinity between human and nCoV proteins to identify the potential spreader nodes using fuzzy semantic scoring methods, discussed previously. The spreader proteins at both levels have been validated by using SIS model. For disease prediction, spreader node identification is essential. The suggested model's identification of the selected human spreading nodes as prospective protein targets for COVID-19 treatment has been seen to advance the work gradually. In the following chapter, a *in-silico* drug repurposing study has been proposed by using the target spreader proteins for COVID-19 treatment. COVID-19 Symptom-based analysis has also been done to identify the list of FDA-approved drugs. A molecular docking study has also been done on the identified drug for validation.

Chapter 5

Drug Repurposing for COVID-19 using a Novel *in-silico* Method

5.1 Background

The world has witnessed several severe epidemics like Spanish flu, ebola, cholera etc. Now the world is in front of the most life-threatening viral outburst with COVID-19. The feature that makes this new coronavirus, nCoV, unique is its ability to quickly transmit through an infected COVID patient [353]. The virus causing COVID-19 is an assimilation of accessory, non-structural and structural proteins [354]. According to WHO coronavirus disease dashboard [355], 162,704,139 confirmed cases of COVID-19, including 3,374,052 deaths, have been reported as of 1:39 pm CEST, 17 May 2021. Based on the prior knowledge of major outbreaks of Ebola, cholera etc., treatments with different antiviral drugs are considered and implemented to terminate COVID-19 based on previous knowledge of significant attacks. With the gradual increase in the COVID-19 mortality rate, thus there is an urgent need for an effective drug/vaccine. Several drugs like Remdesivir, Azithromycin, Favirapir, Ritonavir, Darunavir, etc., are put under evaluation in more than 300 clinical trials to treat COVID-19. On the other hand, several vaccines like Pfizer-BioNTech, Moderna, Johnson & Johnson's Janssen, Sputnik V, Covishield, Covaxin, etc., also evolved from the research study. While few of them already gets approved, others show encouraging results and are still under assessment. In parallel, there are also significant developments in new drugs for related diseases. But, since the approval of new molecules takes substantial time, drug repurposing studies have also gained considerable momentum. A literature survey [356] is recently carried out through a refined computational search in various online repositories like Google Scholar, Science Direct, PubMed, etc., to enlist various COVID-19 drug-related research articles since the onset of this pandemic. Almost 22 most relevant COVID-19 related drug articles [356] have been filtered out from the search results. All these significant researches and some others have been extensively studied. It is noted from the study that the most recommended drugs are azithromycin, lopinavir, ritonavir, remdesivir, and favipiravir. It also appears that the amount of data accessible for these drugs is insufficient to recommend any one of them as a treatment for COVID-19 until and unless the necessary amount of appropriate clinical trials are executed. Relative data comparison is missing in almost all human-related studies about COVID-19. So, it is uncertain whether the COVID infected patient recovers due to applying the suggested drug or recover due to extensive clinical care and isolation. However, some of the *in-vitro* studies have shown favourable results for these drugs. Still, these are all preliminary data, which need much more evidence before putting it in clinical trials.

COVID-19 vaccines are also in the same race as COVID-19 drugs. According to a current report [357]. by the Centers for Disease Control and Prevention (CDC), Pfizer-BioNTech, Moderna, and Johnson & Johnson's Janssen are the recommended and authorized vaccines in the United States COVID-19. In addition, another vaccine Sputnik-V has been developed by Gamaleya National Center in Russia. Though it is found effective in initial trials, it has been recommended only for emergency use by the Technical Advisory Group (TAG) of WHO. In contrast, Covishield [358] and Covaxin [359] are the recommended vaccines in India. All these vaccines might have specific side effects that need further analysis and research. But the two significant areas of concern are:

- It is still unknown whether these vaccines are effective against all COVID-19 strains.

- *"getting vaccinated"* does not guarantee that COVID-19 will not happen again

Still, it says it could save somebody's life by refraining from getting seriously ill if they get infected with COVID-19. So, vaccines do not provide any herd immunity.

Due to the daily increase in deaths [360] there is an urgent need to identify a potential vaccine/drug that will eventually help eradicate COVID-19. So, with no other alternatives left, clinical trials have been started by WHO [137] on all the suggested drugs which have somehow proved to be substantially beneficial in case studies of COVID-19. Drug design needs a proper understanding of disease transmission mechanisms that can be effectively done by analyzing host-pathogen PPIN [137]. Pathogen facilitates disease progression as it has the potential to transform itself by mutation. Infection of pathogen gets broadcasted through the connecting edge of interaction between host and pathogen. Thus, it is essential to explore target proteins and their interactions in host-pathogen PPI [56] networks for potential drug discovery [361]. However, the only recognized *in-vitro* human-nCoV PPIN available to date is in the work of Gordon *et al.* [164]. But UniProt reviewed nCoV proteins cannot be mapped through this *in-vitro* generated PPIN. So, all this led to the development of an *in-silico* human-nCoV PPIN through the SIS model [317] and fuzzy thresholding. Further study of protein targets of potential FDA-approved drugs [227] of COVID-19 in the formed human-nCoV PPIN network also shows that FDA-approved drug, Fostamatinib which is having R406 as its active promoiety [362] can be a potential drug for COVID-19 treatment. Rigel Pharmaceuticals, Inc. [363] got approval for TAVALISSE (Fostamatinib disodium hexahydrate) for the treatment of Chronic immune thrombocytopenia (ITP) [364] from the FDA-approved on 17/04/2018.

The major contribution in this work is designed as follows:

- It uses an *in-silico* model which has been developed to identify potential spreader proteins in a human-nCoV interaction network in the work of Saha *et al.* [158, 226], which was validated using proteins which are the targets of potential FDA-approved drugs [227] for COVID-19 treatment.
- A two-way analysis:
  - Human-nCoV interaction network analysis
  - COVID-19 symptom [225] based analysis (including *"loss of smell"*), have been implemented to detect the potential candidates in the list of FDA-approved drugs for COVID-19
- In both the analyses, Fostamatinib/R406 [362], an FDA-approved drug, commonly used for the treatment of chronic ITP [365] ranks as the top having a maximum overlap of target proteins in the human-nCoV interaction network.
- Fostamatinib/R406 is used for ITP [366] which is also associated with COVID-19 infections [367]
- Molecular docking, has been also performed on Fostamatinib/R406 and other potential FDA-approved drugs [227] with the available major COVID-19 crystal structures having PDB IDs: 6LU7 [368], 6M2Q [369], 6W9C [370], 6M0J [165], 6M71 [371], and 6VXX [372]. While Fostamatinib registers the highest score for 6LU7 and 6M2Q, it obtains a second position to the other COVID-19 structures.
- The active promoiety of Fostamatinib, *i.e.*, R406, generates the highest docking scores compared to all other active metabolites.

5.2 Dataset

The following datasets, given in Table 5.1, are used for the following work. The table depicts that there is no interaction evidence between SARS-CoV - SARS-CoV protein interactions and SARS-CoV-2 - SARS-CoV-2 protein interactions in publicly available databases.

Table 5.1: Description and details of the datasets

Database Name	Description	Nodes	Interactions /Edges
Human PPIN	Human-human protein interactions	21557	342353
SARS-CoV PPIN	SARS-CoV-SARS-CoV protein interactions	7	—-
SARS-CoV - human PPIN	SARS-CoV - human protein interactions	120	118
SARS-CoV-2 proteins	UniProt collected reviewed COVID-19 proteins	14	—-

5.3 Methodology

The proposed methodology involves 4 datasets: 1) Human PPIN [318, 319] 2) SARS-CoV PPIN [313] 3) SARS-CoV - human PPIN [313] and 4) SARS-CoV-2 proteins [33].The overall dataset statistics are highlighted in Table 5.1. The entire proposed methodology of drug repurposing can be categorized into four major sub-sections:

5.3.1 Detection of Spreader Nodes in Human-nCoV PPIN

The only recognized *in-vitro* human-nCoV PPIN available to date is in the work of Gordon *et al.* [164]. But UniProt reviewed nCoV proteins cannot be mapped through this *in-vitro* generated PPIN. So, an attempt has been made to construct a human-nCoV PPIN based on the available PPIN information of SARS-CoV, which has 89% similarity [311, 312] with SARS-CoV-2. Not every protein in a PPIN is a spreader protein/node. Spreader proteins are considered to be those specific proteins that have a unique fast capability of transmitting infection in their neighborhood in a short time [158]. They are identified through a spreadability index computed by the combination of three terminologies s discussed in chapter 4: 1) E_r [332], 2) N_D [332] and 3) N_w [331].

The spreadability index of node i is defined as the ability of node i to mediate a viral infection in a PPIN. With the previously mentioned three terminologies, spreadability index can mathematically defined as:

$$Spreadability\_index(i) = (Edgeratio(i) \times neighborhood\_diversity(i) + \{Node(w_i)\}) \quad (5.1)$$

Nodes having a high spreadability index are termed as spreader nodes, *i.e.*, if the viral proteins establish interactions with these nodes, then the viral infection can be mediated to a more significant number of nodes in a much short amount of time compared to the other nodes in PPIN. Identification of spreader proteins is conducted initially in the SARS-CoV PPIN dataset. Corresponding connected human proteins, *i.e.*, level-1 and level-2 of selected SARS-CoV spreader proteins, are chosen from SARS-CoV - human PPIN and human PPIN datasets. The selected spreader nodes are also validated by SIS model [317]. The schematic diagram for the construction of COVID-19-Human PPIN is given in Figure 5.1.

In a PPIN, level-1 proteins of a node are those proteins that are in direct connection with that node, *i.e.*, its immediate neighbors, whereas level-2 proteins are those proteins that are indirectly connected with level-1 proteins of that node, *i.e.*, its indirect neighbors [152]

This results in forming a PPIN consisting of 7 SARS-CoV, 24 level-1 and 111 level-2 human spreader proteins, respectively, under a low threshold [158]. The potential human-nCoV interactions have been identified using developed *in-silico fuzzy*PPI model [128]. In this model, SARS-COV spreader nodes which include level-1 and level-2 proteins in humans are considered the candidate set of interactors for nCoV [226]. The human-nCoV pair-wise relationships are quantified using the semantic similarity of their annotated GO pairs. A hybrid approach has been applied to assess the semantic similarity between GO target pairs using the topological properties of three GO sub-graphs [261]. These GO-level assessment scores are incorporated to obtain the fuzzy interaction affinity score which ranges between [0, 1] between the target human and nCoV protein pair and results (see Figure 5.1). The high specificity *i.e.* 99.9% has been achieved on a threshold of 0.4 fuzzy interaction affinity score on a benchmark human PPI dataset. Finally, with the high specificity threshold, potential interactions are identified between nCoV bait and human prey [226].

5.3.2 Identification of FDA-approved Candidate Drugs *w.r.t* COVID-19 Spreader Nodes Using Human-nCoV Interaction Network Analysis

Once the COVID-19-human PPIN is formed, all the level-1 and level-2 human proteins of COVID-19 are mapped with their corresponding drugs from DrugBank [330]. DrugBank is an online repository [343] that contains comprehensive data about drugs, drug-protein targets and information about drug metabolism. Due to the high-quality annotation in DrugBank, it becomes the most used database in almost all *in-silico* methodologies involved in drug design, docking of drugs, and drug interaction prediction. It contains about 60% and 10% of FDA-approved and experimental drugs, respectively [330]. On proper analysis, it has been observed that various spreader nodes in COVID-19-human PPIN are the protein targets of potential COVID-19 FDA-approved

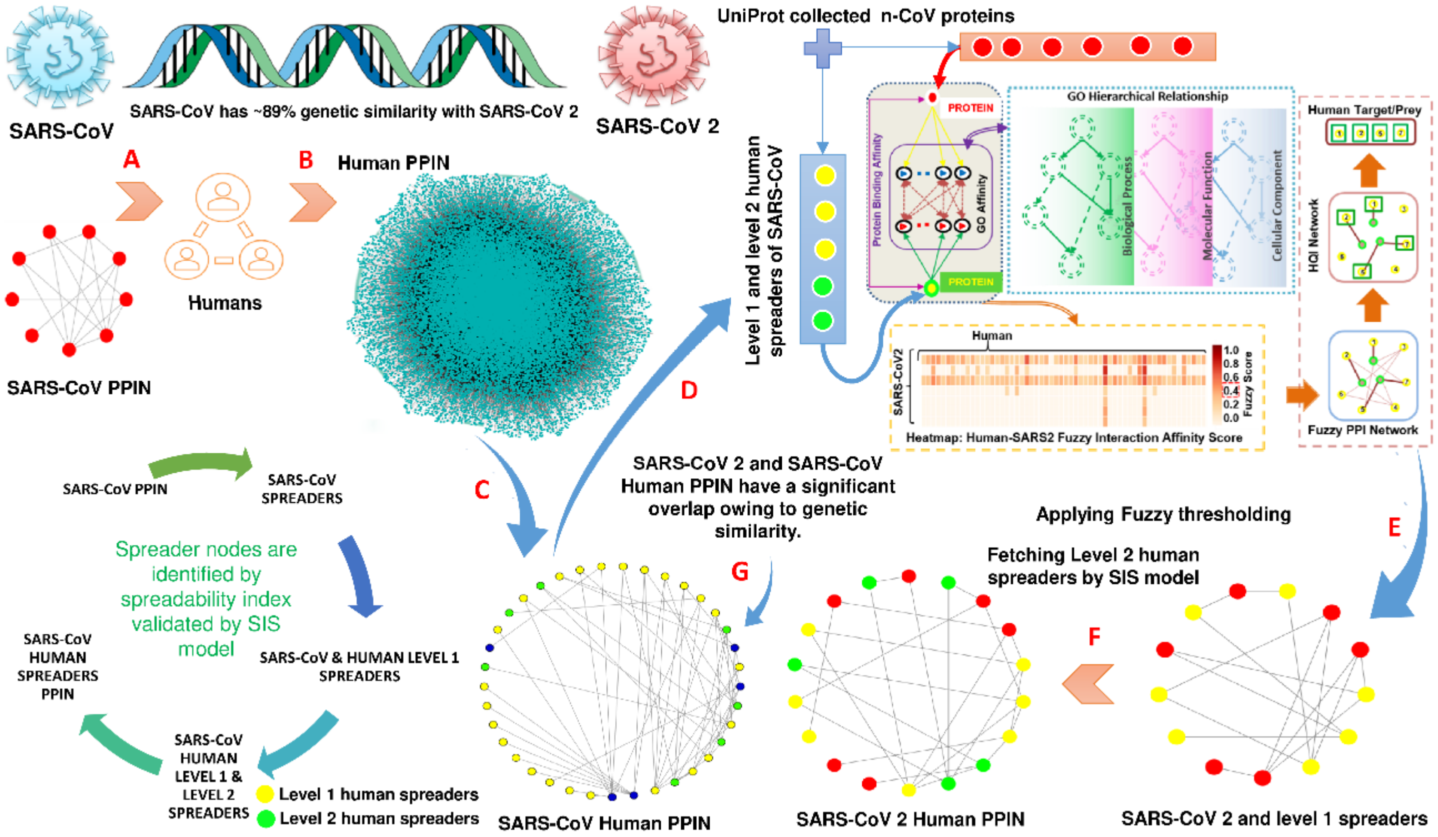

SARS-CoV
SARS-CoV has ~89% genetic similarity with SARS-CoV 2
SARS-CoV 2
UniProt collected n-CoV proteins
A
B
Human PPIN
Humans
SARS-CoV PPIN
PROTEIN
Protein Binding Affinity
GO Affinity
PROTEIN
GO Hierarchical Relationship
Biological Process
Molecular Function
Cellular Component
Human Target/Prey
HQI Network
Fuzzy PPI Network
Level 1 and level 2 human spreaders of SARS-CoV
Human
SARS-CoV2
1.0
0.8
0.6
0.4
0.2
0.0
Fuzzy Score
Heatmap: Human-SARS2 Fuzzy Interaction Affinity Score
D
C
SARS-CoV PPIN
SARS-CoV SPREADERS
Spreader nodes are identified by spreadability index validated by SIS model
SARS-CoV HUMAN SPREADERS PPIN
SARS-CoV & HUMAN LEVEL 1 SPREADERS
SARS-CoV HUMAN LEVEL 1 & LEVEL 2 SPREADERS
Level 1 human spreaders
Level 2 human spreaders
SARS-CoV 2 and SARS-CoV Human PPIN have a significant overlap owing to genetic similarity.
G
Applying Fuzzy thresholding
E
Fetching Level 2 human spreaders by SIS model
F
SARS-CoV Human PPIN
SARS-CoV 2 Human PPIN
SARS-CoV 2 and level 1 spreaders

Figure 5.1: The schematic diagram for the construction of COVID-19-Human PPIN. **A.** Formation of SARS-CoV PPIN (Red nodes) **B.** Formation of human PPIN **C.** Spreader nodes identification in SARS-CoV - human PPIN which is formed by the application of spreadability index and SIS model (Blue nodes represent SARS-CoV spreaders, while yellow and green denote level-1 and level-2 human spreaders) **D.** UniProt collected n-CoV proteins along with level-1 and level-2 human spreader nodes in SARS-CoV - human PPIN are provided as an input to fuzzy thresholding **E.** SARS-CoV-2, and level-1 Human spreaders are identified based on Fuzzy score (Red nodes represent SARS-CoV-2 spreaders, while yellow denotes level-1 human spreaders) **F.** Level-2 Human spreaders are identified by the application of the SIS model (Green nodes represent level-2 human spreaders) **G.** SARS-CoV - human PPIN and SARS-CoV-2-Human PPIN have a significant overlap owing to their genetic similarity.

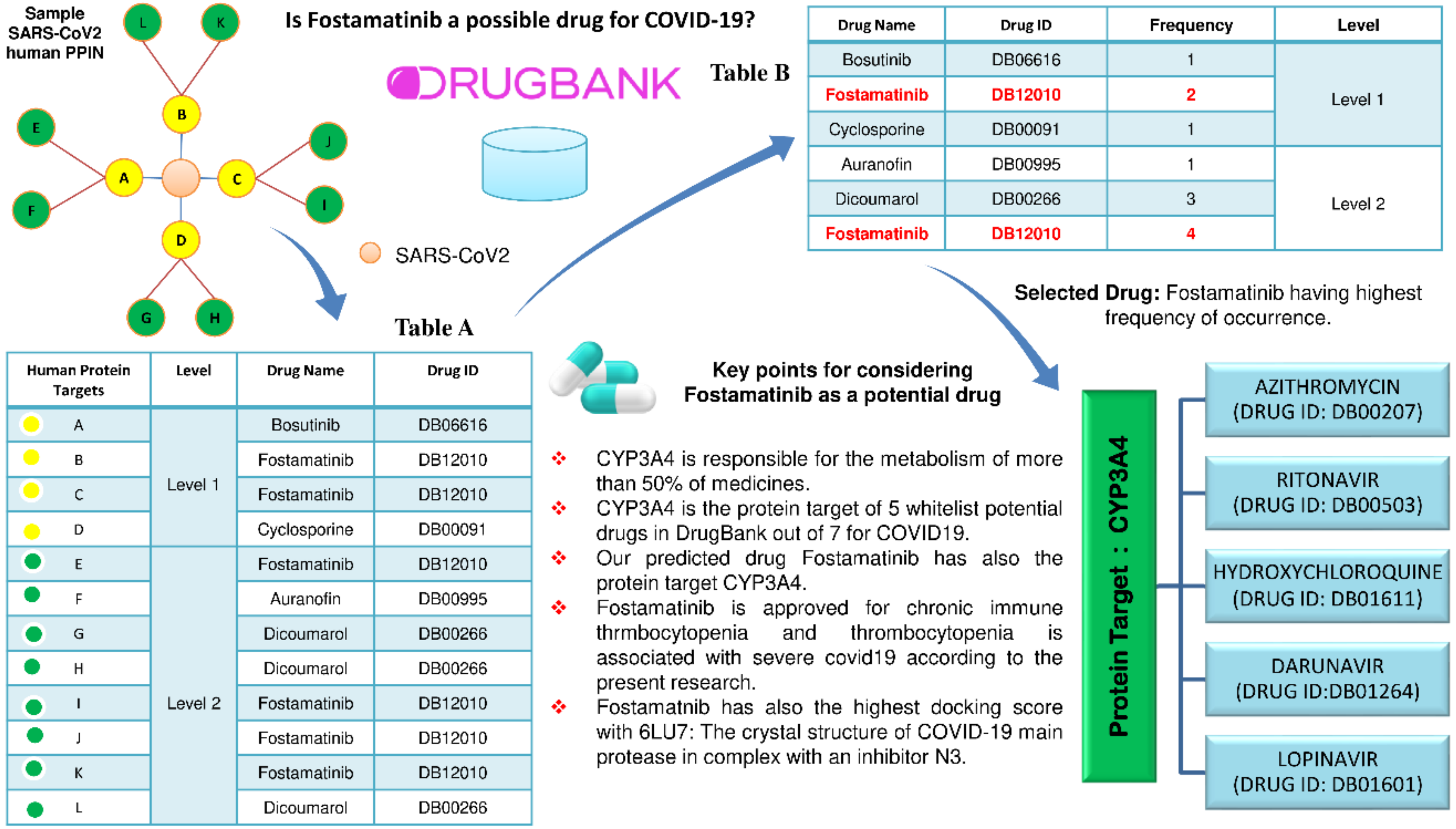
listed drugs [227]: hydroxychloroquine [323, 373], azithromycin [323], lopinavir [321], ritonavir [322], remdesivir [324, 326, 341], and Favipiravir [327, 328]. The details of significant overlap between spreader nodes and drug-protein targets are highlighted in Table 5.2 [226]

Table 5.2: Details of overlap of spreader nodes and potential COVID-19 FDA-approved drugs

Sl. No.	**COVID-19 FDA-approvedlisted Drugs**	**DrugBank ID**	**Drug Protein targets/ Spreader nodes**	**No. of hits**
1	Hydroxychloroquine	DB01611	TLR9, ACE2, CYP3A4, ABCB1	4
2	Azithromycin	DB00207	CYP3A4, ABCB1	2
3	Lopinavir	DB01601	CYP3A4, ABCB1	2
4	Ritonavir	DB00503	CYP3A4, ABCB1	2
5	Remdesivir	DB14761	R1AB_SARS2	1
6	Favipiravir	DB12466	ABCB1	1
7	Darunavir	DB01264	CYP3A4, ABCB1	2

It has been observed from the table that hydroxychloroquine has the highest hit/overlap, *i.e.* four, while each of azithromycin, lopinavir, ritonavir and darunavir has two hits [226]. Remdesivir and favipiravir have one impact individually [226]. Remdesivir is the only drug that acts directly on the COVID-19 protein R1AB_SARS2. Significant overlapping drug targets and spreader nodes in Table 5.2 motivate us to analyze further and develop a consensus strategy to identify potential drugs for COVID-19 treatment. The consensus strategy is described in Algorithm 5.1. Drug consensus score (DCS) is used in Algorithm 5.1, defined as the frequency of occurrences of a drug at a particular level of PPIN. Execution of Algorithm 5.1 is also highlighted in Figure 5.2 by considering a randomly generated COVID-19-Human PPIN. In this PPIN, corresponding linked drugs are mapped with each human protein (marked as green) in level-1 and level-2, as shown in Table A in Figure 5.2.

Hence the DCS, *i.e.*, frequency of each drug, is computed and highlighted in Table B in Figure 5.2. Since Fostamatinib has the highest DCS in both levels, it is considered the potential drug for the target nCoV protein in the randomly generated COVID-19-Human PPIN (marked as red). Algorithm 5.1 is also implemented to the host targets of *in-vitro* generated human-nCoV PPIN of Gordon *et al.* [164].



Is Fostamatinib a possible drug for COVID-19?

Table A

Human Protein Targets		Level	Drug Name	Drug ID
🟡	A	Level 1	Bosutinib	DB06616
🟡	B		Fostamatinib	DB12010
🟡	C		Fostamatinib	DB12010
🟡	D		Cyclosporine	DB00091
🟢	E	Level 2	Fostamatinib	DB12010
🟢	F		Auranofin	DB00995
🟢	G		Dicoumarol	DB00266
🟢	H		Dicoumarol	DB00266
🟢	I		Fostamatinib	DB12010
🟢	J		Fostamatinib	DB12010
🟢	K		Fostamatinib	DB12010
🟢	L		Dicoumarol	DB00266

Table B

Drug Name	Drug ID	Frequency	Level
Bosutinib	DB06616	1	Level 1
Fostamatinib	DB12010	2	
Cyclosporine	DB00091	1	
Auranofin	DB00995	1	Level 2
Dicoumarol	DB00266	3	
Fostamatinib	DB12010	4	

Selected Drug: Fostamatinib having highest frequency of occurrence.

Key points for considering Fostamatinib as a potential drug

- ❖ CYP3A4 is responsible for the metabolism of more than 50% of medicines.
- ❖ CYP3A4 is the protein target of 5 whitelist potential drugs in DrugBank out of 7 for COVID19.
- ❖ Our predicted drug Fostamatinib has also the protein target CYP3A4.
- ❖ Fostamatinib is approved for chronic immune thrmbocytopenia and thrombocytopenia is associated with severe covid19 according to the present research.
- ❖ Fostamatnib has also the highest docking score with 6LU7: The crystal structure of COVID-19 main protease in complex with an inhibitor N3.

Protein Target : CYP3A4

- AZITHROMYCIN (DRUG ID: DB00207)
- RITONAVIR (DRUG ID: DB00503)
- HYDROXYCHLOROQUINE (DRUG ID: DB01611)
- DARUNAVIR (DRUG ID:DB01264)
- LOPINAVIR (DRUG ID: DB01601)

Figure 5.2: A DCS was adopted to choose Fostamatinib/R406 as a potential COVID-19 drug. Other connecting biological links for selecting the same have also been highlighted.

Algorithm 5.1: Potential Drug Selection using Consensus Strategy (DCS)

```
Input  : The set of spreader proteins in COVID-19-human PPIN, as $S^l =$
         $\{S_1, S_2 \ldots..S_n\}$. The set of drugs listed in DrugBank, $D^l =$
         $\{D_1, D_2 \ldots..D_n\}$for spreader protein $S^i$, $(1 \leq i \leq n)$
Output: $mxKey$- a potential drug for COVID-19

    /*  Empty list of drug for level-1 and level-2 spreader   */
$DL^1 \leftarrow [\,],DL^2 \leftarrow [\,]$
$Dc^1 \leftarrow [\,],\ Dc^2 \leftarrow [\,]$
for each spreader protein $S^{i=1}$ do
    if $S^i \in D_t(D_k\ )$ then
        append $D_k$ in $DL^1$
    end
end
for each drug $D_i$ from $DL^1$ do
    compute $DCS^1(D_i)$ and $Dc^1[D_i] \leftarrow DCS^1(D_i)$
end
for each spreader protein $S^i$ from $S^{l=2}$ do
    if $S^i \in D_t(D_k)$ then
        append $D_k$ in $DL^2$
    end
end
$Dc^1_{sort} \leftarrow sort_{desc}(Dc^1)$
$Dc^2_{sort} \leftarrow sort_{desc}(Dc^2)$
if $(Dc^1_{sort}(D^1_0) \geq Dc^2_{sort}(D^2_0))$ then
    mxKey $\leftarrow D^1_0$    /*  drug D from $Dc^1$ with highest $DCS^1$ value in
    $DL^1$   */
    mxKey $\leftarrow D^2_0$  /*  # drug D from $Dc^1$ with highest $DCS^2$ value in
     $DL^1$   */
end
return mxKey
```

$D_t\ (D_k\ )$: Target proteins of drug D_k
$DCS^t\ (D_i\ )$: DCS of drug D_i at level t
$sort_{desc}\ (Dc^f)$: Descending ordered sorting of dictionary Dc^f, based on the values of the dictionary. f-level of spreader protein.

5.3.3 Identification of FDA-approved Candidate Drugs *w.r.t* COVID-19 Spreader Nodes Using COVID-19 Symptoms, Risk Factors and Clinical Outcome-Based Analysis

COVID-19 is associated with specific health symptoms like cough, fever, breathing difficulty, *“loss of smell”*, etc. Usually, the symptom *“loss of smell”* plays a higher significant role in comparison to the other existing symptoms [374–376] due to the following reason:

- According to a correspondence published on April 15, 2020, in *The Lancet Infectious Diseases* [374] it was highlighted by the authors that though the reason of losing smell by COVID-19 patients was not discovered yet, their initial inspection suggests that *"loss of smell" "manifests either early in the disease process or in patients with mild or no constitutional symptoms."*.

- Another correspondence published on June 04, 2020, in The Lancet [375], stated that *"after quantifying the sensitivity, specificity, positive predicted value, and negative predicted value of fever, cough, fever or cough, and "loss of smell" in 76,260 users of the COVID Symptom Study app who underwent the SARS-CoV-2 test (13,863 testing positive; 62,397 testing negative), they found that the predictive ability of "loss of smell" and taste to be higher than fever or persistent cough, which is in line with their previous finding that "loss of smell" and taste was the strongest predictor of having the virus [376]. Moreover, they found that the median duration of anosmia symptoms was 5 days, whereas the median duration of fever was only 2 days."*

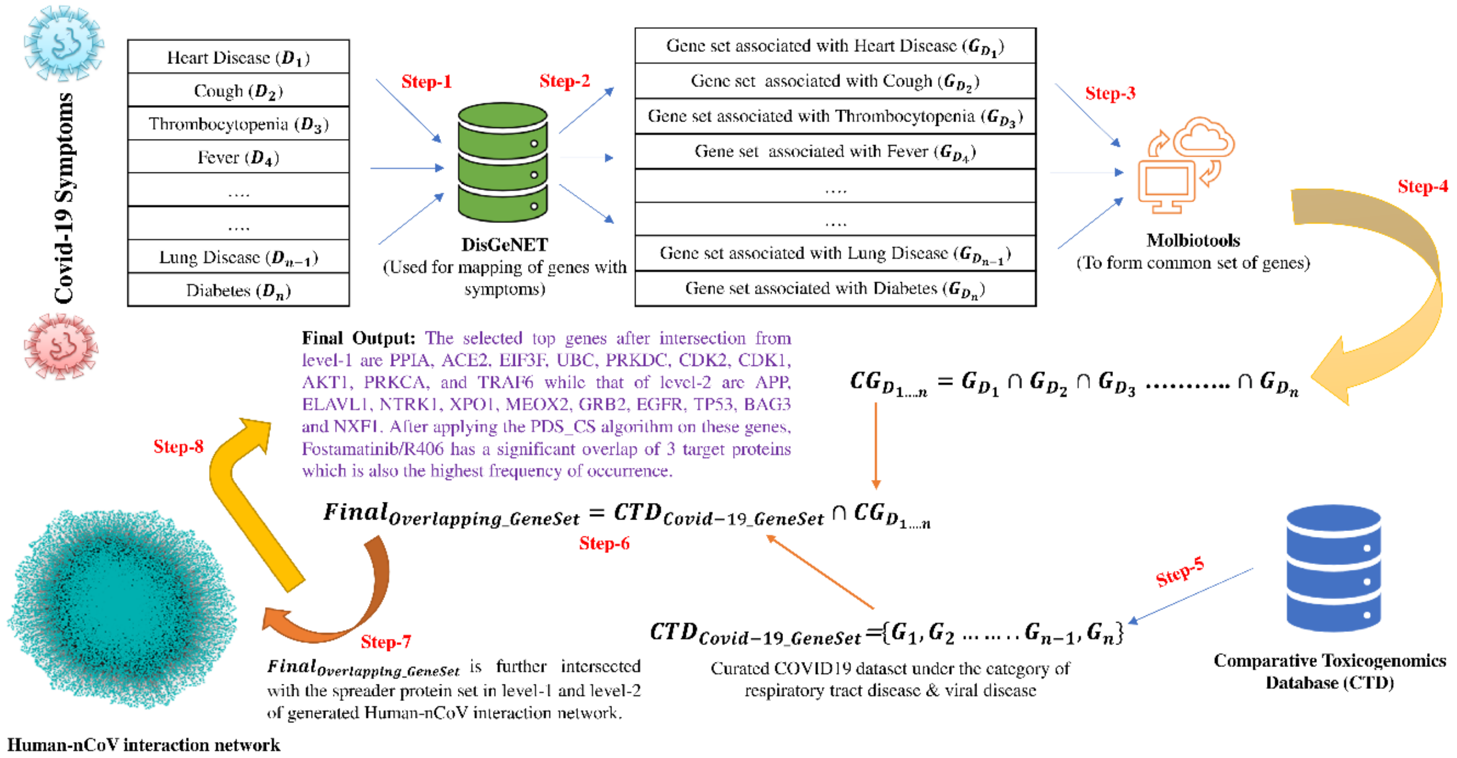
These symptoms are linked with specific human gene sets chosen as the bait's possible target prey, *i.e.*, nCoV. The same is also true for other risk factors and clinical outcomes of COVID-19. So, all these genes under the mentioned categorization are grouped [322] from the disease-gene dataset available from DisGeNET. DisGeNET [377] is considered one of the significant resources covering all the relevant information about various diseases. These multiple gene sets are compared with each other [225] using molbiotools. The resultant gene set is again compared [225] with the curated COVID-19 dataset available in Comparative Toxicogenomics Database (CTD) [378] under respiratory tract disease & viral disease to obtain an overlapping gene set. CTD is yet another significant resource that collects, organizes and stores scientific data which describes the interrelationship between proteins, pathways, interactions, drugs etc. The overlapping gene set is further intersected with the spreader protein set in level-1 and level-2 of generated human-nCoV interaction network [226]. The top 10 key genes are selected from the resultant intersection in each level based on the fuzzy score and spreadability index score in level-1 and level-2. These genes are considered the most significant ones that play an essential role in COVID-19 transmission [176, 379–382] and prevention [383–387] in the human-nCoV interaction network. Potential FDA-approved drugs having these key genes/spreader proteins as known targets are identified from DrugBank data [330, 343]. Then Algorithm 5.1 is executed to determine the most potential candidate FDA-approved drugs for COVID-19. The entire process of the symptom-based analysis is highlighted in Figure 5.3.

5.3.4 Computational Docking of Potential Drugs *w.r.t* COVID-19 Protein Structures

The earlier sections discuss how several SARS-CoV-2 proteins react with the human level-1 and level-2 spreaders to form SARS-CoV-2 - Human PPIN. Hence, a drug repurposing study is done based on network and symptom-based analysis. It reveals Fostamatinib/R406 might be a potential drug for COVID-19. However, a docking study is required to light this further, stating how well Fostamatinib/R406 binds with the SARS-CoV-2 proteins. One of the most powerful approaches for structure-based drug discovery is molecular docking. It is defined as analyzing how more than one molecular structure, *i.e.* drug and protein or enzyme gets attached [388]. In other words, docking can be interpreted as a molecular modeling methodology, which is implemented to anticipate how small molecules, *i.e.*, ligands, interrelate with proteins, *i.e.*, enzymes. But to do docking, protein structures of both SARS-CoV-2 proteins and Fostamatinib/R406 are required. So, protein-ligand docking is executed by using Molegro Virtual Docker (version: 6.0) on potential COVID-19 FDA-approved drugs, Fostamatinib and R406, with all the so far available protein structures on nCoV having PDB IDs: 6LU7 [368], 6M2Q [369], 6W9C [370], 6M0J [165], 6M71 [371], 6VXX [372]. Grid-based cavity prediction is used to identify the potential binding sites. Models involving flexible ligands are taken into consideration. The orientation of ligands usually differs, and ranking for each ligand is based on the energy scores. The entire algorithm is implemented at 1500 iterations with a simplex evolution size of 10 runs. Compounds that take the lowest binding energy in comparison to others are considered to be the best. The molecules of the potential COVID-19 FDA-listed drugs are downloaded from DrugBank [330] in the Structure data file (SDF/PDB) format. These scores assist in the identification of the best molecules docked in the selected target site. All the molecules are sorted based on these scores, representing the lowest energy required to get tied up with amino acid (AA) components. The docking returns two types of scores:

Moldock scores: MolDock is based on a new heuristic search algorithm that combines differential evolution with a cavity prediction algorithm. The docking scoring function of MolDock is an extension of the piecewise linear potential (PLP), including new hydrogen bonding and electrostatic terms. As a result, MolDock has a very high docking accuracy for the identification of ligand-binding modes [389]

Rerank scores: The re-rank score is a linear combination of E-inter (steric, Van


Covid-19 Symptoms
Heart Disease (D_1)
Cough (D_2)
Thrombocytopenia (D_3)
Fever (D_4)
....
....
Lung Disease (D_{n-1})
Diabetes (D_n)
Step-1
Step-2
DisGeNET
(Used for mapping of genes with symptoms)
Gene set associated with Heart Disease (G_{D_1})
Gene set associated with Cough (G_{D_2})
Gene set associated with Thrombocytopenia (G_{D_3})
Gene set associated with Fever (G_{D_4})
....
....
Gene set associated with Lung Disease ($G_{D_{n-1}}$)
Gene set associated with Diabetes (G_{D_n})
Step-3
Molbiotools
(To form common set of genes)
Step-4
$CG_{D_{1....n}} = G_{D_1} \cap G_{D_2} \cap G_{D_3} .... \cap G_{D_n}$
Step-5
Comparative Toxicogenomics Database (CTD)
$CTD_{Covid-19\_GeneSet} = \{G_1, G_2 .... G_{n-1}, G_n\}$
Curated COVID19 dataset under the category of respiratory tract disease & viral disease
Step-6
$Final_{Overlapping\_GeneSet} = CTD_{Covid-19\_GeneSet} \cap CG_{D_{1....n}}$
Step-7
$Final_{Overlapping\_GeneSet}$ is further intersected with the spreader protein set in level-1 and level-2 of generated Human-nCoV interaction network.
Step-8
Human-nCoV interaction network

Final Output: The selected top genes after intersection from level-1 are PPIA, ACE2, EIF3F, UBC, PRKDC, CDK2, CDK1, AKT1, PRKCA, and TRAF6 while that of level-2 are APP, ELAVL1, NTRK1, XPO1, MEOX2, GRB2, EGFR, TP53, BAG3 and NXF1. After applying the PDS_CS algorithm on these genes, Fostamatinib/R406 has a significant overlap of 3 target proteins which is also the highest frequency of occurrence.

Figure 5.3: COVID-19 Symptoms based analysis. The analysis consists of the following steps. **Step-1:** Symptoms are searched in the DisGeNET database. **Step-2**: Symptoms associated gene sets are fetched from the DisGeNET database. **Step-3:** All gene sets are provided as input to Molbiotools online. **Step-4:** Common overlapping sets of genes are obtained from Molbiotools. **Step-5:** Curated COVID-19 dataset is extracted from the CTD under respiratory tract disease & viral disease. **Step-6:** These sets of genes intersect with the common overlapping set of genes obtained from Molbiotools to form a key set of COVID-19 related genes. **Step-7**: This key set of COVID-19 related genes, after mapping to their corresponding protein IDs, are compared with the spreader proteins in human-nCoV PPIN. **Step-8:** After comparison, top genes are selected in both level-1 and level-2 of human-nCoV PPIN, which are finally used for Algorithm 5.1 to determine the most potential candidate FDA-approved drug for COVID-19. Here, Fostamatinib/R406 has the highest DCS.

der Waals, hydrogen bonding, electrostatic) between the ligand and the protein and E-intra. (torsion, sp2-sp2, hydrogen bonding, Van der Waals, electrostatic) of the ligand weighted by pre-defined coefficients. The re-ranking procedure is adequate for identifying high-quality binding modes in place of more advanced scoring schemes [389]

5.4 Experimental Results

Computational studies and results of associated drugs with human proteins in human-nCoV PPIN show that there is a probability that Fostamatinib/R406 may act as one of the potential candidates for COVID-19 treatment.

5.4.1 Drug Consensus Results for COVID-19 Spreader Nodes Using Human-nCoV Interaction Network Analysis

Drugs and their corresponding IDs are mapped with all human spreader proteins in human-nCoV PPIN by matching the related drug-protein targets with spreader nodes. Using Algorithm 5.1, it is seen that Fostamatinib/R406 has the highest frequency of occurrence in the whole PPIN among all human proteins, with an overlap of 155 target proteins in human-nCoV PPIN, *i.e.*, highest DCS as discussed earlier in section 5.3. It has a DCS score of 7, *i.e.*, count of level-1 protein targets of Fostamatinib as shown in Table 5.3, and 148, *i.e.*, count of level-2 protein targets of Fostamatinib as shown in Table 5.4, in level-1 and level-2 human spreader proteins. This establishes that the algorithm has succeeded in detecting the appropriate drug molecules with the highest protein targets in both levels. Protein targets corresponding to the DCS score of Fostamatinib in level-1 is highlighted in Figure 5.4, while that of level-2 is shown in Figure 5.5. In Figure 5.4, green nodes represent level-1 protein targets of Fostamatinib, while blue and yellow nodes denote COVID-19 and other level-1 human proteins, respectively. In Figure 5.5, green nodes represent level-2 protein targets of Fostamatinib, while blue and yellow nodes denote COVID-19 and other level-1 spreader human proteins, respectively. Other level-2 human spreaders in Figure 5.5 are not shown to avoid visual complexity. The highest frequency of Fostamatinib/R406 is observed when Algorithm 5.1 is implemented on human-nCoV PPIN of Gordon et al. [164]

5.4.2 Drug Consensus Results for COVID-19 Spreader Nodes Using COVID-19 Symptoms, Risk Factors and Clinical Outcome-based Analysis

Grouping genes based on various categories of COVID-19 symptoms, risk factors, and clinical outcomes [225] is done using DisGeNET [377]. The numerical statistics of

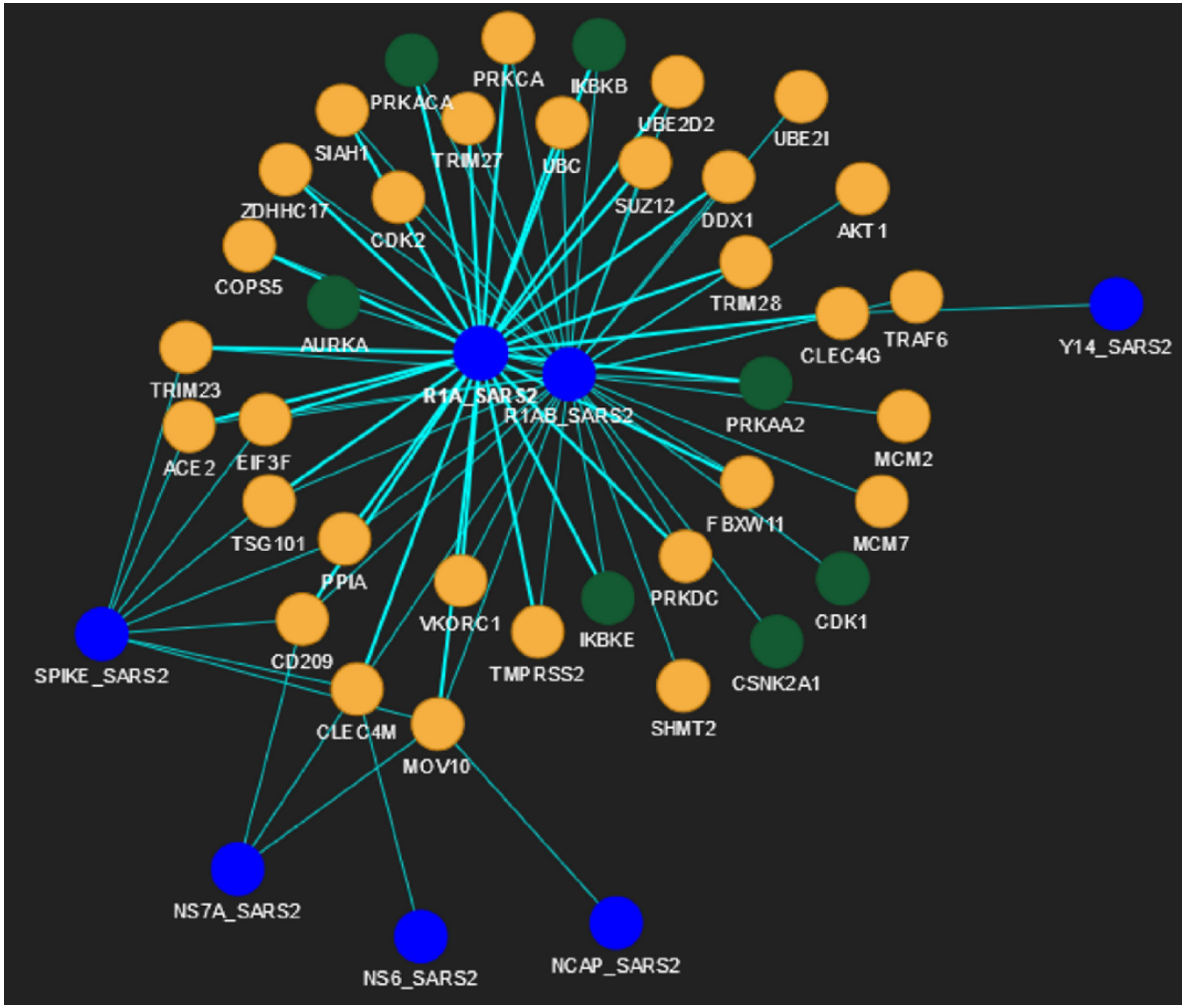


Figure 5.4: Application of Algorithm 5.1 in Human-CoV PPIN (level-1). DCS score of Fostamatinib/R406 in level-1 is 7 (green nodes). Blue and yellow nodes denote COVID-19 and other level-1 proteins, respectively.

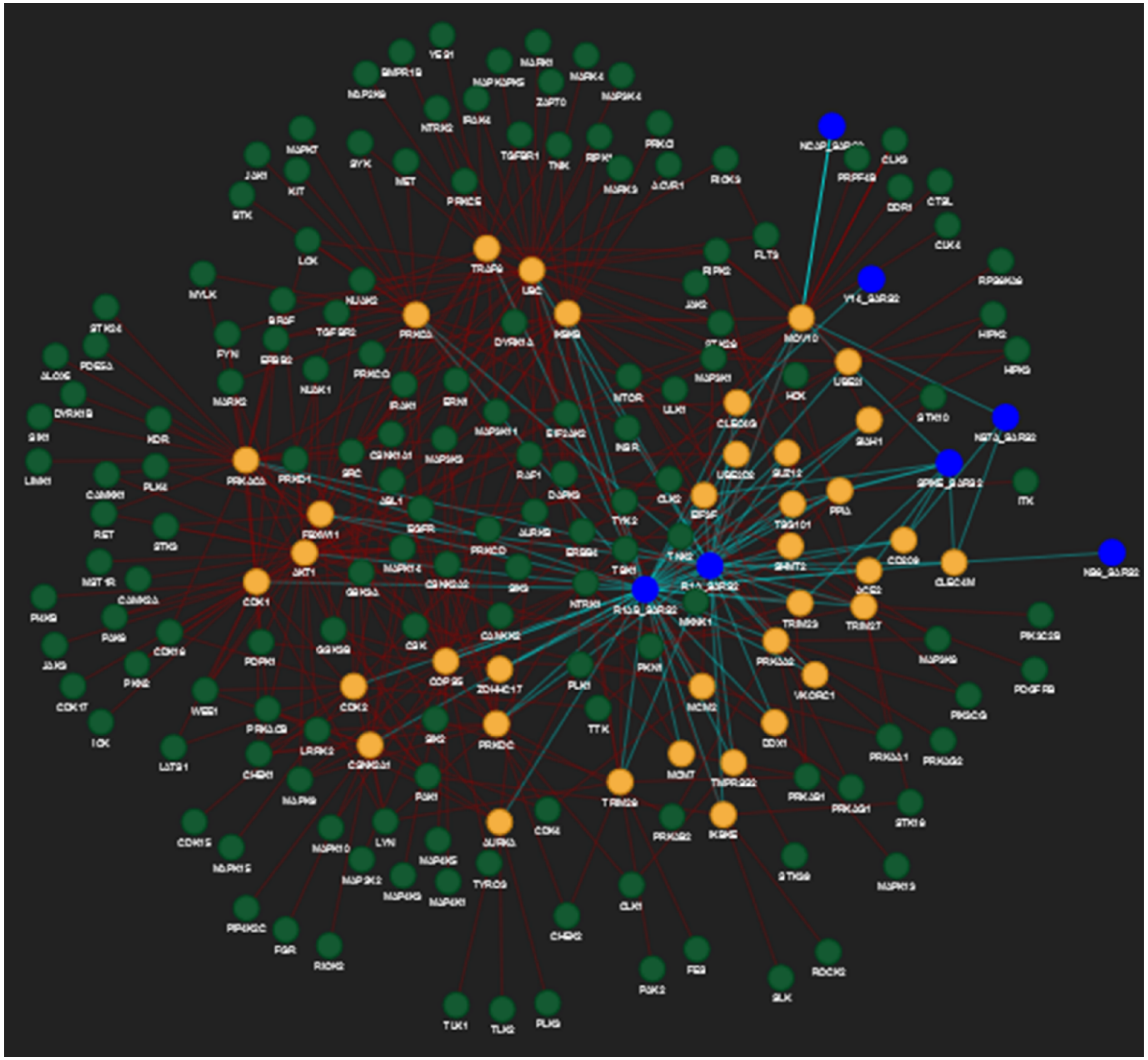


Figure 5.5: Application of Algorithm 5.1 in Human-CoV PPIN (level-2). DCS score of Fostamatinib/R406 in level-2 is 148 (green nodes). Blue and yellow nodes denote COVID-19 and other level-1 spreader proteins, respectively. Other level-2 spreaders are not shown to avoid visual complexity.

Table 5.3: Detailed analysis of DCS score at level-1 (Top 5 DCS have been shown)

Drug	Drug ID	DCS (level-1)
Fostamatinib/R406	DB12010	7
Arsenic trioxide	DB01169	2
Acetylsalicylic acid	DB00945	2
Resveratrol	DB02709	2
Tamoxifen	DB00675	1

Table 5.4: Detailed analysis of DCS score at level-2 (Top 5 DCS have been shown)

Drug	Drug ID	DCS (Level-2)
Fostamatinib/R406	DB12010	148
Copper	DB09130	88
Zinc acetate	DB14487	57
Zinc	DB01593	57
Zinc chloride	DB14533	57

the result are highlighted in Table 5.5. Mobiotools [390] are used to compare these gene sets to obtain 4931 unique genes. These genes are further compared with the curated COVID-19 dataset of CTD [378], containing 12672 genes. The comparison generates an overlapping gene set containing 3525 genes. When used for validation against the spreader proteins in the human-nCoV interaction network, these genes produce a significant overlap of 1448 genes in both level-1 and level-2. This highlights the fact that 1448 out of 3525 genes are selected as spreader nodes in the network. Hence, the top 10 key genes are selected from 1448 in each level based on the fuzzy and spreadability index scores in level-1 and level-2. The selected top genes from level-1 are PPIA, ACE2, EIF3F, UBC, PRKDC, CDK2, CDK1, AKT1, PRKCA, and TRAF6 level-2 are APP, ELAVL1, NTRK1, XPO1, MEOX2, GRB2, EGFR, TP53, BAG3 and NXF1. Potential FDA-approved drugs having these key genes/spreader proteins as known targets are identified from DrugBank data. It is also observed that after applying Algorithm 5.1 on the obtained result, Fostamatinib/R406 has a significant overlap of 3 target proteins which is also the highest frequency of occurrence. Similarly, for the symptom *"loss of smell"*, 12 overlapping genes are detected, and mapping with known drug targets in DrugBank has also been done as shown in Table 5.6. After applying Algorithm 5.1, only Fostamatinib/R406 and copper emerge based on their frequency of occurrence.

Table 5.5: Few statistical analyses of genes in COVID-19 symptoms, risk factors and clinical outcome

Categorizations	Symptoms	Total no. of genes
COVID-19 symptoms [225]	Cough	270
	Fever	1743
	Dyspnea	323
	Pneumonia	1416
Risk factors [225]	Heart Disease	1964
	Kidney Disease	2131
	Lung Disease	1018
	Diabetes	5078
	Hypertension	1573
	Cancer	4747
Clinical Outcomes [225] (Mild & Moderate Case)	Lymphopenia	241
	Pulmonary infiltrate	43
Clinical Outcomes (Severe Case) [225]	Leukocytosis	179
	Neutrophilia	152
	Sepsis	1506
	Kidney injury	228
	Coagulopathy	21
	Thrombocytopenia	774
	Multiple organ failure	25

5.4.3 Docking Results for Potential COVID-19 Drugs *w.r.t* COVID-19 Protein Structures

Molecular docking is used in the proposed methodology to measure the binding capability of the potential COVID-19 drugs on 6LU7, 6M2Q, 6W9C, 6M0J, 6M71 and 6VXX. The detailed procedure of execution has been already discussed in the methodology section. In this work, SDF/PDB format is used for the docking of all the COVID-19 drugs. The details of Best dock poses for potential COVID-19 drugs and interactions of hydrogen bonds with respect to 6LU7 are given in Table 5.7. The Moldock schore and Rerank score are depicted in Table 5.8. At the same time, docking results with others are shown in Table 5.9 and Table 5.10. It is observed from the results that

Table 5.6: Mapping of FDA-approved drug of DrugBank with selected key genes of level-2 associated with *"loss of smell"* symptom of COVID-19

Genes in L2 n-CoV	Target drugs	DrugId	Approved
DCC 2	N/A	N/A	N/A
EIF4G1	N/A	N/A	N/A
GIGYF2	N/A	N/A	N/A
HTRA2	N/A	N/A	N/A
LRRK2	Fostamatinib	DB12010	TRUE
PARK7	Copper	DB09130	TRUE
PINK1	N/A	N/A	N/A
PODXL	N/A	N/A	N/A
PTPN11	Dodecyltrimethylammonium	DB02779	FALSE
SNCA	Copper	DB09130	TRUE
UCHL1	Phenethyl Isothiocyanate	DB12695	FALSE
VPS35	N/A	N/A	N/A

while Fostamatinib/R406 registers the highest score for 6LU7 and 6M2Q, it obtains the second position in comparison to the other COVID-19 structures.

5.4.4 Docking Results for Active Metabolites/Promoieties of COVID-19 Prodrugs *w.r.t* COVID-19 Protein Structures

Several drug molecules consist of pharmacologically inactive compounds, which are known as Prodrugs [391]. These drugs get metabolized after entering the human body to liberate the active drug. On careful observation, it has been observed that Fostamatinib is also a prodrug. Fostamatinib (R788) is considered to be an orally induced prodrug in humans that releases active metabolite/promoiety R940406 (R406) [392]. R406 is a spleen tyrosine kinase (SYK) inhibitor responsible for treating rheumatoid arthritis [392]. Similar instances have also been observed in the case of remdesivir and favipiravir. So, the binding capability of these active metabolites/promoieties must be validated against 6LU7, 6M2Q, 6W9C, 6M0J, 6M71, and 6VXX by molecular docking for consideration of any prodrug as a COVID-19 drug. The results of this docking with 6LU7 are highlighted in Table 5.11 and Table 5.12. The result draws the reference that R406 also shows high binding affinity scores compared to the others, which promotes the fact that Fostamatinib/R406 can be a potential COVID-19 drug. Molecular docking results of Fostamatinib and its corresponding promoiety, R406, have also been highlighted in Figure 5.6.

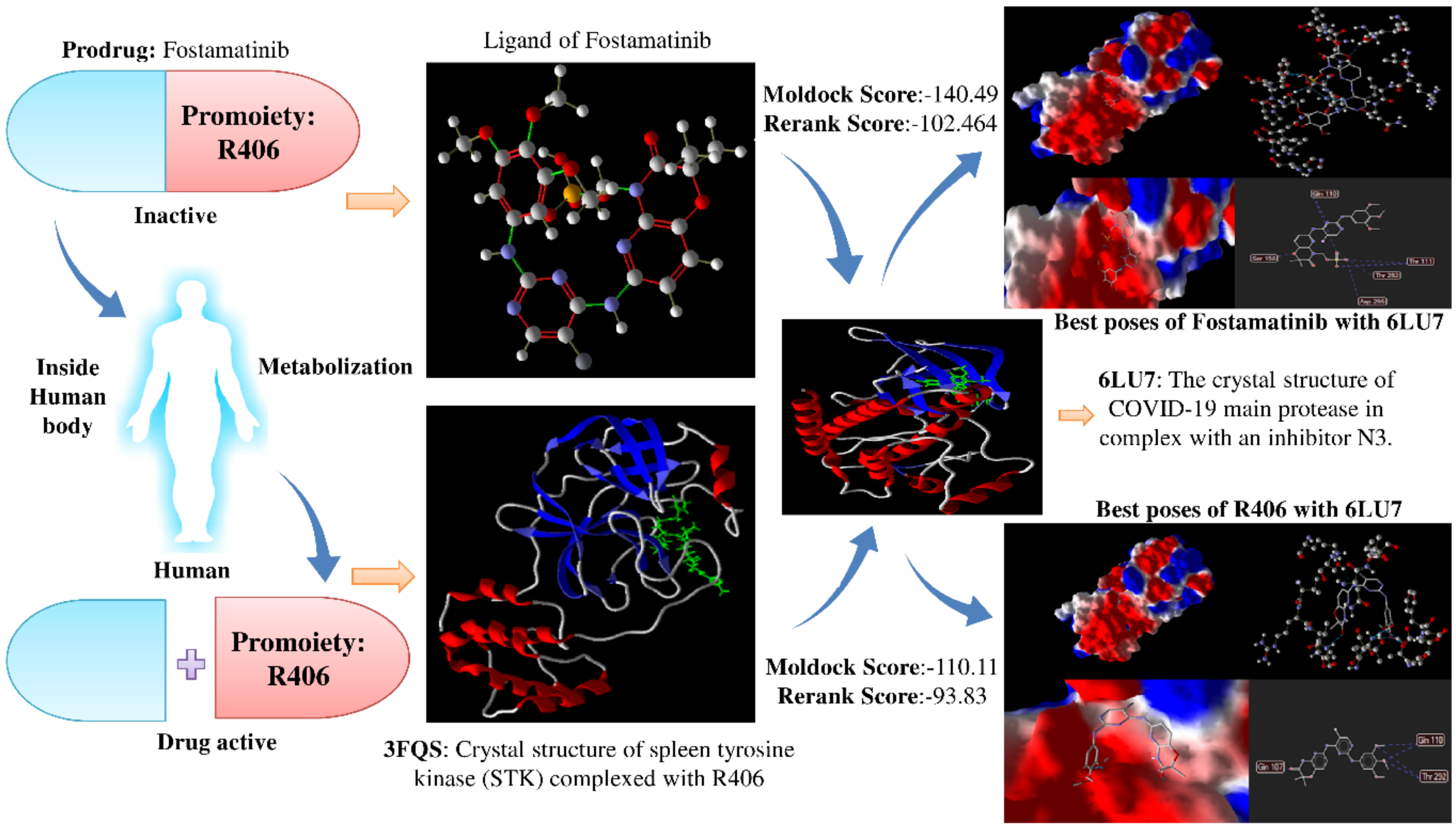

Prodrug: Fostamatinib
Promoiety: R406
Inactive
Inside Human body
Metabolization
Human
Promoiety: R406
Drug active
Ligand of Fostamatinib
3FQS: Crystal structure of spleen tyrosine kinase (STK) complexed with R406
Moldock Score:-140.49
Rerank Score:-102.464
Moldock Score:-110.11
Rerank Score:-93.83
Best poses of Fostamatinib with 6LU7
6LU7: The crystal structure of COVID-19 main protease in complex with an inhibitor N3.
Best poses of R406 with 6LU7

Figure 5.6: Molecular docking results of prodrug Fostamatinib and its corresponding promoiety, R406. Fostamatinib and R406 both have high binding affinity scores in comparison to the other potential COVID-19 drugs.

Table 5.7: Best dock poses for potential COVID-19 drugs and interactions of hydrogen bonds with respect to 6LU7

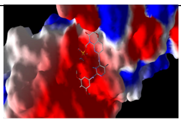
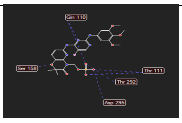
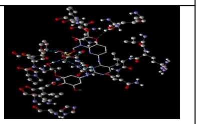
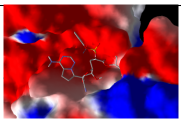
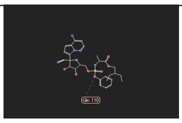
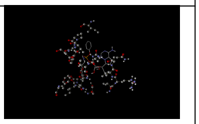
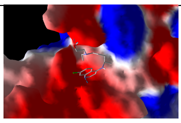
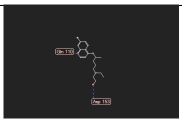
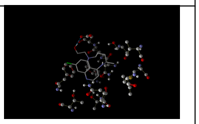
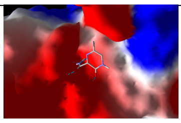
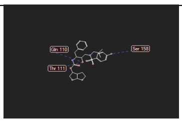
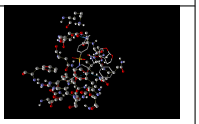
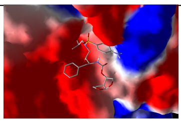
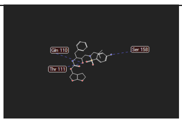
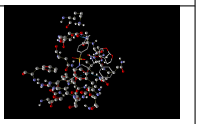
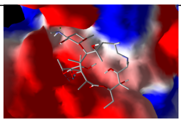
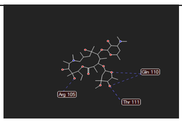
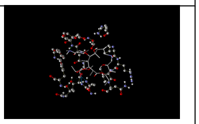
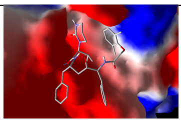
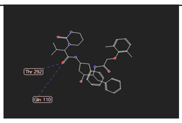
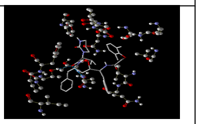
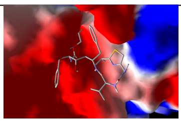
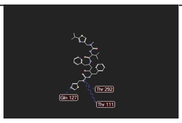
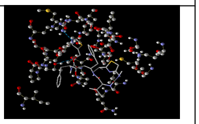
Drugs	**Drug ID**	**Best docked poses**	**H-bond interaction details best pose**	**H-bond interaction best pose**
Fostamatinib	**DB12010**			
Remdesivir	DB14761			
*HCQ**	DB01611			
Favipiravir	DB12466			
Darunavir	DB01264			
Azithromycin	DB00207			
Lopinavir	DB01601			
Ritonavir	DB00503			

Table 5.8: Moldock Score and Rerank Score for potential COVID-19 drugs and interactions of hydrogen bonds with respect to 6LU7

Drugs	**Drug ID**	**Moldock score**	**Rerank Score**
Fostamatinib	**DB12010**	**-140.495**	**-102.464**
Remdesivir	DB14761	-134.19	-56.312
Hydroxychloroquine	DB01611	-106.266	-69.417
Favipiravir	DB12466	-62.855	-55.371
Darunavir	DB01264	-128.798	-80.316
Azithromycin	DB00207	-86.77	29.53
Lopinavir	DB01601	-83.09	-30.19
Ritonavir	DB00503	-110.36	103.40

Table 5.9: Docking scores (Moldock Score) of drugs for 6VXX, 6M71, 6M2Q, 6W9C, 6M0J

Drugs	**Drug ID**	**6VXX**	**6M71**	**6M2Q**	**6W9C**	**6M0J**
Ritonavir	DB00503	-228.318	-201.076	-121.515	-230.852	-197.609
Fostamatinib	DB12010	-172.859	-179.834	-120.397	-185.156	-176.483
Lopinavir	DB01601	-166.217	-166.388	-114.563	-181.474	-170.136
Remdesivir	DB14761	-152.678	-160.739	-112.938	-172.778	-165.157
Darunavir	DB01264	-144.579	-152.622	-90.9735	-156.453	-152.907
HCQ	DB01611	-127.194	-110.183	-80.4403	-132.054	-122.721
Azithromycin	DB00207	-120.155	-105.759	-66.3967	-127.711	-115.321
Favipiravir	DB12466	-76.41	-73.7678	-14.436	-72.6467	-70.8629

Table 5.10: Docking scores (Rerank Score) of drugs for 6VXX, 6M71, 6M2Q, 6W9C, 6M0J

Drugs	**Drug ID**	**6VXX**	**6M71**	**6M2Q**	**6W9C**	**6M0J**
Ritonavir	DB00503	-154.226	-140.533	-92.2327	-160.902	-130.098
Fostamatinib	DB12010	-57.5164	-110.545	-87.5182	-130.02	-130.509
Lopinavir	DB01601	-81.6619	-138.449	12.9996	-147.683	-125.165
Remdesivir	DB14761	-101.76	-124.925	-84.1368	-108.703	-119.626
Darunavir	DB01264	-103.522	-122.378	-66.1872	-115.516	-73.643
HCQ	DB01611	-100.612	-88.6121	18.5798	-102.419	-97.8097
Azithromycin	DB00207	-88.4657	-72.8345	-54.8591	-100.182	-31.7329
Favipiravir	DB12466	-62.055	-52.7567	793.471	-59.1369	-57.8456

Table 5.11: Docking scores (Moldock Score) of Prodrugs for 6VXX, 6M71, 6M2Q, 6W9C,6M0J

Prodrugs	**Active promoieties**	**6VXX**	**6M71**	**6M2Q**	**6W9C**	**6M0J**
Fostamatinib	RP406 (using 3FQS)	-143.34	-131.064	-115.229	-150.184	-134.057
Remdesivir	GS-441524	-115.54	-106.624	-106.108	-107.333	-120.417
Favipiravir	RdRp complex (6K32)	-92.5709	-100.143	-54.3239	-93.9083	-71.9692

5.4.5 Analysis of 3 Key Target Genes of Fostamatinib/R406 in Human-nCoV Interaction Network in Symptom-based Analysis

The three key target genes of Fostamatinib/R406, as identified in Table 5.13 and Table 5.14, are CDK1 (level-1) and NTRK1, EGFR (level-2). It is noted that these three

Table 5.12: Docking scores (Rerank Score) of Prodrugs for 6VXX, 6M71, 6M2Q, 6W9C,6M0J

Prodrugs	Active promoieties	6VXX	6M71	6M2Q	6W9C	6M0J
Fostamatinib	RP406 (using 3FQS)	-115.106	-109.152	-76.3059	-111.853	-111.887
Remdesivir	GS-441524	-91.7647	-80.5349	-70.1108	-87.910	-94.3608
Favipiravir	RdRp complex (6K32)	-66.4667	-84.9671	-49.3377	-81.8392	-57.9054

Table 5.13: Mapping of FDA-approved drug of DrugBank with selected key genes of level-1

Level-1 Key Genes	Approved/Approved & Investigational Drug	
	Drug	Drug ID
PPIA	Cyclosporine	DB00091
	Copper	DB09130
ACE2	Hydroxychloroquine	DB01611
	Chloroquine	DB00608
EIF3F	No approved drug	
UBC	No approved drug	
PRKDC	Caffeine	DB00201
CDK2	Bosutinib	DB06616
CDK1	*Fostamatinib*	DB12010
PRKCA	D-alpha-Tocopherol acetate	DB14002
	Midostaurin	DB06595
	alpha-Tocopherol succinate	DB14001
	Phosphatidyl serine	DB00144
	Vitamin E	DB00163
	Tamoxifen	DB00675
	Ingenol mebutate	DB05013
AKT1	Arsenic trioxide	DB01169
TRAF6	No approved drug	

genes are related to the most significant COVID-19 symptoms, risk factors, and clinical outcomes, which are highlighted in Table 5.15. Moreover, these three genes also play an essential role in response to viral infections (*see* Table 5.16). All these depict the fact that Fostamatinib/R406 might be a potential drug treatment for COVID-19 treatment.

Table 5.14: Mapping of FDA-approveddrug of DrugBank with selected key genes of level-2

Level-1 Key Genes	**Approved/Approved & Investigational Drug**	
	Drug	**Drug ID**
APP	Aluminium phosphate	DB14517
	Dimercaprol	DB06782
	Copper	DB09130
	Florbetapir (18F)	DB09149
	Flutemetamol (18F)	DB09151
	Deferoxamine	DB00746
	Zinc	DB01593
	Zinc sulfate, unspecified form	DB14548
	Florbetaben (18F)	DB09148
	Zinc acetate	DB14487
	Aluminum acetate	DB14518
	Aluminium	DB01370
	Zinc chloride	DB14533
ELAVL1	No approved drug	
NTRK1	Entrectinib	DB11986
	Fostamatinib	DB12010
	Cenegermin	DB13926
	Amitriptyline	DB00321
	Imatinib	DB00619
	Regorafenib	DB08896
	Larotrectinib	DB14723
XPO1	Selinexor	DB11942
MEOX2	No approved drug	
GRB2	Pegademase	DB00061
EGFR	Lidocaine	DB00281
	Gefitinib	DB00317
	Fostamatinib	DB12010
	Zanubrutinib	DB15035
	Cetuximab	DB00002
	Erlotinib	DB00530
	Vandetanib	DB05294
	Osimertinib	DB09330
	Dacomitinib	DB11963

5.4.6 Application of Algorithm 5.1 on the Host Targets of *in-vitro* Generated Human-nCoV PPIN of Gordon et al. [164]

Gordon *et al.* [164] cloned, tagged, and expressed 26 of the 29 SARS-CoV-2 proteins in human cells and identified the human proteins that are physically associated with each

Table 5.15: Mapping of CDK1 and NTRK1, EGFR with COVID-19 symptoms, risk factors and clinical outcomes

Drug	**Key Target Genes**	**Level**	**COVID-19 symptoms**	**Clinical outcome (severe case)**	**Risk Factor**
Fostamatinib	CDK1	Level-1	pneumonia	−−	diabetes
					cancer
	NTRK1	Level-1	fever	−−	kidney disease
					cancer
					hypertension
			pneumonia	−−	lung disease
					diabetes
	EGFR	Level-2	pneumonia	neutrophilia	heart disease
					hypertension
					cancer
			dyspnea	neutrophilia	kidney disease
			fever	kidney injury	lung disease
			cough	thrombocytopenia	diabetes

Table 5.16: Role of CDK1 and NTRK1, EGFR in viral infections

Drug	**Key Target Genes**	**Level**	**Role in viral infections**
Fostamatinib	CDK1	Level-1	Viruses can express some oncoproteins. Genome replication gets induced inside the hosts' cells due to some signals generated due to the interference of these proteins with CDK and CIKs function [391].
	NTRK1	Level-2	NTRK1 has an active role in the immune response against viral infection [392]
	EGFR	Level-2	Hindrance of EGFR signaling might prevent an excessive fibrotic response to SARS-CoV.

of the SARS-CoV-2 proteins by affinity-purification mass spectrometry. As a result, 332 high-confidence PPI between SARS-CoV-2 and human proteins are identified. These 332 host targets are collected, and Algorithm 5.1 is implemented on the same. It is observed from the implementation that Fostamatinib/R406 has a significant overlap of 10 target proteins (i.e., DCS of 10) in human-nCoV PPIN, which is also the highest

Table 5.17: Detailed analysis of DCS score (Top 6 DCS have been shown)

Drug	Drug ID	DCS (Level-1)
Fostamatinib	DB12010	10
NADH	DB00157	5
Flavin adenine dinucleotide	DB03147	5
Romidepsin	DB06176	2
Glutamic acid	DB00142	2
Atorvastatin	DB01076	1

frequency of incidence across the entirety of the PPIN when contrasted with the other human proteins related to drugs. The result is highlighted in Table 5.17.

The ten host target genes associated with Fostamatinib/R406 in Table 5.17 are TBK1, CIT, NEK9, RIPK1, COQ8B, CSNK2A2, MARK1, MARK3, MARK2 and PRKACA. In addition, on careful observation, it has been noted that these genes also play an essential role in response to viral infections, which has been discussed below:

1. **TBK1:** TBK1 (TANK-binding kinase 1) plays a highly significant role in developing natural immunity against antiviral activities. It activates IRF (interferon regulatory factor) 3, which in turn induces type I interferon (IFNs) (IFN-α/β) proteins regulating immune activity [393].

2. **CIT:** Encoded serine/threonine protein kinases are unique features in a specific set of giant DNA viruses. However, their role in the replication of virus vary. But different viral serine/ CIT (Citron Rho-Interacting Serine/Threonine Kinase) has the potential to act as the targets of antiviral drugs [363].

3. **NEK9:** Nek9 exhaustion leads to the reduction of virus replication centers within which it remains confined. However, Nek9 overexpression will increase the number of viral genomes in the infected cell [394].

4. **RIPK1:** Enhancement of plasma pro-inflammatory cytokines and lymphopenia is considered to be one of the significant predictors in increasing COVID-19 severity. Activating RIPK1 promotes the growth of these cytokines. In addition, it leads to the exhaustion of T cell populations (lymphopenia) in patients who get infected with HIV, which might pave the way for the entrance of SARS-CoV-2 in them [395].

5. **COQ8B:** Mitochondrial metabolism is executed as a part of the metabolic pathway through the interaction of SARS-CoV-2's M protein and COQ8B [396].

6. **CSNK2A2:** CSNK2A2 is involved in the regulation of primary cellular processes as well as viral infection [397].

7. **MARK1:** MARK1 plays an active role in viral responses [398]. .

8. **MARK3:** MARK3 controls cardiovascular disease, which is considered one of the significant symptoms of COVID-19 [399].

9. **MARK2:** MARK2 is engaged in stimulating FEZ1 (Fasciculation And Elongation Protein Zeta 1) phosphorylation on the central cores of viruses [400].

10. **PRKACA:** PRKACA also plays a similar role as that of MARK3 [401]

5.5 Discussion

In this computational study, the human-nCoV PPIN has been analyzed and attempted to identify the candidate drugs for the level-1 and level-2 spreader proteins. The study identifies Fostamatinib/R406, an FDA-approved drug, as the most promising drug with the best chances to target the COVID-19 spreader proteins. The work relies on the hypothesis that SARS-CoV-2/nCoV has 89% genetic resemblance with SARS-CoV. Based on this, human-nCoV PPIN has been developed, and its spreader nodes have been identified using the SIS model and fuzzy thresholding. Furthermore, a consensus strategy by a two-way analysis has been utilized to analyze drugs based on the overlap of spreader proteins and drug-protein targets. The consensus scores for Fostamatinib/R406 are the highest in analyzing the candidate drugs for COVID-19 spreader proteins. Besides, Fostamatinib/R406 also generates satisfactory results in molecular docking with the available COVID-19 protein structures. It also targets CAYP34A [366, 402], a common target for almost all the FDA-approved drugs [227] for COVID-19. Moreover, recent studies also suggest that it is used for thrombocytopenia [366] which is also associated with COVID-19 infections [367]. A clinical test is needed as the FDA approves Fostamatinib/R406 in ITP [403] and to determine its efficacy against SARS-CoV-2. Rigel Pharmaceuticals have already started the clinical trials of Fostamatinib/R406 [371]. The results obtained are quite encouraging and positive regarding reports published to date [370, 403]. According to the report [370, 403], Fostamatinib meets the *"primary endpoint of Safety in Phase 2 Clinical Trial"* conducted in hospitalized patients affected with COVID-19. In addition to this, they have also enrolled themselves for a Phase 3 clinical trial of fostamatinib/R406 to treat the same. But arriving at a specific conclusion needs time and more research analysis. In a nutshell, our computational research evidence discovers that Fostamatinib/R406 may be considered one of the strong contenders for COVID-19 treatment.

This chapter represents an *in-silico* study on repurposing of drugs for COVID-19 treatment. Identified spreader proteins for both the levels *i.e.* level-1 and level-2 which are validated from SIS model are used as target proteins for re-purposed drugs. Still, now, it has only been studied how nCoV interacts with human proteins, and suitable repurposed drugs are identified from the target proteins. But not only nCoV but other COVID variants also interact with human proteins. In the following chapter, an assessment has been done to identify the GO-based interaction affinities between the human-coronavirus family interactome. The chapter also identifies level-1 potential spreader proteins to find out FDA-approved drugs which are used to treat COVID related diseases.

Chapter 6

Assessment of GO-based Protein Interaction Affinities in the Large-Scale Human-Coronavirus Family Interactome

6.1 Background

The emerging coronavirus pandemic has Sparked a flurry of research into the SARS-CoV-2 virus and the COVID-19 disease it causes in people [404]. As discussed in Chapter 4, COVID-19 was identified in Wuhan (Hubei province) [299] and was soon declared as global emergency by WHO . A coronavirus is a member of the family Coronaviridae. SARS-CoV-2 is a novel coronavirus that replicates itself by interacting with the host proteins. As a result, identifying virus and host PPIs could help researchers better understand the virus disease transmission behavior and identify possible COVID-19 drugs. Along with humans, it also affects mammals and birds. Even though the coronavirus typically causes the common cold, cough, etc., it also causes severe acute, chronic respiratory disease, multiple organ failure, and, ultimately, human mortality.

Apart from SARS-CoV-2, coronavirus family have 44 different variants. . Based on the availability of the GO annotation of the proteins, 11 viral variants, *viz.*, SARS-CoV-2, SARS-CoV, MERS-CoV, Bat coronavirus HKU3, Bat coronavirus Rp3/2004, Bat coronavirus HKU5, Murine coronavirus (M-CoV), Bovine coronavirus (BCoV), Rat coronavirus (RCoV), Bat coronavirus HKU4, Bat coronavirus 133/2005, are considered from 44 viral variants. Before SARS-CoV-2, the two primary outbreaks were MERS-CoV and Severe Acute Respiratory Syndrome (SARS). Southern China was the location of SARS's inception. Its fatality rate was between 14 and 15% [297]. The MERS-CoV outbreak was supposed to start in Saudi Arabia. In the fight against the MERS-CoV virus, 858 out of 2494 afflicted cases prevailed. As a result, it produced a substantially higher death rate of 34.4% compared to the SARS-CoV.

Regarding biology, the three epidemic-starting viruses, SARS, MERS, and SARS-CoV-2, belong to Coronaviridae's genus Beta coronavirus. Proteins that are both structural and non-structural contribute to the development of SARS-CoV-2. Out of the two, structural proteins such as the spike protein, nucleocapsid protein, membrane protein, and envelope protein play a crucial part in spreading the disease by binding with receptors after entering the human body [304].

The primary factor that needs to be considered while examining the disease transmission process from SARS-CoV-2 to humans is PPIN. It is critical for determining essential proteins and functions [137, 152, 153, 157, 158, 180, 305–307, 405–409] responsible for various diseases. The primary focus of research has changed from the study of the PPIN underlying various types of human diseases to the study of the PPIN due to the improvement in the availability of human PPIN data [309]. According to the report, SARS-CoV-2 has $\sim 89\%$ similarity with SARS-CoV [312, 410]. SARS-CoV, a disease that initially appeared in the Guangdong Province of China in November 2002, spread to 28 regions worldwide in 2003 and resulted in 774 fatalities among the 8096 people with COVID-19 [411, 411]. According to phylogenetic analysis, it was assumed that SARS-CoV was different from previously known coronaviruses [412, 413]. Even though the etiological agent was discovered and molecular research on the SARS-CoV advanced quite quickly, the mystery surrounding the disease's cause remained unsolved. Data indicated that SARS-CoV was an animal-borne disease from the beginning [411, 414, 415]. After the surge of SARS-CoV in 2012, there was another coronavirus surge, MERS-CoV, in Jordon. A bat and numerous dromedary camels have been reported to have MERS-CoV sequences. MERS-CoV is an enzootic disease in the Arabian Peninsula, portions of Africa, and the Middle East. It affects camels as its primary reservoir and occasionally, but infrequently, infects humans [416]. MERS-CoV is a member of the Beta coronavirus family. WHO confirmed 2220 people with COVID-19 along with 790 deaths for MERS-CoV [417]. There is a 35% fatality rate from MERS. MERS-CoV is not specifically treated. MERS-CoV outbreaks in hospitals and homes are brought on by person-to-person transmission [418]

A beta-CoV prevalent in wild mice, the MHV or M-CoV is similar to SARS-CoV-2. In-depth research has been done on laboratory MHV strains to understand host antiviral defense systems and coronavirus virulence factors [419]. Murine-CoV contains several strains that induce variable symptoms in the respiratory, digestive, hepatic, and neurological systems [420–422]. The genus of beta-CoVs includes all MHV strains and certain human CoVs *viz.* HCoV-OC43, HCoV-HKU1, SARS-CoV, MERS-CoV, and SARS-CoV-2. The tropism and pathogenicity of various MHV strains vary, and research on recombinant MHV variations has uncovered host and viral variables that affect viral propagation or evade immune Identification [423].

The wide variety of mammalian and avian species that coronaviruses have been found to infect and the highly varied disease syndromes they cause are well known. One of the well-known traits of several coronaviruses is variable tissue tropism, which also allows them to overcome inter-species boundaries easily. Beta-coronaviruses, known as BCoVs, cause shipping fever, winter dysentery in older cattle, and neonatal calf diarrhea. Interestingly, there have not been any specific genetic or antigenic markers found in BCoVs linked to these unique clinical disorders. BCoVs, on the other hand, are quasispecies that coexist with other coronaviruses. In addition to cattle, BCoVs and coronaviruses resembling cattle were found in several domestic and wild ruminant species, dogs, and humans [424]. The pneumoenteric virus known as the BCoV is a member of the Betacoronavirus genus. Because of several instances of genetic recombination and interspecies transmission, members of the Betacoronavirus species appear to be host-range variants descended from the same parental virus due to their close antigenic and genetic relatedness [425–428]

Two separate teams reported finding SARS-like CoVs (SL-CoVs) in bats in 2005, and they hypothesized that bats were SARS-CoV natural reservoirs [429, 430]. Most bat SL-CoVs were discovered in rhinolopus bats, especially Rhinolophus sinicus. They share 87 to 92% of their nucleic acid and 93 to 100 % of their amino acid sequences with the SARS-CoV [429–433]. According to a phylogenetic study, MERS-CoV is a member of lineage C of the Betacoro-navirus genus. It resembled the pipistrelle bat (*Pipistrellus pipistrellus*) and lesser bam-boo bat (*Tylonycteris pachypus*) most closely, as well as the bat coronaviruses HKU4 and HKU5 [417, 434]. The whole genomic sequences of HKU4 and HKU5 and the RNA-dependent RNA polymerase (RdRp) gene show nucleotide identity with MERS-CoV of 50% and 82%, respectively. A recent study established that CD26, also known as dipeptidyl peptidase 4 (DPPIV), is a functional receptor for MERS-CoV. Additionally, it has been demonstrated that this molecule is evolutionarily conserved among mammals and that MERS-CoV can infect a wide variety of mammalian cells (including those from humans, pigs, monkeys, and bats), indicating ease of transmission between hosts [186, 435].

A large-scale PPI network of an organism provides valuable clues for understanding cellular and molecular functionalities, and signaling pathways can provide crucial insights into the disease mechanism, etc. Much biological information is available and encoded in different ontologies called GO. Semantic similarity is the degree of relatedness between the two biological entities (Gene/Protein) based on GO annotations that provide a quantitative measure of their GO-level relationship [280]. Different combinations of edge-based and node-based semantic similarity measures have been applied over the years from GO graphs [270, 272–275, 279, 335, 336, 340, 436, 437]. These methods have specific shortcomings concerning their designed GO semantic features. Some of them have used topological properties of the GO graph, some have used only the information content (IC) of the most informative common ancestor [272–274,436], and some have used DCA based approach [279,335,336]. To define the interaction affinity of any two proteins from their GO information, this hybrid approach is more effective as it incorporates topological features and average IC-based DCA techniques. Much work [128] has already been done to analyze host-pathogenic interactions [56, 136], disease detection [438], and disease-specific multi-omics network analyses [132].

6.2 Dataset

Alpha-, Beta-, Gamma-, and Delta-coronavirus are the four genera that comprise the enormous family of enveloped positive-strand RNA viruses known as coronaviruses. Among all the 44 organisms of coronavirus, here in this work, only 11 organisms have been considered based on the available GO-annotated proteins. The human is considered the host, and the work mainly suggests the affinity of host-pathogen interaction for different coronavirus organisms. Below, a brief description of all selected organisms is given.

1. ***Human Protein:*** All potential interactions between human proteins that have been experimentally verified in humans make up the dataset [318, 439]. The proteins in the Human organism are represented by nodes, whereas the edges represent the respective interactions between the organism. The proteins and their GO annotations are collected from Uniprot, the protein repository [33]. Uniprot contains 20,386 reviewed Human proteins, among which 19,283 proteins are associated with GO annotations.

2. ***SARS-CoV-2 Proteins:*** SARS-CoV-2 is a biological member of the Coronaviridae, which belongs to the genus beta coronavirus. The virus contains four structural proteins, namely envelop protein, membrane protein, nucleocapsid protein, and spike protein which help in binding with receptors after entering the human body and have a crucial function in spreading the disease [304]. Here the work is carried out by collecting the dataset of available SARS-CoV-2 protein from UniProtKB. The repository includes 16 reviewed SARS-CoV-2 proteins as of date.

3. ***SARS-CoV Proteins:*** SARS-CoV is a highly pathogenic and zoonotic virus that causes severe respiratory illness, gastrointestinal, neurological, and fatalities among humans [440–442]. The 2002-2003 SARS-CoV pandemic showed how susceptible humans are to coronavirus epidemics [313]. However, the dataset is collected from UniprotKB, which holds 15 reviewed SARS-CoV proteins.

4. ***MERS-CoV Proteins***: MERS-CoV is also a member of Betacoronavirus. It is an even more pathogenic and zoonotic virus in comparison to SARS-CoV. MERS-CoV emerged around 2012 in the Arabian Peninsula with very high transmissibility by affecting more than 2000 people [443]. The dataset has been retrieved from UniProtKB, which holds around 10 MERS-CoV proteins.

5. ***Bat coronavirus HKU3 Proteins:*** Surveillance research in Hong Kong among non-caged animals from wild regions found that a closely similar bat coronavirus, SARS-related Rhinolophus bat coronavirus HKU3, was the natural animal host [346]. We have retrieved a protein set of Bat coronavirus HKU3 from UniProtKB, having 12 proteins.

6. ***Bat coronavirus RP3/2004 Proteins:*** With the high geographic spread and species variety, bats represent an order with significant evolutionary success. Bats are the natural reservoirs of several viruses closely related to SARS-CoV [444]. A search for ACE2 sequence similarities in domestic and wild animals in Italy revealed domestic mostly horses, cats, cattle, and sheep and wild mostly European rabbits and grizzly bears animal species as potential SARS-CoV-2 secondary reservoirs. Molecular docking of these species ACE2 against the S protein of the Bat coronavirus (Bt-CoV/Rp3/2004) suggests that the primary reservoir *Rhinolophus ferrumequinum* may infect secondary reservoirs, domestic and animals living in Italy [445].

7. ***Bat coronavirus HKU5 Proteins:*** An enclosed, positive-sense single-stranded RNA mammalian Group 2 Betacoronavirus called bat coronavirus HKU5 was found in Japanese Pipistrellus in Hong Kong. This coronavirus strain is closely related to the recently discovered novel MERS-CoV, which is to blame for the coronavirus outbreaks linked to the Middle East respiratory illness in 2012 [188, 417].

8. ***Bat coronavirus HKU4 Proteins:*** *Tylonycteris* bat coronavirus HKU4, a member of Betacoronavirus, is an enveloped, single-stranded virus having a genetical similarity with MERS-CoV or HCoV-EMC. The main difference between HCoV-EMC and bat coronavirus HKU4 lies in between the spike protein and envelop protein, where HCoV-EMC have five ORFs instead of four with low amino acid identities to Bat-CoV HKU4 [446]. The human CD26 (hCD26) receptor is engaged explicitly by a receptor binding domain (RBD) in the MERS-CoV envelope-embedded spike protein to start viral entry. Due to the viral spike protein's great sequence identity, we looked into whether or not HKU4 and HKU5

can detect hCD26 for cell entrance. We discovered that HKU4-RBD binds to hCD26, but not HKU5-RBD and that pseudotyped viruses incorporating HKU4 spike can infect cells by recognizing hCD26. The overall hCD26-binding mechanism of the HKU4-RBD/hCD26 complex was identical to that of the MERS-RBD, according to the structure. However, HKU4-RBD has a lower affinity for receptor binding than MERS-RBD because it is less suited to hCD26 [447].

9. ***Bat coronavirus 133/2005:*** The spike and RdRp proteins of MERS-CoV were sub-jected to phylogenetic analysis, which indicated that the virus is linked to bat viruses. Coronavirus surveillance investigations in several populations of bats have shown that they are potential reservoirs for this unique virus [448]. Different phylogenetic studies reveal that MERS-CoV was grouped with the Betacoronavirus genus, particularly near BtCoV/133/2005 and BtCoV HKU4-2, which had the most significant S1 amino acid sequence similarity (60%) with MERS-CoV [449].

10. ***Murine coronavirus:*** M-CoV, a member of the Betacoronavirus family having Emba-covirus subgenus, is mainly found responsible for infecting rats [450, 451]. Enterotropic and Polytropic are the two strains of M-CoV. MHV strains D, Y, RI, and DVIM are examples of enterotropic strains. In contrast, hepatitis, enteritis, and encephalitis are the leading causes of illness caused by polytropic strains like JHM and A59 [452]. M-CoV comes in over 25 distinct strains. These viruses, which spread by the fecal-oral or respiratory routes and infect mice's livers, have been utilized as an animal disease model for hepatitis [453]. The strains MHV-D, MHV-DVIM, MHV-Y, and MHV-RI, which are transmitted in fecal matter, primarily affect the digestive tract. However, they can occasionally affect the spleen, liver, and lymphatic tissue [454].

11. ***Bovine coronavirus:*** BCoV is a member of Betacoronavirus, and it can infect both cattle and humans [455, 456]. It is also an enveloped single-stranded RNA virus that enters the host cell by binding itself with the N-acetyl-9-O-acetylneuraminic acid receptor [457, 458]. BCov is mainly responsible for causing gastroenteritis in calves resulting in massive economic damage [459]. BCoV consisted of five structural proteins, namely spike glycoprotein; integral membrane protein; hemagglutinin-esterase glycoprotein; small membrane pro-tein, and nucleocapsid phosphoprotein [460]. A phosphoprotein with a high content of essential amino acids, the N protein joins the genomic RNA directly to create a helicoidal nucleocapsid. The N protein carries out numerous activities related

to viral pathogenicity, transcription, and replication. Because it is a highly conserved protein expressed in significant amounts during viral replication, it is frequently employed for molecular diagnosis of BCoV [461].

12. ***Rat coronavirus:*** RCoV, subset of Murine coronavirus, is also a single-stranded RNA virus belonging to the Betacoronavirus family which is responsible for infecting rats [462]. The respiratory disease in adult rats is caused by RCoV in adult rats, which is characterized by an early Polymorphonuclear neutrophils (PMN) response, viral multiplication, inflammatory lung lesions, modest weight loss, and efficient infection resolution [463]. When a virus is present, PMN in the respiratory tract is typically associated with severe disease pathology [464–467]

6.3 Methodology

A GO-based Graph theoretic model is proposed to determine the interaction affinity between the host-pathogen protein pairs for humans and different coronavirus organisms. Currently, 19,281 human proteins have GO annotations, whereas around 242 viral proteins are obtained from a selected organism having GO annotations. Based on the above data, Level-1 interactors generate $\sim$ 4.5 million potential host-pathogen interactions. The variety and veracity issue plays a significant role in such a large-scale dynamic PPI network. Handling large, dynamic, heterogeneous networks using in-silico methods is tedious. Therefore, an Apache Spark-Based analytical study is proposed to compute the interaction affinity in large-scale PPIN using the GO graph.

6.3.1 GO Graph-based Scoring for Potential Host-Pathogen Protein Interaction Identification

Combining the similarity scores of the GO terms connected to the proteins will yield an estimate of the semantic similarity between two interacting proteins [56,272,317,336]. The greater the similarity between two GO pairs, the greater the interaction affinity between the proteins. The GO hierarchy's independent DAGs represent three distinct features of proteins: CC, MF, BP. Each node represents GO terms, and edges indicate various hierarchical relationships. The two fundamental relations *"is_a"* and *"part of"* GO graphs are considered for semantic score computation. Considering the similarity between all the GO pairs, the semantic similarity of the protein pairs can be estimated. The shortest path length between a pair of terms in a GO graph and the average information content IC [340] of the DCA of the respective GO term [336,337] measures the similarity of the pair.

$$PrM\left(t\right)=\frac{\left|An\left(t\right)\right|+\left|Dn\left(t\right)\right|}{\left|N_O\right|} \tag{6.1}$$

where *An(t)* is the ascendant term for t and *Dn(t)* is the descendent term of t. N_o is the total number of GO terms in ontology O, and *PrM(t)* is the proportion measure of term t. The GO keywords chosen as cluster centers are those for which this proportion metric is higher than a certain threshold. The cluster centers in this study are selected using the proposed threshold values [56, 132, 317, 438]. Once the cluster centers have been chosen, the shortest path lengths between each term in the ontology and the cluster centers have been calculated. The membership value of a GO term decreases with the increase in the shortest path length. The membership function of a GO term is given by.

$$MmFc(t)=e^{-\ \frac{-(x-c_i)^2}{2\ k^2}} \tag{6.2}$$

Where c_i is the i^{th} cluster center, x is the shortest path length and k is the width of the membership function. If no path from any GO term to a cluster center is found, then the membership of the GO term with respect to that cluster center will be considered 0. Similar membership for any target GO pair indicates very closely related concepts of GO functionality and widely related membership value represents separated concepts. For any target pair of GO-term *(t_i,t_j)*, a weight parameter is introduced to estimate this difference in membership. The weight parameter is thus defined by

$$WT(t_i,t_j)=1-maxD((ti,tj) \tag{6.3}$$

Where *maxD(t_i,t_j)* represents the maximum difference in membership values of GO pair (t_i,t_j) across all cluster centers of any particular GO graph type(CC/MF/BP). The information content (IC) based information of the disjunctive common ancestor *(DsjCAs)* of any GO graph is more significant in the semantic similarity assessment of two GO terms [279] IC of any GO term t, with respect to a GO graph g is defined as

$$ICg(t)=-log(Pr(t)) \tag{6.4}$$

The probability *Pr(t)* is the occurrences of the term t with respect to the total annotations of GO graph g. The occurrences of term t depend on its annotations over the protein cor-pus. Using the IC of the DsjCA, the shared information content (SIC) is computed for the target GO term pair (t_i,t_j). The SIC is computed as

$$SIC\left(t_i, t_j\right) = \frac{\sum_{a \in DCA(t_i,t_j)} IC\left(a\right)}{\left|DCA\left(t_i, t_j\right)\right|} \tag{6.5}$$

Finally, the semantic similarity, *sim*, between two GO pairs t_i and t_j is calculated as

$$sim(t_i, t_j) = WT(t_i, t_j)\ SIC\left(t_i, t_j\right) \tag{6.6}$$

When comparing the annotations of the proteins P_i and P_j for each type of GO, the maximum similarity of all possible GO pairs is used to determine the semantic similarity of the protein pair (P_i, P_j) for each GO-type *i.e.* CC, MF, and BP. The average of the CC, MF, and BP-based semantic similarity is used to define the protein pair's interaction affinity (P_i, P_j). Figure 6.1. refers to the schematic diagram of our proposed model where the host-pathogen interaction affinity between humans and organisms from the coronavirus family is calculated using the GO information, resulting in high-quality interactions for retrieving vulnerable human Prey for coronavirus hosts.

6.4 Experimental Result

The *in-silico* model proposed here, contains protein interaction affinity between humans and different organisms from the coronavirus family. The *in-silico* model is validated by identifying the overlapped edges *w.r.t.* the *state-of-the-art* datasets. Any computational model must always consider the input and output source, and the proposed model is no exception.

6.4.1 Identification of Host-Pathogen Protein Interactions for the Different Organisms of the Coronavirus Family

Identification of host-pathogen protein interactions for the different organisms of the coronavirus family [128]. The proposed GO-based *in-silico* model is applied to find the interaction affinity between the host protein and different organisms of the coronavirus family. Among 44 different organisms of the coronavirus family, based on the availability of the proteins, 11 organisms are considered. The proposed model is created from the ontological relationship graphs by comparing the affinities of all potential GO pairings that may be annotated from any target protein pair. Finally, the score of interaction affinity of protein pair based on their annotated GO pair-wise interaction is computed within a range of [0, 1]. Table 6.1 gives a detailed description of the number of proteins available for the respective coronavirus organism and the number of

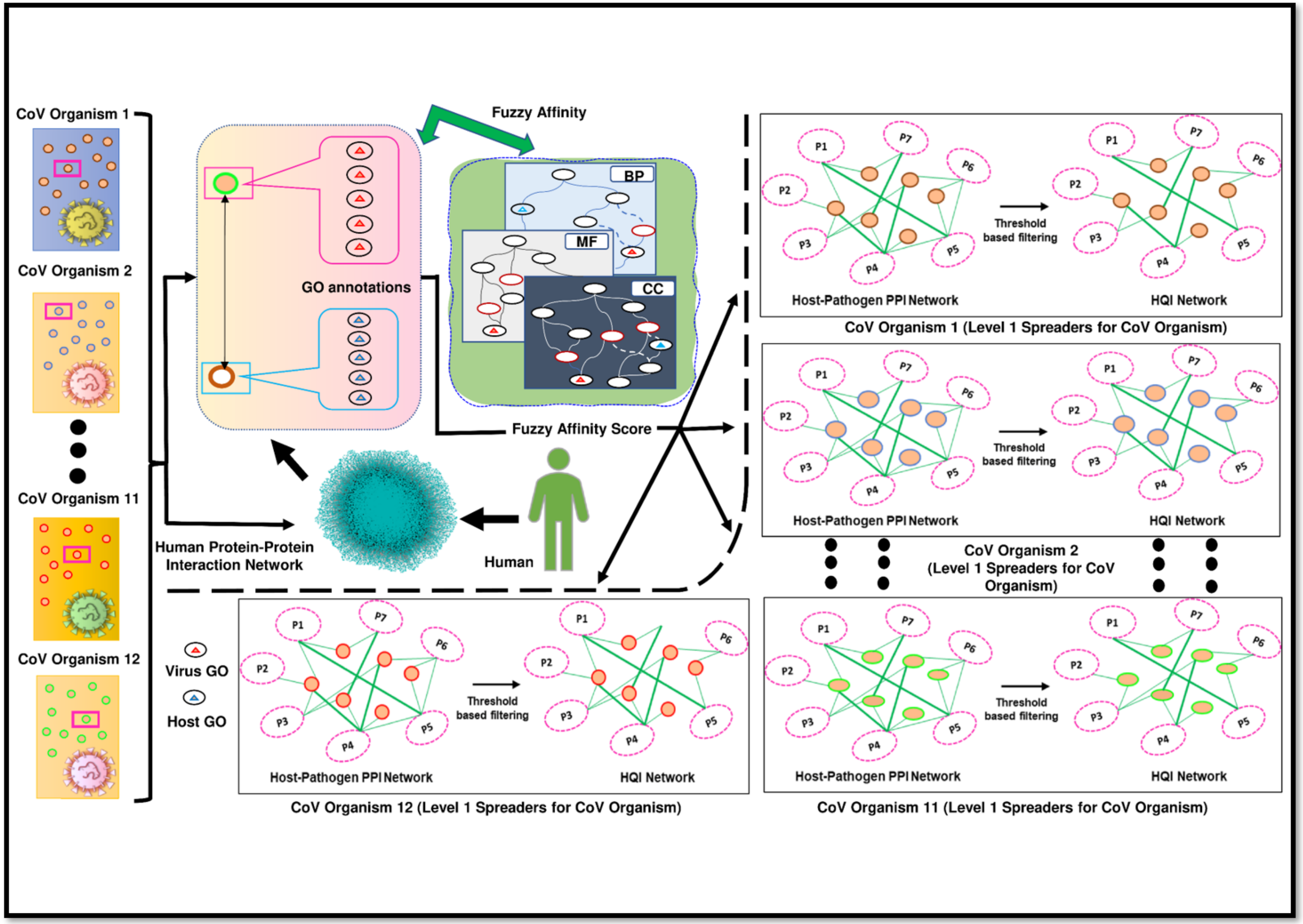

CoV Organism 1
CoV Organism 2
CoV Organism 11
CoV Organism 12
GO annotations
Fuzzy Affinity
BP
MF
CC
Fuzzy Affinity Score
Human Protein-Protein Interaction Network
Human
Virus GO
Host GO
P1
P2
P3
P4
P5
P6
P7
Threshold based filtering
Host-Pathogen PPI Network
HQI Network
CoV Organism 1 (Level 1 Spreaders for CoV Organism)
CoV Organism 2 (Level 1 Spreaders for CoV Organism)
CoV Organism 11 (Level 1 Spreaders for CoV Organism)
CoV Organism 12 (Level 1 Spreaders for CoV Organism)

Figure 6.1: Schematic diagram of our proposed model. The coronavirus and human proteins' interaction affinities are determined by the model using GO information of the proteins. Three different GO-relationship graphs, CC, MF, and BP, are used to evaluate all GO pair-wise interaction affinities. A protein pair's fuzzy interaction affinity is calculated using the three pair-wise scores of all GO-pair affinities.

Table 6.1: Detailed description of proteins and host-pathogen interaction for all organisms from the coronavirus family

Organism	**# of Proteins**	**# of Host-pathogen Interaction**
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	14	205,140
Severe acute respiratory syndrome coronavirus(SARS-CoV)	15	233,411
Bat coronavirus HKU3	12	125,904
Bat coronavirus Rp3/2004	13	125,904
Murine coronavirus	40	425,162
MERS-CoV	10	174,136
Bovine coronavirus	94	688,115
Bat coronavirus HKU5	10	117,090
Rat coronavirus	12	92,508
Bat coronavirus HKU4	10	117,090
Bat coronavirus 133/2005	10	98,494

possible host-pathogen interaction networks that can be generated for each organism.

6.4.2 Detailed Description of Human–nCoV Protein Interaction Network

The 2019 coronavirus disease pandemic was brought on by the novel coronavirus known as SARS-CoV-2/nCoV. It affected over 12 million people and caused over 560,000 fatalities in 213 nations [468]. To infect a host, the nCoV protein, like other virus proteins, must interact with the host protein and replicate the genome. At the time of our experiment, Uniprot [33] holds around 19,283 human proteins and 16 nCoV proteins (*see* Table 6.3) having GO annotations. Here, through our proposed *in-silico* model, we compute all the possible protein interactions between human-nCoV for all the proteins having GO annotations (*see* Table 6.4). Detailed descriptions for all types of possible interactions are given in Table 6.2.

6.4.3 Validation Through the *State-of-the-art* Dataset

Gordon *et al.* [164] proposed a host-pathogen interaction dataset physically connected with the human cell by cloning, tagging, and expressing 27 out of 29 proteins using affinity-purification mass spectrometry. Up to 14 open-reading frames can be encoded by a 30-kb genome (ORFs). In order to create the 16 non-structural proteins (NSP1-

Table 6.2: Detailed statistics of Human-nCoV protein interactions computed by our proposed model

Intersection Type	**Organism**	**Proteins**	**Interactions**
All	Total Dataset	19,297	164,701,415
Host-Pathogen	Human-nCoV	19,297	206,516
Pathogen-Pathogen	nCoV-nCoV	14	83
Host-Host	Human-Human	19,283	164,494,816

Table 6.3: Details of nCoV proteins collected from Uniprot [33]

Entry	**Entry Name**	**Protein Name**
P0DTD1	R1AB_SARS2	Replicase polyprotein (ORF1ab polyprotein)
P0DTC1	R1A_SARS2	Replicase polyprotein (ORF1a polyprotein)
P0DTC2	SPIKE_SARS2	Spike glycoprotein (Peplomer protein)
P0DTD8	NS7B_SARS2	ORF7b protein (Accessory protein 7b)
P0DTC6	NS6_SARS2	ORF6 protein, ORF6 (Accessory protein 6)
P0DTC8	NS8_SARS2	ORF8 protein, ORF8 (Non-structural protein 8, ns8)
P0DTF1	ORF3B_SARS2	Putative ORF3b protein
P0DTC5	VME1_SARS2	Membrane protein (E1 glycoprotein)
P0DTD3	ORF9C_SARS2	Putative ORF9c protein
P0DTC3	AP3A_SARS2	ORF3a protein
P0DTG0	ORF3D_SARS2	Putative ORF3d protein
P0DTG1	ORF3C_SARS2	ORF3c protein (ORF3h protein, ORF3h)
P0DTC7	NS7A_SARS2	ORF7a protein, ORF7a
P0DTD2	ORF9B_SARS2	ORF9b protein, ORF9b
P0DTC9	NCAP_SARS2	Nucleoprotein
P0DTC4	VEMP_SARS2	Envelope small membrane protein

NSP16) that make up the replicase transcriptase complex, ORF1a and ORF1ab encode polyproteins. This produces a dataset of 332 high-confidence host-pathogen PPIN. However, while validating our computational model, we discovered that the protein sequences provided by Gordon *et al.* do not have any mapping with the corresponding UniProt id. The proposed method has exclusively focused on the SARS-CoV-2 proteins published on UniProt. The proposed method has used a mathematical model

Table 6.4: Details of Human-nCov Interactions at different threshold values

Interaction Type	Organism	Threshold	Nodes	Edges	Human	nCoV
Host-Pathogen	Human-nCoV	0.2	109	592	10	12
		0.15	245	1,174	128	13
		0.1	886	2,909	768	13
		0.09	1,193	3,586	1,075	13
		0.08	1,754	4,619	1,636	13
		0.05	7,397	16,209	7,278	13
		0.02	15,551	74,560	15,431	13
		0.001	18,936	166,382	18,816	14

to determine the binding affinities of a portion of the evaluated human proteins listed on UniProt. Because SARS-CoV-2 proteins could not be directly mapped into corresponding UniProt accession IDs, direct comparison and validation concerning Gordon *et al.* was impossible. Thus, the nCoV proteins from Gordon *et al. were mapped to the corresponding Uniprot IDs. As the proposed research heavily depends on the underlying* GO network of the host-pathogen protein interaction network, those proteins are selected with all three GO annotations. To validate the proposed method, all possible interactions are computed in the proposed computational environment, which gives 57,615 possible interactions, which are their respective fuzzy score from 27 bait and 332 prey. Among these interactions, 129 existing host-pathogen from high confidence dataset proposed by Gordon *et al.* whose scores are calculated.

Apart from the high-confidence host-pathogen protein interaction network dataset, Gordon *et al.* also provided a host-pathogen interaction dataset that contains a human-nCoV protein interaction network without any threshold. This mainly contains scoring results of all bait and all prey proteins showing spectral counts of experimental samples. The dataset contains 22,153 interactions, including 27 bait and 2753 host proteins. The proposed model generates an interaction network with the said protein, which generates all-vs-all interactions. Among those 22,153 interactions, there are 7,866 existing host-pathogen interactions whose scores are calculated. Table 6.5 gives detailed information regarding the host-pathogen interaction for the high-confidence human-nCoV dataset and the generic human-nCoV dataset proposed by Gordon *et al.*

6.4.4 Comparison with *Gordon.et.al.* [164]

To validate the proposed computational model, the data set has been compared with that proposed by Gordon *et al.* [164]. To experiment with the proposed computational

Table 6.5: Overall statistics for interaction affinity score of high confidence human-nCov dataset and all human-nCov dataset proposed by Gordon *et al.* computed by the proposed model

Dataset	**# of Interactions**	**# of Bait**	**# of Prey**	**Total Interaction Score Computed**
High Confidence Host-Pathogen PPI	332	27	332	57,615
All Host-Pathogen PPI	22,153	27	2,753	2,156,507

model, a dataset of human and SARS-CoV-2/nCoV proteins has been constructed by retrieving from the UniProt protein repository, as discussed above. The computation results in fuzzy scoring of the protein pair *viz.* human-human ppin, human-nCoV ppin, and nCoV-nCoV ppin. The edge-overlapping has shown the validation of our computational model between two datasets at different threshold values set on the fuzzy score. Edge overlapping signifies the common edges present in both datasets. For the experiment, the fuzzy score threshold ranging from 0.1-0.001 has been kept. At first, the network of the proposed model has been compared with the high-confidence human-nCoV network proposed by Gordon et al. The dataset contains 332 host proteins and 27 viral proteins. Table 6.7 compares two datasets at different threshold values and produces the intersected nodes and edges between the two datasets, along with the common host and viral proteins.

The high-confidence dataset and the other dataset proposed by Gordon et al., which contains scoring results of all bait and all prey proteins showing spectral counts of experimental samples, are also being compared in the same manner discussed here with varying threshold values imposed on fuzzy interaction affinity score. The threshold ranges from 0.1-0.001. The dataset proposed by Gordon et al. contains 2753 host proteins and 27 viral proteins. Table 6.6 represents the comparison between the two datasets at different threshold values and produces the intersected nodes and intersected edges between the two datasets.

6.4.5 Comparison with *Dick. et al* [469]

Protein-protein Interaction Prediction Engine (PIPE) is a sequence-based PPI prediction approach that looks at sequence windows on each query protein proposed by Dick

Table 6.6: Detailed validation of the model compared to all human-nCov Datasets proposed by Gordon *et al.* [164]

Data (Gordon et.al)	**Proposed Dataset**					
Number of Host	**No. of Bait**	**Threshold**	**Number of Host**	**No. of Bait**	**# Intersected Nodes**	**# Intersected Edges**
2753	27	0.1	17875	13	88	149
2753	27	0.09	18064	13	104	176
2753	27	0.08	18218	13	128	214
2753	27	0.05	19838	14	381	626
2753	27	0.02	19123	14	1129	2513
2753	27	0.001	19193	14	1817	6634

Table 6.7: Detailed validation of our model compared to High confidence human-nCoV proposed by Gordon *et al.* [164]

Data (Gordon et.al)	**Proposed Dataset**					
Number of Host	**No. of Bait**	**Threshold**	**Number of Host**	**No. of Bait**	**# Intersected Nodes**	**# Intersected Edges**
332	27	0.1	768	13	8	5
332	27	0.09	1075	13	8	5
332	27	0.08	1636	13	8	5
332	27	0.05	7278	13	20	14
332	27	0.02	15431	13	60	51
332	27	0.001	18816	14	109	99

et.al. [469]. The evidence for the putative PPI is strengthened if the two sequence windows have a lot in common with other pairs of proteins that have been found to interact. Normalization is used in a similarity-weighted scoring system to consider common sequences unrelated to PPIs. A PPI is anticipated, given enough supporting data [470–472]. For understudied species, the PPI Prediction Engine (PIPE4) iteration

Table 6.8: Detailed validation of the model compared to all Human-nCov Datasets proposed by Dick *et.al*

Dataset Dick *et.al*	**# of Interactions**	**# of Bait**	**# of Prey**	**Total Interaction Score Computed**
PIPE4	702	13	518	575
SPRINT	510	15	368	413

has recently been modified [473].

Like PIPE, the SPRINT predictor gathers data from previously reported PPI interactions based on window similarity with the query protein pair to determine its prediction scores [474]. SPRINT uses a spaced seed method to compare the sequences of protein windows, where only certain places in the two windows must match, as determined by the bits of the spaced seeds. Additionally, because proteins are encoded with five bits per amino acid, it is possible to quickly compute protein window similarities and, consequently, forecast scores using very efficient bitwise operations [474].

Here, the two datasets produced by Dick *et al.* [469] are being compared, and an interaction affinity pair is being generated by using our proposed method. Table 6.8 shows the details of the comparison with both datasets. The table shows that PIPE4 contains 702 interactions, among which the proposed model identifies 575 interactions, and the score has been generated. On the other hand, the SPRINT dataset contains 510 interactions, among which 413 are identified by the proposed method.

6.4.6 Vulnerable Host Protein

One of the main focuses of the research is to identify the common vulnerable host proteins at different threshold values. As discussed in subsection 6.4.1, the proposed computational model efficiently computes the interaction affinity and can generate a fuzzy score for any host-pathogen interaction pair for any organism from the Corona family. The host-pathogen network has been experimented for the entire corona family with the selected organism, as mentioned in section 6.2 and retrieved the network at different threshold values ranging from 0.1-0.001. At each threshold score, the network for each COVID organism is segregated and constructed in their respective networks. Thus for each threshold score, a separate host-pathogen network is obtained for each coronavirus organism. So, for each threshold score, some common host protein interacts with all the coronavirus organisms. As the value of the score decreases from a high threshold to a low threshold value, the number of common host proteins increases. These host proteins are the level-1 spreader nodes. These spreader nodes are identified by fuzzy thresholding, and these host proteins are vulnerable to the propagation or contamination of the diseases caused by the viral proteins. Table 6.9 represents the number of vulnerable host proteins at different fuzzy threshold scores. Figure 6.2 and Figure 6.3 represent the Venn diagram of the vulnerable host proteins at 0.1 and 0.001 threshold values, respectively. For simplicity and ease of the process, the viral organism is divided into three subsets. SARS-CoV-2, SARS-CoV and MERS-CoV forms one group, all the different organism from BAT-CoV *viz., Bat coronavirus HKU3, Bat coronavirus Rp3/2004, Bat coronavirus HKU5, Bat coronavirus HKU4, Bat coronavirus 133/2005* forms one group, and M-CoV, BCoV and RCoV forms the third group. Then the common host proteins from all three groups are identified separately. Intersected host protein sets from all three groups are identified and again intersected. This results in the common vulnerable host proteins at the specified threshold value. For visualization, a threshold value of 0.1 is selected arbitrarily for constructing the Venn diagram. 0.1 threshold value gives 191 vulnerable host proteins interacting with all selected coronavirus organisms.

All level 1 human proteins of the coronavirus family are mapped with their matching medicines from DrugBank once the coronavirus family-human PIN has been created [330]. DrugBank is an online database that offers extensive information on medicines, drug-protein targets, and drug metabolism [352]. Most *in-silico* approaches used in drug design, drug docking, and drug interaction prediction use DrugBank as their most frequently used database because of its high-quality annotation.

It has around 60% FDA-approved medications and 10% of investigational drugs. It has been determined through adequate analysis that some spreader nodes in COVID-19-human PPIN are the protein targets of possible COVID-19 FDA-approved medicines [227]: hydroxychloroquin [323], azithromycin [323], lopinavir [321], remdesivir [226, 324], etc. Not only the list of drugs for COVID-19, but we have obtained a list of FDA-approved drugs from level-1 vulnerable host proteins for the entire coronavirus family by using DCS. The algorithm is defined as the number of times a drug occurs at a specific PPIN level. Each human protein is mapped with the appropriate related medicines in this level-1 PPIN.

The DCS, or frequency of each drug, is therefore calculated. Table 6.10 represents the top-5 FDA-approved drug at different fuzzy threshold values and the number of vulnerable host proteins at that corresponding threshold value, Drug ID, and corresponding DCS score for each drug. Fostamatinib is thought to be a promising med-

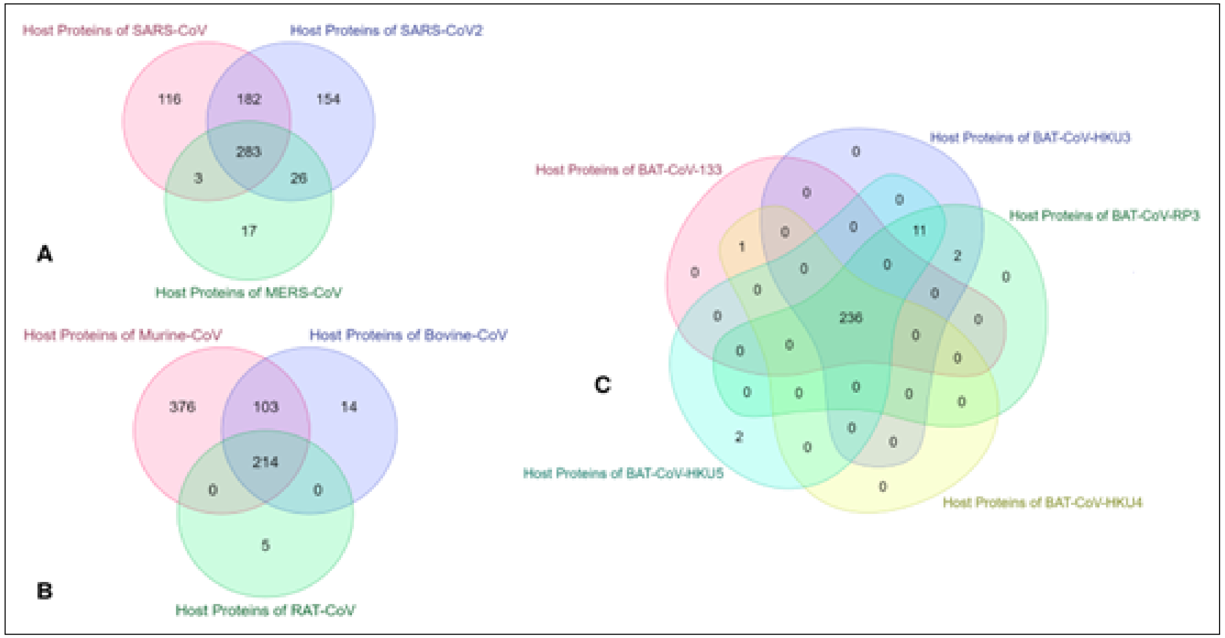


Figure 6.2: Venn diagram of number for vulnerable host proteins obtained from host-pathogen interaction for all selected coronavirus organisms at 0.1 fuzzy threshold value. (A). The intersection of host protein identified from SARS-CoV-2, SARS-CoV, and MER-CoV. (B) Intersected host proteins from Murine-CoV, Bovine-CoV, and Rat Coronavirus. (C). Intersected host proteins of the different viral organisms of Bat Coronavirus.

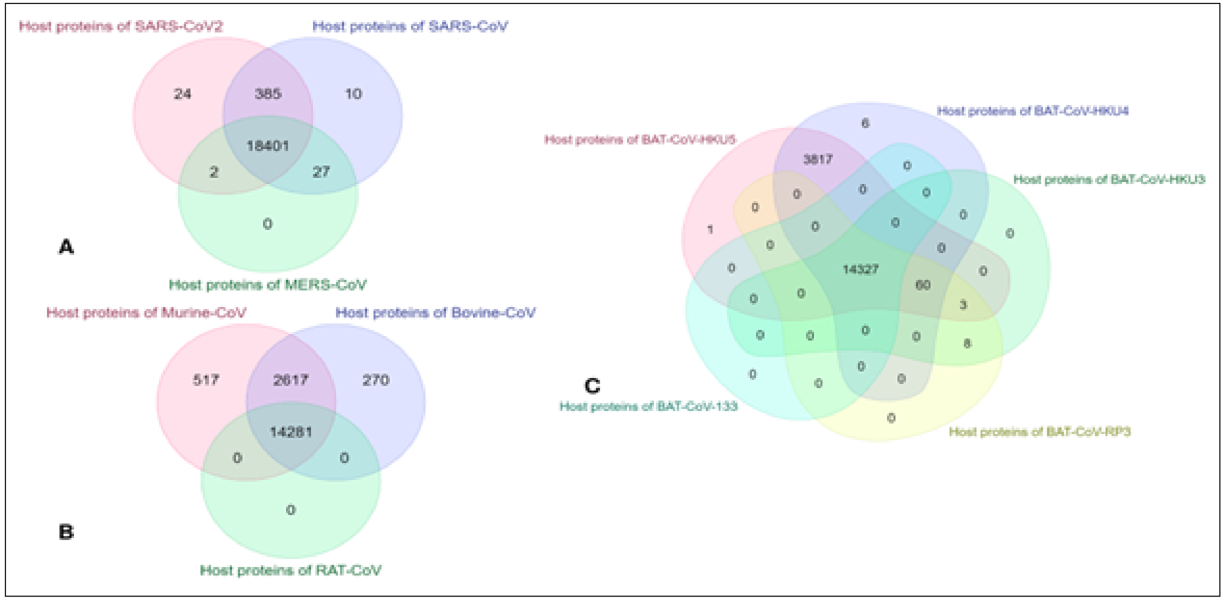


Figure 6.3: Venn diagram of a number of vulnerable host proteins obtained from host-pathogen interaction for all selected coronavirus organisms at 0.001 fuzzy threshold value. (A). The intersection of host protein identified from SARS-CoV-2, SARS-CoV, and MER-CoV. (B) Intersected host proteins from Murine-CoV, Bovine-CoV and Rat Coronavirus. (C). Intersected host proteins from different viral organisms of Bat Coronavirus.

Table 6.9: Number of Vulnerable host proteins identified from the host-pathogen network for all selected coronavirus organisms at a different fuzzy threshold score

Threshold	**# of Vulnerable Human Proteins**
0.001	14279
0.005	11208
0.03	3889
0.05	526
0.07	351
0.1	191

ication for the target nCoV protein in the randomly created COVID-19 Human PPI since it has the highest DCS in most cases.

6.5 Discussion

The number of vulnerable host proteins at different threshold values is represented in Table 6.10, and the list of the top 5 drugs, along with their drug-id based on the DCS score, are listed. This leads us to the analysis with the application of the lowest threshold values *(i.e., 0.001)*, based on which the possible repurposed drugs are proposed.

Drug repurposing is a powerful strategy that gives new therapeutic alternatives by identifying other uses for already-approved medications, as vaccine and drug development can take years [475]. The traditional conservative drug development approach, which is restricted to ”one drug, one target” paradigms, does not take into account or assess the off-target effects or the likelihood of numerous drug indications, even though some of them have since been confirmed to exist [476]. Upon the formation of the coronavirus-human PPIN, all level-1 Coronavirus human proteins are mapped with the appropriate medications via DrugBank [330].

DrugBank is an online database that provides detailed information on pharmaceuticals, drug-protein targets, and drug metabolism. DrugBank is the most often utilized database in practically all *in-silico* approaches used in drug design, drug docking, and drug interaction prediction because of the high-quality annotation in the database. It includes 10% and 60% of FDA-approved and investigational medications. It is observed that the above list of drugs at the threshold value 0.001 are listed in Table 6.9.

Table 6.10: Top 5 target drugs with their respective DCS score at different threshold values

Threshold	Vulnerable Human Proteins	Drug ID	DCS Score	Drug Name
0.001	14279	DB12010	181	Fostamatinib
		DB09130	47	Copper
		DB14533	45	Zinc chloride
		DB14487	45	Zinc acetate
		DB01593	45	Zinc
0.005	11208	DB12010	173	Fostamatinib
		DB01069	45	Promethazine
		DB01593	39	Zinc
		DB09130	39	Copper
		DB14487	39	Zinc acetate
0.03	3889	DB12010	25	Fostamatinib
		DB09130	6	Copper
		DB04464	5	N-Formylmethionine
		DB14487	5	Zinc acetate
		DB11638	5	Artenimol
0.05	526	DB12010	7	Fostamatinib
		DB12267	2	Brigatinib
		DB00041	2	Aldesleukin
		DB00074	2	Basiliximab
		DB09130	2	Copper
0.07	351	DB00041	2	Aldesleukin
		DB12010	2	Fostamatinib
		DB11638	2	Artenimol
		DB00004	2	Denileukin diftitox
		DB02240	1	Quinacrine mustard
0.1	191	DB12267	1	Brigatinib
		DB00111	1	Daclizumab
		DB11942	1	Selinexor
		DB08804	1	Nandrolone decanoate
		DB00047	1	Insulin glargine

When compared to the remaining human protein-associated medications, fostamatinib has the highest frequency of occurrence in the entire PPIN and has a sizable overlap of target proteins in the human-coronavirus PPIN with highest DCS of 181. It was already discussed and proposed in [352] that Fostamatinib has the highest DCS score *w.r.t* level-1 and level-2 human spreader proteins.

Thus, the drug of concern for the proposed work has shifted to the one with the next highest score, copper. Copper has an enormous effect in defeating COVID-19 which helps it to dominate with a high DCS score. The study proposed in [226] aims to investigate the effects of a highly specialized drug, "Hinokitiol Copper Chelate" on enormous quantities of 2019-nCoV Spike Glycoprotein with a single RBD. This investigation offers a superior version of Hinokitiol Copper Chelate for *in-vitro* testing against 2019-nCoV Main Protease.

The authors suggest combining copper, NAC, colchicine, NO, and the experimental antivirals remdesivir or EIDD-2801 as a potential treatment for SARS-COV-2 [477]. *In-silico* docking study of copper complexes with SARS-CoV-2 viruses shows a steady binding with SARS-CoV-2 main protease (M^{pro}) active-site region [478].

Zinc supplements also play a crucial role in combating different organisms of coronavirus. The essentiality of Zinc lies in the preservation of natural tissue barriers such as the respiratory epithelium, preventing pathogen entry, for a balanced functioning of the human immune system. The deficiency of Zinc can probably lead to the infection and detrimental progression of COVID-19 [479]. The body's tissue barriers, which contain cilia, mucus, anti-microbial peptides like lysozymes, and interferons, stop infectious organisms from entering. The primary mechanisms for SARS-CoV-2 entering cells are the cellular protease TMPRSS2 and the angio-tensin-converting enzyme 2 (ACE2) [347]. People with COVID-19 are accompanied by ciliated epithelium destruction and ciliary dyskinesia, which limit mucociliary clearance [480]. The quantity and length of bronchial cilia increased after zinc supplementation in zinc-deficient rats [481]

In COVID-19, zinc supplementation was hypothesized to reduce mortality. Supplementing with zinc had no positive effects on how the illness progressed. The zinc-supplemented group's hospital stay was lengthier. There is no evidence to back up regular zinc supplementation in COVID-19 [482]. The confounding variables impacting zinc's bio-availability may be avoided by administering zinc intravenously, enabling zinc to fulfill its medicinal potential. If effective, intravenous zinc might be quickly incorporated into clinical practice due to benefits such as lack of toxicity, cheap cost, and accessibility of supply [483].

Promethazine, an antipsychotic agent showing clathrin-mediated endocytosis, is one most effective drugs for SARS-CoV and MERS-CoV, which has been repurposed for the treatment of COVID-19 as there are almost 89% genetic similarity with SARS-CoV-2 and SARS-CoV [354]. Two pills were offered as an intervention, one with Aspirin and Promethazine and the other with vitamins D3, C, and B3, together with zinc and selenium supplements [484]. A randomized clinical trial has been conducted to recover mildly to moderate COVID-19 patients.

Based on this validation, further research on the repurposed drug, docking study, and other symptomatic analyses will help to identify the potential drug for the entire coronavirus family. A clinical study on Promethazine and Fostamatinib [352, 484] is also in progress. Even though the research is in its early stages, it in some way partially corroborates our findings.

In this chapter, an assessment model has been proposed that finds the interaction affinity between human-coronavirus interactome. It also identifies the potential repurposed drugs for the treatment of COVID-19 from level 1 spreader proteins. From the above study, it has been elucidated that protein interaction network plays a significant role in disease mechanisms. It has also been observed that proteins which undergo PTM are more involved in interacting with other proteins [26]. In the following chapter, an *in-silico* model has been proposed to predict the S-palmitoylation PTM sites in synaptic protein sequences in male/female mouse data.

Chapter 7

Prediction of PTM Sites in Protein Sequences - A Case Study with S–Palmitoylation for Synaptic Proteins

7.1 Background

The chapter presents a prediction approach for predicting the S-Palmitoylation site for Synaptic proteins in mouse data. S-palmitoylation is a reversible covalent post-transnational modification of cysteine thiol side chain by palmitic acid. S-palmitoylation plays a critical role in a variety of biological processes and is engaged in several human diseases. Therefore, identifying specific sites of this modification is crucial for understanding their functional consequences in physiology and pathology. An RF-classifier-based prediction strategy has been proposed to predict palmitoylated cysteine sites on synaptic proteins from male/female mouse data. A heuristic strategy for selection of the optimum set of physicochemical features from the AA-Index dataset using (a) K-Best (KB) features, (b) genetic algorithm (GA), and (c) a union (UN) of KB and GA based features have been introduced. A class of neuron-specific phosphoric proteins known as synaptic proteins is connected to synaptic vesicles. They attach to the cytoskeleton and are found on the surface of practically all synaptic particles. Synaptic proteins are mainly found in central nervous system and are distributed over the brain. Brain functions strictly depend on precise structural and functional synaptic integrity regulation. Among the mechanisms governing synaptic protein functions, PTM [485,486] play a pivotal role. PTM's may influence synaptic protein activity and turnover, localization at the synapse, and signaling cascades [487–490].

One of the PTMs is protein S-palmitoylation involving covalent attachment of palmitic acid (C16:0) to cysteine residue(s) via a thioester bond. Recent studies showed that S-palmitoylation can modulate protein localization, stability, activities, and trafficking and play an essential role in various biological processes, including synaptic plasticity [491,492], cell signaling, cellular differentiation [493], and apoptosis [494].

Unlike other fatty acid modifications, S-palmitoylation is a reversible process, tightly regulated by two groups of enzymes: palmitoyl acyltransferases (PATs) which is a palmitoylating enzymes and palmitoyl thioesterases which is a depalmitoylating enzyme. It is widely accepted that repeated cycles of palmitoylation/depalmitoylation are critically involved in regulating multiple protein functions. The molecular mechanisms that lie behind site-specific protein S-palmitoylation remain largely unknown. Several human diseases are often associated with the atypical activity of PATs together with changes in the pattern of S-palmitoylation. S-palmitoylation has been implicated in a wide range of human disease states such as cancer [193], Alzheimer's disease [495], Parkinson's disease, cardiovascular disease, schizophrenia [496], or major depressive disorder (MDD) [208]. Therefore, identifying substrates that undergo S-palmitoylation and specific sites of these modifications may provide candidates for targeted therapy.

Twenty-three PATs have been identified in mammalian cells, which mediate the majority of protein S-palmitoylation. One of the known PATs is a zinc finger DHHC domain-containing protein 7 (Zdhhc7, abbreviated ZDHHC7). This enzyme palmitoylates various synaptic proteins involved in the regulation of cellular polarity and proliferation [497, 498]. Moreover, Zdhhc7 is responsible for S-palmitoylation of sex steroid receptors such as estrogen and progesterone receptors [232, 498, 499]. Importantly, *Zdhhc7-/-* mice developed symptoms characteristic of human Bartter syndrome (BS) type IV because ZDHHC7 protein may affect ClC-K-barttin channel activation [233]. Thus, targeting ZDHHC7 activity may offer a potential therapeutic strategy in certain brain pathophysiological states. Most recently, using the mass spectrometry approach, we have identified sex-dependent differences in the S-palmitoylation of synaptic proteins potentially involved in the regulation of membrane excitability and synaptic transmission as well as in the signaling of proteins involved in the structural plasticity of dendritic spines in the mice brain [232]. For the first time, sex-dependent action of ZDHHC7 acyltransferase is being manifested. Furthermore, it has been revealed that different S-palmitoylation proteins control the same biological processes in male and female synapses [232, 233].

Several methods have been developed for the identification of S-palmitoylation target proteins. However, site-specific identification of S-palmitoylation is less studied. Large-scale identification of S-palmitoylation sites mainly relies on mass spectrometry-based methods such as PANIMoni [231] or PALMPiscs or ssABE [500]. These methods have been successfully used to identify a large number of S-palmitoylated proteins in different species, such as rats, mice, or humans. For instance, PANIMoni has been used to describe endogenous S-palmitoylation and S-nitrosylation of proteins in the rat brain excitatory synapses at the level of specific single cysteine in a mouse model of depression [231]. In recent years, results of large-scale proteome databases obtained with PANIMoni, PALMPiscs, or ssABE methods were used to develop tools to predict sites of specific S-palmitoylation in other biological complexes. Several machine learning-based algorithms [121, 228–230] have been developed for predictions of S-palmitoylation sites such as; NBA-PALM [214] and CSS-PALM [228], but their accuracy is uncertain. Therefore, with the growing number of publicly available large-scale proteome databases of the brain and somatic tissues, there is a need for the development of reliable and accurate computational tools to process them.

Considering the growing recognition for the importance of PTMs of proteins in cell physiology, this study aims to develop a computational tool for predicting S-palmitoylation sites using proteomic data obtained by the mass spectrometry-based method PANIMoni [231]. Most recently, the approach has been successfully used to create a detailed ZDHHC subtype-specific and sex-mouse S- palmitoylome [232, 233]. Here, in this work, these protein databases are used for validation of the computational tool.

This chapter has proposed a RF [501] classifier-based consensus strategy, which can predict the S-palmitoylated cysteine sites on synaptic proteins of the male/female mouse dataset. Different heuristic selection strategies have been applied to the physico-chemical features from the AAIndex feature database [502] along with position-specific amino acid (AA) propensity information, which eventually generates three different sets of features: (a) KB features, (b) GA based features [503], and (c) UN of K-Best and GA based features. The experiment has been carried out on three categorized synaptic protein datasets originally described in [232,233]; *viz.*, male, female, and combined *i.e.*, combination of male and female. In each experimental group, the weighted data is used as the training set, and the knock-out is used as the hold-out test set for performance evaluation and comparison. A novel RF-driven consensus strategy with efficient feature selection shows significant performance in predicting S-palmitoylation sites in mouse data.

7.2 Dataset Description

Experimental S-Palmitoylated datasets, used here, are categorized into three groups, male, female, and combined (includes both male and female), where each category contains two types of data: weight (WT) and knock-out (KO). WT is used for training, and KO data is considered for testing. The dataset was derived using the mass spectrometry-based PANIMoni method from WT and KO from ZDHHC7 mouse brains. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD025286.

Table 7.1: Dataset details of positive and negative sites for all three benchmark data; *Male*, *Female*, and *Combined.*

Category	**Type**	**# of Proteins**	**# of Cystene Sites**
Male	Positive(P_D)	1077	1877(Experimental)
	Negative(N_D)	1175	9279(Identified)
Female	Positive(P_D)	1036	1773(Experimental)
	Negative(N_D)	1131	8934(Identified)
Combined	Positive(P_D)	1180	2083(Experimental)
(Male+Female)	Negative(N_D)	1293	10,403(Identified)

The benchmark dataset for this experiment is constructed with the data available in the said repository. In this experiment, for all three benchmarking datasets, namely, *male*, *female*, and *combined*, WT data is considered a train set, and KO data is considered the test set for classification. Both male and female datasets contain peptides, modified sites, and assigned proteins. All the modified cysteines are labeled. The cysteines which are labeled with *Carbamidomethyl* are palmitoylated and are considered as positive data. The cysteines which are labeled as *N-ethylmaleimide* are not palmitoylated and they constitute the negative data. In this approach, to retrieve the high-quality negative samples, the cysteine positions, which are not within the selected fragments of positive samples, are considered. The cysteine positions with both *Carbamidomethyl* and *N-ethylmaleimide* modification create ambiguity in S-palmitoylation identification and thus are discarded from this experiment. In all experiments, the positive and negative ratio is kept as 1:1 for balanced classification. The details of the three benchmark datasets are shown in Table 7.1.

7.3 Feature Set

To design the features for the classification task, physicochemical properties of the amino acid are being incorporated [502]. The position-specific amino acid propensity is computed from the primary sequence of proteins and extracting α-length sequence window for each cysteine site with the cysteine at the center of the subsequence.

7.3.1 Position-Specific Amino Acid Propensity ($\boldsymbol{PSAAP}$)

The position-specific feature of amino acids is introduced for feature design. First, the position-specific amino acid composition is computed for all λ-length sub-sequences in the positive dataset (say P_D). Initially, the positive data set is divided into five different non-overlapping subsets. For any subset of positive data, the amino acid composition for *i-th* position is defined as, $(A_{1,i}^P, A_{2,i}^P, A_{3,i}^P, A_{4,i}^P......, A_{20,i}^P)^T$ where, i=1,2,3,...,λ and 20 amino acids are ordered alphabetically according to their single letter code. Then, the position-specific amino acid composition is computed as the position-wise average over all five subsets, denoted as $A_{1,i}^{-P}$. Similarly, the negative dataset is partitioned into five equal partitions where each subset size $=|N_D|=|P_D|$. The position-wise amino acid composition is computed for all negative subsets (as done in the case of P_D). The position-wise amino acid composition for individual negative subsets is calculated as, $(A_{1,i}^N, A_{2,i}^N, A_{3,i}^N, A_{4,i}^N......, A_{20,i}^N)^T$ where, i=1,2,3,...,λ is the average of amino acid composition over five negative subsets is represented as $A_{1,i}^{-N}$ Finally, the propensity of the *j-th* amino acid at position i in the cysteine sites is computed as:

$$X_{i,j} = \frac{\bar{A}_{1,i}^{-P} - \bar{A}_{1,i}^{-N}}{\bar{\sigma}_{j,i}^{N}} \tag{7.1}$$

where, $\bar{\sigma}$ represents the standard deviation of *j-th* amino acid at position i overall negative subsets. With these propensity values, final propensity matrix $ProP_{20\times\lambda}$ is constructed as:

$$\begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,\lambda} \\ \cdots & \cdots & \cdots & \cdots \\ \vdots\ddots & \vdots\ddots & \vdots\ddots & \vdots \\ X_{20,1} & X_{20,2} & ... & X_{20,\lambda} \end{bmatrix}$$

7.3.2 Physicochemical Properties Based *PSAAP*

In the next level, a physicochemical property-based feature is generated by incorporating the PSAAP *(ProP)*. Currently, there are 566 physicochemical features in the AAIndex database [502] A numeric score is assigned to each amino acid in the AAIndex database representing any particular physicochemical property scale. Then, the scores are normalized by [0, 1] for all amino acids for individual AAIndex using max–min normalization. From any target subsequence (*length*=λ), the final feature for any amino acid θ at position ι is for amino acid property φ defined as

$$\tau(\theta, \iota) = ProP((Ordx(\theta), \iota) \times PHY_{\varphi}(\theta, \iota) \tag{7.2}$$

where, $Ordx(\theta)$ represent the ordering index of amino acid θ in *ProP* matrix and $PHY_{\varphi}(\theta, \iota)$

7.4 Sub-Sequence Length Selection

To prepare the dataset, protein sequences are segmented into equal-length windows containing the cysteine at the center position. Amino acid sequences before and after

Table 7.2: Performance with different lengths of sub-sequences.

Length(n)	**Precision**	**Recall**	**Accuracy**	**F1**	**AUC**
15	0.657	0.792	0.69	0.718	0.765
16	0.701	0.731	0.709	0.715	0.781
17	0.699	0.722	0.706	0.71	0.777
18	0.72	0.731	0.722	0.725	0.788
19	0.724	0.717	0.723	0.72	**0.79**
20	0.715	0.731	0.719	0.723	0.789

the cysteine position in the sequence window are referred to as backward (BW) and forward (FW) subsequences, respectively. The window size (λ) is varied from 31 to 41 ($i.e.$,$|BW|$=$|FW|$=n) is varied from 5 to 20 and ($\lambda = (2 \times n + 1)$). Different length-wise experimental analysis has been carried out to find the optimal subsequence length (window size). Based on the AUC score, it has been found that the performance is optimum when n=19 (window size=2 × 19 + 1) as depicted in Table 7.2. Thus, the length of the subsequence in this approach is set to 19 for all consecutive experiments.

7.5 Feature Selection

To select the features, two different types of feature optimization strategies have been used for predicting the S-palmitoylation sites in mouse protein. The method includes a *KB* feature optimization strategy and a *GA* feature optimization strategy. Both strategies have been applied to three types of datasets as discussed above and their performances are being recorded and evaluated on the cross-validated test set, and hold-out test set. A detailed discussion of each feature optimization strategy is discussed below.

7.5.1 KB Feature Selection

KB feature selection strategy has been introduced to identify significant and non-redundant features from 566 physicochemical property-based PSAAP features. Initially, individual physicochemical property-wise performance has been evaluated with different varying subsequence lengths (31 to 41). Based on these performances, AUC score, physicochemical properties are sorted/ranked for individual subsequence lengths. Top-performing K features are extracted from each subsequence length-wise evaluation with four different thresholds of K *i.e.* top 25, 50, 75, and 100. Finally, two sets of features are constructed by considering the intersection of KB *i.e.* IB-k and the union of KB *i.e.* (UB-K) features from different length-wise evaluations.

Once retrieving these KB feature sets, performance has been evaluated with the merged feature where individual features are concatenated into a single feature vector

Table 7.3: . Performance of top K features.

Feature	Precision	Recall	Accuracy	F1	AUC
IB25	**0.724**	**0.717**	**0.722**	**0.72**	**0.79**
IB50	0.715	0.713	0.715	0.714	0.784
IB75	0.702	0.673	0.694	0.687	0.772
IB100	0.707	0.702	0.705	0.704	0.775
UB25	0.72	0.722	0.72	0.721	0.789
UB50	0.714	0.715	0.714	0.714	0.782
UB75	0.709	0.706	0.708	0.707	0.778
UB100	0.703	0.700	0.702	0.701	0.771

for final representation. The concatenated feature is generated for the window length 39 ($2 \times n + 1$, where n = 19) as it shows superior performance compared to other window lengths. The Union and Intersection-based performance evaluation with four different thresholds *i.e.* 25, 50, 75, and 100 are depicted in Table 7.4. Based on AUC and accuracy scores, it has been concluded that window length 39 with IB25 gives the best result with the highest AUC score among all as depicted in see Table 7.3, thus constituting the KB features. Figure 7.1 shows the detailed workflow for selecting the KB feature from the 566 feature set. Finally, the KB feature results in 19, 20, and 21 features in male, female, and the combined datasets, respectively.

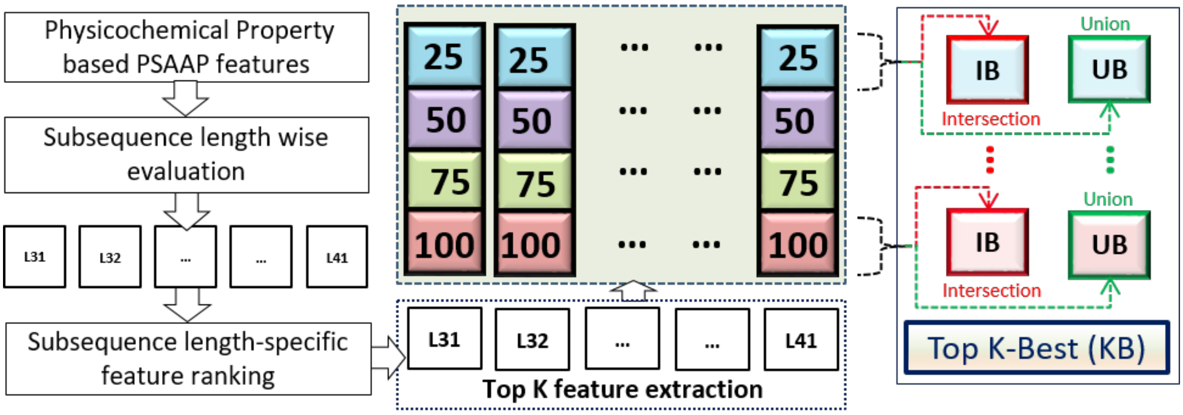


Figure 7.1: A detailed flow chart for K-Best feature selection.

Table 7.4: Detailed performance of K-Best features on different length of sub-sequences

Length	Intersection based Feature Set						Union based Feature Set					
	IB	**Pre**	**Rec**	**Accu**	**F1**	**AUC**	**UB**	**Pre**	**Rec**	**Accu**	**F1**	**AUC**
15	25	0.657	0.792	0.69	0.718	0.765	25	0.684	0.719	0.694	0.701	0.77
	50	0.69	0.718	0.698	0.703	0.766	50	0.686	0.724	0.696	0.704	0.764
	75	0.687	0.715	0.694	0.7	0.762	75	0.684	0.713	0.692	0.698	0.761
	100	0.681	0.709	0.689	0.695	0.757	100	0.683	0.698	0.686	0.69	0.758
16	25	0.701	0.731	0.709	0.715	0.781	25	0.699	0.724	0.706	0.711	0.78
	50	0.692	0.724	0.701	0.707	0.773	50	0.7	0.716	0.705	0.708	0.776
	75	0.697	0.716	0.702	0.706	0.772	75	0.689	0.723	0.699	0.705	0.769
	100	0.679	0.709	0.686	0.693	0.763	100	0.694	0.703	0.697	0.698	0.768
17	25	0.699	0.722	0.706	0.71	0.777	25	0.699	0.721	0.706	0.71	0.777
	50	0.698	0.713	0.702	0.705	0.772	50	0.696	0.715	0.702	0.705	0.77
	75	0.684	0.722	0.694	0.702	0.766	75	0.692	0.706	0.695	0.698	0.767
	100	0.688	0.706	0.693	0.697	0.766	100	0.685	0.712	0.693	0.698	0.763
18	25	0.72	0.731	0.723	0.725	0.788	25	0.717	0.733	0.722	0.724	0.787
	50	0.711	0.728	0.716	0.719	0.781	50	0.709	0.726	0.714	0.717	0.787
	75	0.704	0.725	0.71	0.714	0.774	75	0.709	0.728	0.715	0.718	0.777
	100	0.71	0.724	0.714	0.717	0.774	100	0.707	0.72	0.711	0.713	0.773
19	25	0.724	0.717	0.722	0.72	**0.79**	25	0.72	0.722	0.72	0.721	**0.789**
	50	0.715	0.713	0.715	0.714	0.784	50	0.714	0.715	0.714	0.714	0.782
	75	0.702	0.673	0.694	0.687	0.772	75	0.709	0.706	0.708	0.707	0.778
	100	0.707	0.702	0.705	0.704	0.775	100	0.703	0.7	0.702	0.701	0.771
20	25	0.715	0.731	0.719	0.723	0.789	25	0.716	0 .73	0.72	0.722	0.787
	50	0.709	0.728	0.715	0.718	0.783	50	0.712	0.728	0.717	0.72	0.782
	75	0.703	0.719	0.708	0.711	0.776	75	0.704	0.719	0.708	0.711	0.778
	100	0.707	0.719	0.71	0.712	0.775	100	0.702	0.713	0.705	0.707	0.775

N.B.:

Pre represents precesion.

Accu represents Accuracy.

IB represents Intersection based Feature Set.

UB represents Union based Feature Set.

Rec represents Recall.

7.5.2 GA Based Feature Selection

GA, which is inspired by the natural selection and evolution process, is a guided random optimized search technique that results in an excellent semi-optimal solution to the feature selection problem [504]. Under GA, fitter children, *i.e.* chromosomes, populated from the earlier generation *i.e.* parents have a better chance of survival. The feature subsets are encoded as chromosomes are considered as individuals and the collection of such chromosomes represents the population. Here, the chromosomes are encoded as a binary string where '1' at any position i of represents the selection of i-th feature and '0'represents the refusal. Each chromosome representing a subset of features is given a fitness score, which is obtained as the AUC in predicting the correct S-palmitoylation modification using this feature subset and RF classifier.

Initially, the 566 physicochemical properties are hierarchically clustered based on the amino acid properties. Then, the hierarchical cluster tree is partitioned into 331 non-singleton and 185 singleton clusters using the same splitting strategy proposed in [28]. In this experiment, GA has been used in two steps:

- First, GA is employed over the non-singleton clusters to obtain the best-performing feature among the cluster members.
- Second, GA is applied with the newly identified features from the non-singleton clusters and with the remaining features from singleton clusters.

In the method proposed here, RF is used for classification purposes while evaluating the performance of feature(s) at each generation. However, the AUC score is incorporated in fitness/objective computation. In this experiment, a roulette wheel selection strategy and uniform crossover are employed. The crossover probability (p) and uniform mutation probability (q) are set to 0.7 and 0.01, respectively, to populate the next generation chromosome. The positive and negative data ratio is kept as 1:1 for evaluation purposes. The tie between equally performing chromosomes, the one with the lesser number of features, is retained. The method results in the globally best chromosomes. Finally, the GA-based approach identified 6 features in males, 7 in females, and 21 features in the combined dataset, respectively, for final classification. The overall workflow of GA-based feature design is detailed in Figure 7.2.

7.5.3 Classifier Selection

To check the efficacy of the classifier, the PTM prediction method has been evaluated on a subset of the dataset using the KB feature s described above with two machine learning algorithms SVM [251] and RF [501]). The performance of both classifiers is presented in Table 7.5. Based on the AUC, F1, and Accuracy scores, RF outperforms SVM. Thus, all the experiments presented in the main manuscript are carried out

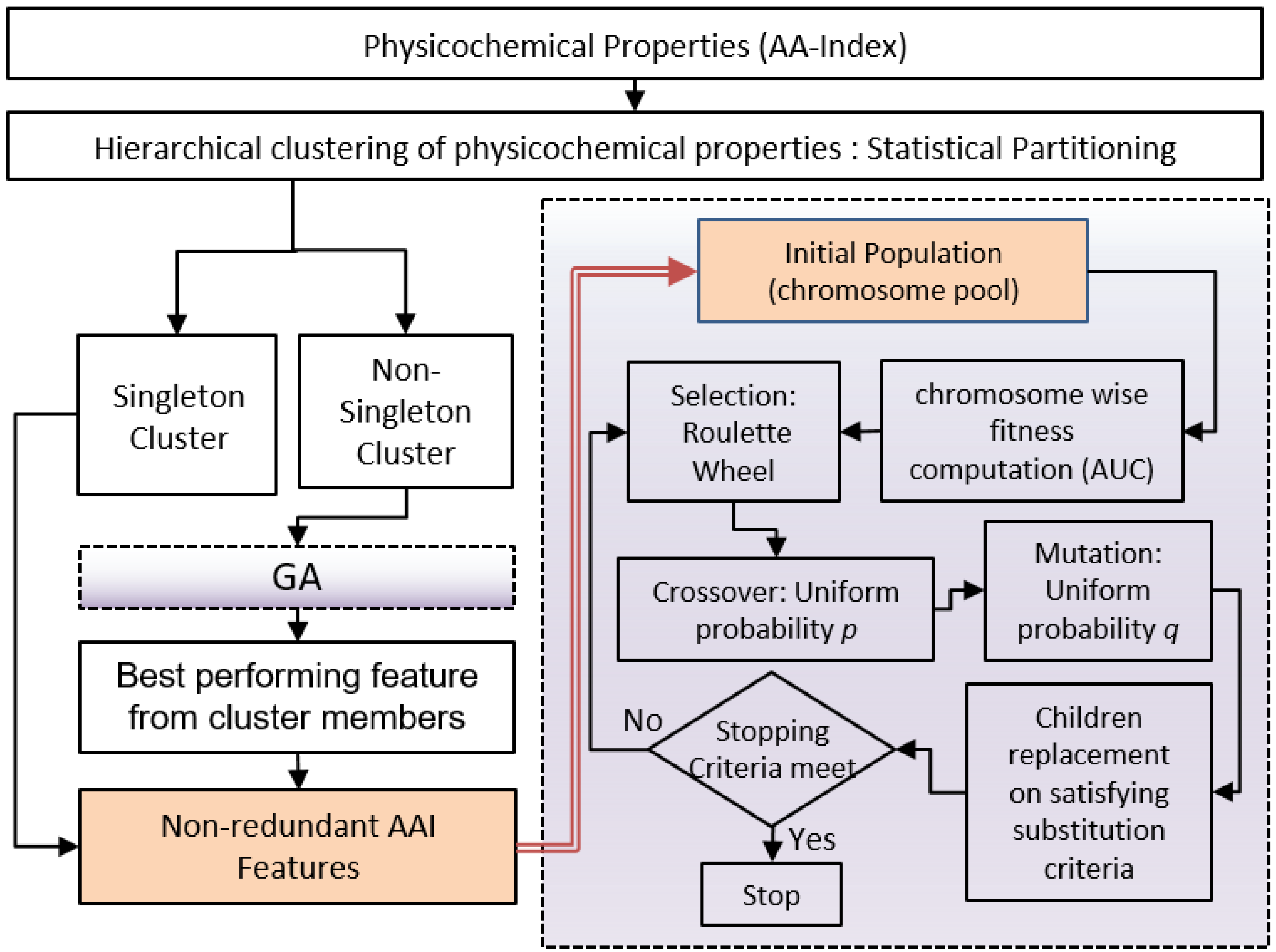

Physicochemical Properties (AA-Index)
Hierarchical clustering of physicochemical properties : Statistical Partitioning
Singleton Cluster
Non-Singleton Cluster
GA
Best performing feature from cluster members
Non-redundant AAI Features
Initial Population (chromosome pool)
chromosome wise fitness computation (AUC)
Selection: Roulette Wheel
Crossover: Uniform probability *p*
Mutation: Uniform probability *q*
Children replacement on satisfying substitution criteria
Stopping Criteria meet
No
Yes
Stop

Figure 7.2: A Detailed workflow of GA-based feature selection.

Table 7.5: Performance evaluation on SVM and RF classifier.

Classifier	Precision	Recall	Accuracy	F1	AUC
SVM	0.671±0.03	0.620±0.08	0.652±0.01	0.720±0.018	0.636±0.03
RF	0.669±0.08	0.650±0.02	**0.664±0.01**	**0.728±0.01**	0.659±0.01

using RF on all three different types of mouse data *i.e.* Male, Female, and Combined. For each fold, the positive and negative samples are selected in equal ratios for both train and test sets to obtain an unbiased classification.

7.6 Experimental Results

The proposed method, predicts the S-palmitoylation sites from the primary sequence information of synaptic proteins. In the mouse model experiments, three categories of data, *viz.,* Male, Female, and Combined, and three different feature sets, *viz.,* KB, GA, UN, along with the RF classifier, have been used. The rationale behind the choice of the RF classifier is elaborated in Table 7.5. Features are extracted from the sequence motifs of variable length, and detailed experiments are conducted to select the optimum length of such sequence motifs. These experiments are summarized in section 7.3 and detailed results are described in Table 7.4. Finally, the proposed approach presents a three-star consensus model for the final classification task. The efficacy of PTM prediction depends heavily on selecting appropriate feature sets, the choice of the classifier, and the underlying evaluation strategy. In this work, GA-based features show better the AUC score for male, female, and combined datasets. The UN features show promising performances for the female dataset with higher accuracy, whereas KB and GA features achieve the highest accuracy in male and combined datasets, respectively. Finally, a three-star approach has been presented for the final classification task The consensus model significantly improved the performance compared to individual feature-specific models. The proposed consensus-based approach, has been further considered with two *state-of-the-art methods.*

7.6.1 Performance Evaluation

The performance of the proposed model has been evaluated with five-fold cross-validation on three different feature sets *viz.* KB, GA, UN, using an RF classifier. Five-fold cross-validation has been introduced to estimate the model's strength on all three categories of datasets, male, female, and combined, and the performances are reported in Table 7.6. The individual fold-wise performances on all three datasets are reported in Table 7.9. In all three datasets, the GA-based feature outperforms the rest two in AUC score. However, in the proposed method, for fold-wise testing,

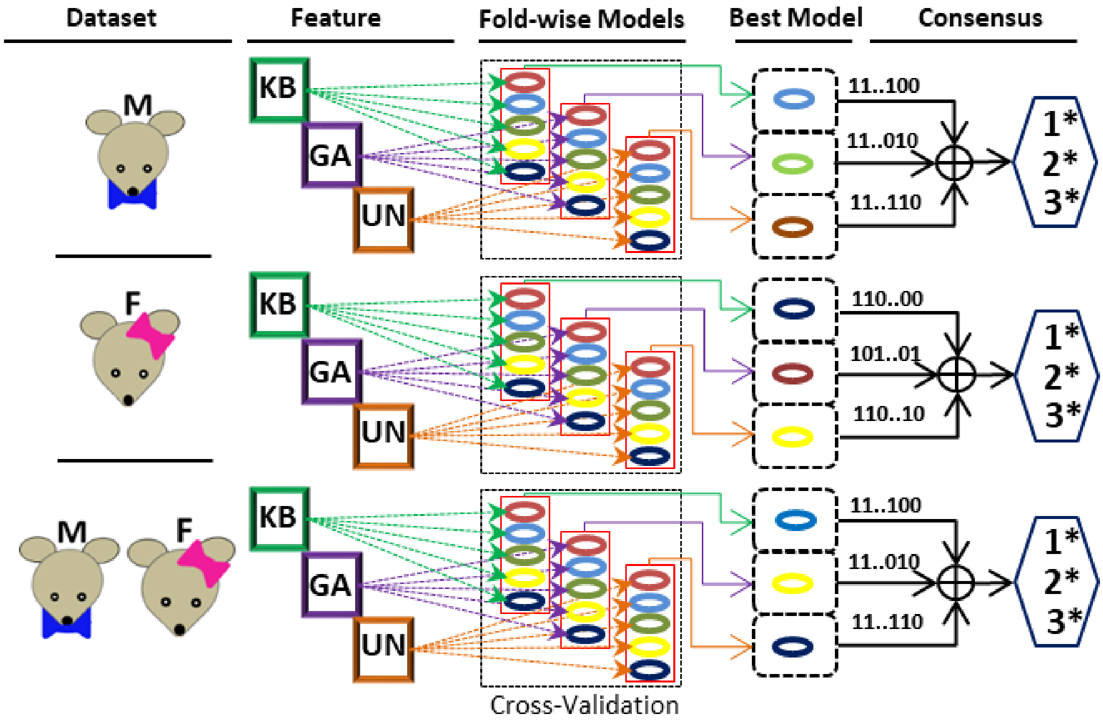


Figure 7.3: A schematic diagram depicting the underlying consensus strategy for S-palmitoylation prediction.

the GA-based feature shows a ∼ 79% Average AUC (AvgAUC) score for both male and combined datasets, and ∼80% AUC on the female dataset, surpassing the other two features. For female data, the UN-based feature outperforms KB and GA-based features, having an accuracy score of ∼71.9% and F1 of ∼71.3% (see Table 7.6). The AUC and AUPRC curves from training models are shown in Figure 7.4.

The KO data has been used as the hold-out test set from three categories of data *viz.*Male, Female, and Combined individually. In the KO hold-out test set, the GA-based feature shows better performance for all the datasets than other features with an AUC score of ∼ 66.4% in males. ∼68.6% in females, and ∼62.5% in combined dataset (see Table 7.7). Moreover, GA has higher accuracy in all hold-out test data except the males set, where the KB-based model achieves ∼62% accuracy. Furthermore, we have introduced a consensus strategy for the final classification of S-palmitoylation on the hold-out test set. Initially, the best models are extracted from the cross-validation strategy for each feature set on the three categories of data set independently.

7.6.2 Comparison with the *State-of-the-Art* Approaches

To demonstrate the performance of the proposed method, the proposed approach has been compared with existing PTM prediction models. Three *state-of-the-art* method have been identified for benchmark purposes.

Table 7.6: : Performance evaluation of 5-fold cross-validation on Male, Female, and Combined dataset with KB, GA and UN features.

Type	**Feature**	**5-Fold Cross-Validation**					
		AvgAUC	**MaxAUC**	**Pre**	**Recall**	**Accu**	**F1**
Male	KB	0.785 ± 0.0137	0.801	0.732±0.02	0.666 ± 0.02	0.711 ± 0.01	0.697 ± 0.016
	GA	0.790 ± 0.013	0.812	0.726 ± 0.02	0.675 ± 0.02	0.710 ± 0.02	0.700 ± 0.02
	UN	0.786 ± 0.013	0.798	0.726 ± 0.02	0.662 ± 0.01	0.706 ± 0.01	0.693 ± 0.01
Female	KB	0.796 ± 0.02	0.82	0.715 ± 0.02	0.701 ± 0.02	0.708 ± 0.02	0.706 ± 0.016
	GA	0.801 ± 0.018	0.827	0.732 ± 0.02	0.69 ± 0.04	0.718 ± 0.02	0.709 ± 0.02
	UN	0799 ± 0.018	0.821	0.729 ± 0.02	0.698 ± 0.03	0.719 ± 0.02	0.713 ± 0.02
Combined	KB	0.791 ± 0.02	0.830	0.718 ± 0.04	0.689 ± 0.02	0.708 ± 0.03	0.703 ± 0.03
	GA	0.795 ± 0.02	0.830	0.733 ± 0.03	0.684 ± 0.03	0.717 ± 0.02	0.707 ± 0.03
	UN	0.793 ± 0.02	0.820	0.734 ± 0.02	0.670 ± 0.01	0.714 ± 0.02	0.701 ± 0.02

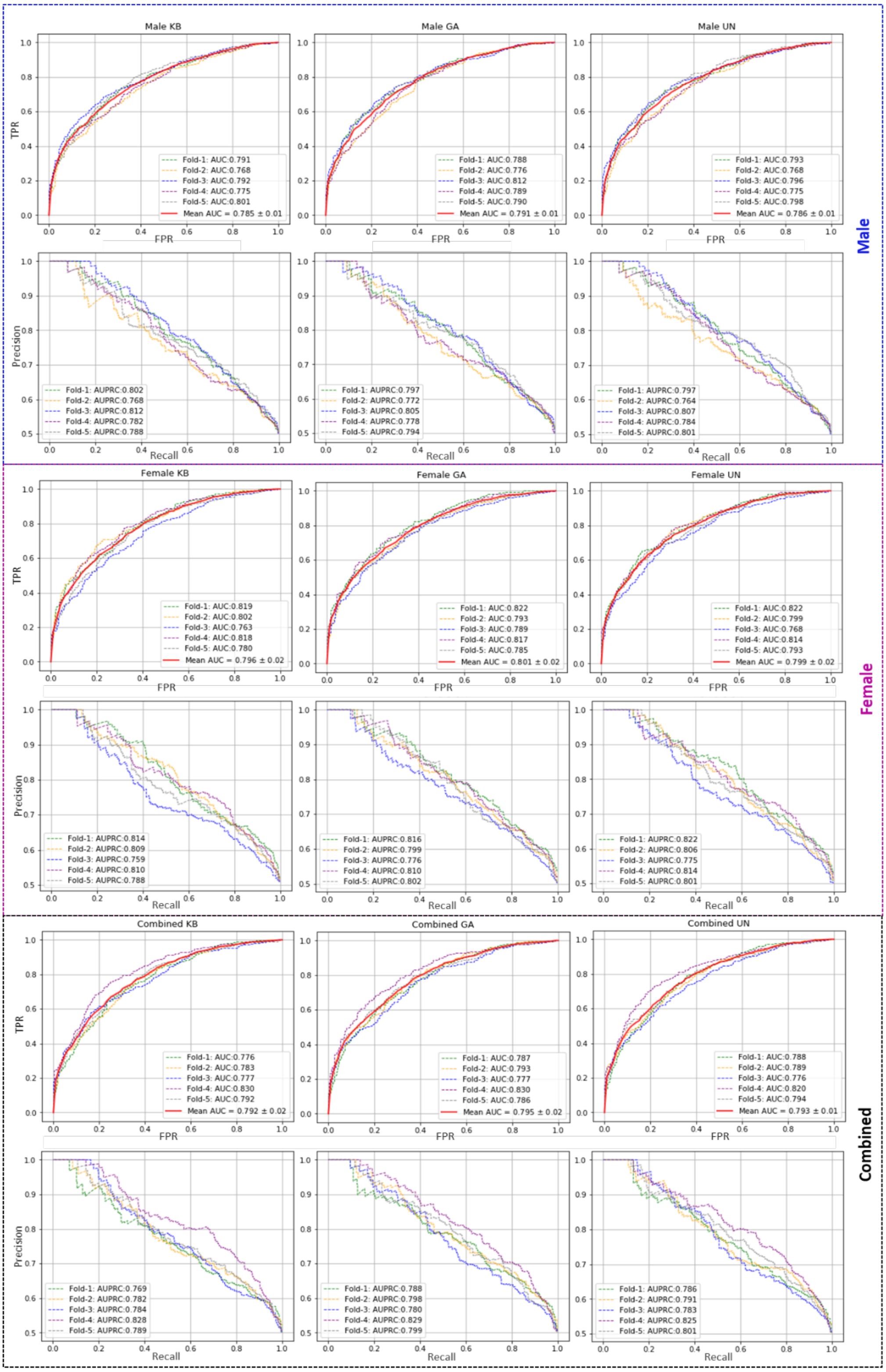


Figure 7.4: Performance evaluation on three datasets, Male, Female, and Combined. Plots in the 1st, 3rd, and 5th rows show the AUC, and the 2nd, 4th, and 6th rows represent AUPRC, respectively. The 1st, 2nd, and 3rd column-wise plots represent KB, GA, and UN type features-based evaluation.

Table 7.7: Performance evaluation using fold-wise and consensus strategy on hold-out test data.

Dataset		Feature	Pre	Rec	Accu	F1	MCC	AUC
Male	FW	KB	0.643 ± 0.01	0.54 ± 0.02	0.620 ± 0.01	0.587 ± 0.01	0.244 ± 0.02	0.661 ± 0.01
		GA	0.629 ± 0.01	0.535 ± 0.02	0.609 ± 0.01	0.578 ± 0.01	0.222 ± 0.02	0.664 ± 0.01
		UN	0.634 ± 0.02	0.532 ± 0.01	0.612 ± 0.01	0.579 ± 0.01	0.227 ± 0.03	0.661 ± 0.01
	Con	1* Con	0.585	0.812	0.618	0.68	0.255	0.639
		2* Con	0.667	0.713	0.678	0.689	0.357	0.679
		3* Con	0.676	0.423	0.610	0.520	0.238	0.628
Female	FW	KB	0.617 ± 0.01	0.566 ± 0.01	0.608 ± 0.01	0.591 ± 0.01	0.216 ± 0.02	0.667 ± 0.01
		GA	0.641 ± 0.01	0.600 ± 0.01	0.632 ± 0.01	0.62 ± 0.01	0.265 ± 0.01	0.686 ± 0.004
		UN	0.622 ± 0.01	0.566 ± 0.02	0.611 ± 0.01	0.593 ± 0.01	0.223 ± 0.02	0.684 ± 0.004
	Con	1* Con	0.593	0.792	0.624	0.678	0.264	0.64
		2* Con	0.799	0.706	0.764	0.749	0.532	0.768
		3* Con	0.800	0.447	0.668	0.573	0.373	0.708
Combined	FW	KB	0.586 ± 0.02	0.475 ± 0.01	0.57 ± 0.01	0.525 ± 0.01	0.142 ± 0.02	0.597 ± 0.01
		GA	0.608 ± 0.02	0.486 ± 0.02	0.586 ± 0.02	0.54 ± 0.02	0.176 ± 0.03	0.625 ± 0.01
		UN	0.605 ± 0.02	0.472 ± 0.02	0.581 ± 0.02	0.53 ± 0.02	0.167 ± 0.03	0.615 ± 0.01
	Con	1* Con	0.654	0.719	0.669	0.685	0.340	0.671
		2* Con	0.679	0.669	0.676	0.674	0.353	0.676
		3* Con	0.612	0.374	0.568	0.464	0.148	0.580

FW=Fold-Wise Score
Con=Consensus Score

Table 7.8: Performance comparison with the *state-of-the-art methods* for S-palmitoylation prediction

Methods	AUC	AUPRC	Accu	F1	MCC
CapsNet [229]	0.780 ± 0.02	0.500 ± 0.07	NA	NA	NA
MusiteDeep [230]	0.771 ± 0.02	0.484 ± 0.05	NA	NA	NA
ModPred [221]	0.8553 ± 0.01	0.5973 ± 0.04	NA	NA	NA
Proposed Method (1:1)	0.936 ± 0.01	0.889 ± 0.02	0.824 ± 0.03	0.799 ± 0.04	0.669 ± 0.05
Proposed Method (1:2)	0.928 ± 0.02	0.785 ± 0.04	0.816 ± 0.02	0.645 ± 0.06	0.577 ± 0.06

The three *state-of-the-art* method that have been identified for benchmark are CapsNet [229], MusiteDeep [230], and ModPred [221]. The CapsNet [229] is a deep learning-based architecture that provides prediction models for different PTM sites. MusiteDeep [230] [505] is a deep-learning-based system that can predict general and kinase-specific phosphorylation sites from primary sequence information. ModPred [221] is a sequence-based PTM prediction tool developed on the structural and functional signatures of proteins. The CapsNet, provides a 10-fold cross-validation result on the benchmark dataset of animal species *viz.* metazoa, extracted from the NCBI taxonomy database [506] which has been curated by collecting annotations from Uniprot/Swiss-Prot [507] with less than 30% sequence similarity.

The proposed approach has also been trained with the similar dataset used in CapsNet [229] for S-palmitoylated cysteine prediction for comparison purposes. When compared with all three existing approaches on similar datasets, the performance scores are directly incorporated from Wang et al. [229]. In the proposed model, class-imbalanced learning has also been proposed by imposing a positive-negative ratio at 1:2 along with the balanced learning (1:1). The performance has been compared with the existing approaches concerning the AUC and AUPRC scores (see Table 7.8). The proposed method outperforms the *state-of-the-art* methods in both metrics. The AUC and AUPRC have improved by 8% in comparison with the earlier best-performing method. Additionally, the proposed approach has surpassed the prior approaches by 32% in the AUPRC score, as depicted in Table 7.8. The detailed fold-wise evaluation scores are shown in Table 7.10 and Table 7.11 with balanced and imbalanced datasets respectively.

Table 7.9: : Performance evaluation of 5-fold cross-validation on Male, Female, and Combined dataset with KB, GA, and UN features.

Dataset	Feature	Fold	Pre	Rec	Accu	AUC	F1	MCC
Male	KB	0	0.748	0.65	0.715	0.791	0.695	0.434
		1	0.717	0.644	0.695	0.768	0.679	0.392
		2	0.736	0.684	0.719	0.792	0.709	0.44
		3	0.699	0.671	0.691	0.775	0.685	0.383
		4	0.761	0.682	0.734	0.801	0.719	0.47
	GA	0	0.714	0.668	0.701	0.788	0.691	0.402
		1	0.702	0.65	0.687	0.776	0.675	0.375
		2	0.762	0.701	0.741	0.812	0.73	0.483
		3	0.714	0.687	0.706	0.789	0.7	0.412
		4	0.74	0.668	0.717	0.79	0.702	0.435
	UN	0	0.734	0.658	0.71	0.793	0.694	0.422
		1	0.716	0.639	0.693	0.768	0.675	0.387
		2	0.735	0.674	0.715	0.796	0.703	0.432
		3	0.701	0.663	0.69	0.775	0.681	0.38
		4	0.747	0.679	0.725	0.798	0.711	0.451
Female	KB	0	0.725	0.723	0.725	0.819	0.724	0.449
		1	0.718	0.698	0.712	0.802	0.708	0.424
		2	0.676	0.715	0.686	0.763	0.695	0.373
		3	0.74	0.709	0.73	0.818	0.724	0.461
		4	0.698	0.658	0.686	0.78	0.677	0.373
	GA	0	0.742	0.684	0.723	0.822	0.712	0.448
		1	0.718	0.698	0.712	0.793	0.708	0.424
		2	0.712	0.746	0.722	0.789	0.728	0.444
		3	0.762	0.686	0.736	0.817	0.722	0.474
		4	0.727	0.638	0.699	0.785	0.68	0.401
	UN	0	0.744	0.723	0.737	0.822	0.734	0.475
		1	0.713	0.675	0.702	0.799	0.694	0.405
		2	0.697	0.729	0.706	0.768	0.713	0.413
		3	0.755	0.698	0.736	0.814	0.725	0.473
		4	0.738	0.667	0.715	0.793	0.7	0.431
Combined	KB	0	0.708	0.617	0.697	0.776	0.689	0.395
		1	0.691	0.695	0.692	0.783	0.693	0.385
		2	0.673	0.688	0.677	0.777	0.68	0.353
		4	0.774	0.726	0.757	0.83	0.749	0.515
		4	0.747	0.668	0.721	0.792	0.706	0.445
	GA	0	0.724	0.673	0.708	0.787	0.697	0.417
		1	0.731	0.659	0.708	0.793	0.693	0.418
		2	0.699	0.692	0.697	0.777	0.696	0.394
		3	0.779	0.731	0.762	0.83	0.754	0.525
		4	0.731	0.666	0.71	0.786	0.697	0.422
	UN	0	0.727	0.685	0.714	0.788	0.705	0.429
		1	0.722	0.656	0.702	0.7989	0.688	0.406
		2	0.709	0.656	0.694	0.776	0.682	0.388
		3	0.763	0.688	0.737	0.82	0.723	0.476
		4	0.753	0.668	0.725	0.794	0.708	0.452

To investigate the significance of the proposed model on a novel S-palmitoylation dataset, it has been evaluated and compared the performance with two web servers MusiteDeep [230, 505] and CSS-Palm [228]. MusiteDeep [230, 505] is a web resource with a deep-learning framework that can predict and visualize different PTM sites

Table 7.10: : Performance evaluation of proposed methods for S-palmitoylation prediction on the dataset given in [229] with balanced learning (Positive: Negative=1:1)

Fold	Precision	Recall	Accuracy	AUC	AUPRC	F1	MCC
1	0.922	0.714	0.827	0.94	0.89	0.671	0.805
2	0.897	0.696	0.808	0.92	0.872	0.632	0.784
3	0.934	0.749	0.848	0.941	0.904	0.71	0.831
4	0.963	0.678	0.826	0.934	0.901	0.682	0.796
5	0.954	0.649	0.809	0.945	0.889	0.652	0.772
6	0.942	0.649	0.804	0.942	0.883	0.641	0.768
7	0.945	0.828	0.89	0.953	0.929	0.786	0.883
8	0.909	0.633	0.785	0.917	0.863	0.598	0.746
9	0.893	0.726	0.819	0.927	0.878	0.65	0.801
10	0.924	0.708	0.825	0.945	0.889	0.668	0.802

Table 7.11: : Performance evaluation of proposed methods for S-palmitoylation prediction on the dataset given in [229] with class imbalanced learning (Positive: Negative=1:2)

Fold	Precision	Recall	Accuracy	AUC	AUPRC	F1	MCC
1	0.971	0.508	0.746	0.938	0.862	0.667	0.56
2	0.945	0.525	0.747	0.913	0.854	0.675	0.552
3	0.965	0.525	0.753	0.935	0.864	0.68	0.568
4	0.98	0.552	0.77	0.935	0.878	0.706	0.601
5	0.985	0.463	0.728	0.949	0.858	0.63	0.538
6	0.96	0.488	0.734	0.933	0.852	0.647	0.537
7	0.953	0.63	0.8	0.949	0.884	0.759	0.637
8	0.914	0.422	0.691	0.908	0.812	0.578	0.454
9	0.934	0.452	0.71	0.904	0.83	0.61	0.491
10	0.97	0.51	0.747	0.94	0.863	0.668	0.561

Table 7.12: Performance comparison with MusiteDeep [230] [505] and CSS-Palm [228] web server with holdout dataset.

Method	Type of Data	Pre	Rec	Accu	F1	MCC
MusiteDeep	**Male**	0.827	0.088	0.535	0.159	0.155
	Female	0.808	0.107	0.51	0.188	0.151
	Combined	0.555	0.0719	0.507	0.127	0.029
CSS-Palm High Threshold	**Male**	0.857	0.132	0.555	0.229	0.206
	Female	0.783	0.147	0.524	0.247	0.168
	Combined	0.75	0.129	0.543	0.22	0.153
CSS-Palm Medium Threshold	**Male**	0.768	0.158	0.555	0.262	0.182
	Female	0.761	0.177	0.532	0.288	0.173
	Combined	0.735	0.179	0.557	0.289	0.176
RFCF-PALM	**Male**	0.628	0.539	0.609	0.58	0.222
	Female	0.639	0.583	0.627	0.61	0.254
	Combined	0.623	0.504	0.599	0.556	0.202

from protein sequence information. CSS-Palm [228] is developed based on a clustering and scoring strategy (CSS) algorithm and Group-based Prediction System (GPS) algorithm. CSS-Palm is evaluated with two high-performing thresholds, as stated by the authors in [228]. The novel hold-out test data from male, female, and combined sets has been submitted to the above two web servers, and performances have been recorded for comparison purposes (see Table 7.12). The proposed method has achieved a better result in more balanced metrics F1, and MCC compared to each of these web servers in S-palmitoylation prediction depicting the efficacy of the proposed method on S-palmitoylation prediction. In all three datasets, male, female, and combined, the proposed approach has improved the F1 score by 54%, 52%, and 48%, and MCC score by 7%, 32%, and 13%, respectively.

In this novel hold-out data set, both web servers show high precision having 0.827 in MusiteDeep and 0.857 in CSS-Palm and very low recall having 0.0882 in MusiteDeep and 0.1324 in CSS-Palm. A high precision score depicts low false positivity, and low recall depicts the increase in false-negative data, which can be interpreted as a failure for predicting the positive data. This may lead to a biased classification. Low recall also results in a low F1 score, the harmonic mean of precision and recall. Not only the recall score, but the MCC score for both the web servers are low, which depicts the failure of the class imbalance issue [508]. In contrast, the proposed method achieves 0.638 precision, and 0.583 recall scores on this hold-out dataset, which shows a more balanced scenario of classification outcome. In addition, the proposed method shows the highest accuracy for all three categories of the data, which outperforms the other two where accuracy improvement is by 9%, 15%, and 7% in male, female, and the combined dataset.

7.7 Discussion

The method proposed here, computationally predicts the S-palmitoylation sites using the primary sequence information of the synaptic group of proteins from three categories of mouse data, designed as sex-dependent *i.e.* male and female and sex-independent *i.e.* combined mode. The computational model has been developed through a rigorous feature selection strategy and optimal model selection for predicting the S-palmitoylation modification sites in a given sub-sequence window. The proposed model has been evaluated with five-fold cross-validation, and model performances have been compared with the *state-of-the-art* approaches using three different feature sets; KB, GA, and UN. Finally, a consensus strategy is designed based on the feature-specific best models from their cross-validated models. The performance of the consensus model improved significantly compared to *state-of-the-art*. The significant performance improvement in predicting S-palmitoylation modification sites portrayed the efficacy of the proposed method. In a nutshell, the proposed method has been developed with effective feature selection and consensus strategy for *in-silico* prediction of S-palmitoylation in mouse protein and shows significant improvement.

Chapter 8

Conclusion

In order to find answers to fundamental concerns about how life functions, biology provides a wealth of knowledge that may be exploited. The majority of the genetic instructions required to produce proteins are found in the DNA of a cell. Numerous crucial biological activities occurring inside a cell are carried out by proteins interacting with one another. In computational biology, examining the contact affinities of protein pairs is an important research issue for revealing cellular and molecular capacities, signaling cascades, and crucial insights into disease causes.

After the Human Genome Project was finished, genetic sequence data was massively increased. Proteomic data has grown at the same exponential rate as genomic data in many biological contexts thanks to the rise of high-throughput studies. Sequence annotation, structural details, ontology links, functional descriptions, interaction networks, PTMs, diseases linked to specific genes, etc. UniProt, a freely accessible protein data database, has around 180 million unreviewed and roughly 500,000 manually annotated and reviewed proteins.

PPINs are the most important factor in determining which proteins are responsible for various disease conditions and for studying how diseases spread from pathogen to host. They are also very helpful in determining the function of certain proteins. Analyzing host-pathogen networks provides the necessary insight into disease transmission processes for drug creation. Because of their capacity for mutation, pathogens aid in the spread of illness. When a pathogen infects a host, it spreads along the interface between the two. Therefore, it is crucial to investigate target proteins and their interactions within the host-pathogen network in order to identify new therapeutic targets.

Proteins may undergo covalent modifications at specific amino acid residues via a biochemical process known as PTM. Proteins have the ability to break into pieces or add sugar, phosphate, or sulfate groups to surface residues through covalent bonds. Protein interactions influence and control many different protein activities through PTM. However, in biological knowledge databases, only 4% of PPIs have PTM annotations [509].

In a dynamic PPIN of this magnitude, diversity and trustworthiness are major factors. Using current computational tools to manage such a huge, dynamic, heterogeneous network is a time-consuming task.

With these considerations in mind, the research presented in this thesis focuses on four key areas: 1) analyzing computational strategies for large-scale biological data, 2) *in-silico* computation of interaction affinity of human proteome in *fuzzy* semantic space, 3) *in-silico* analysis of host-pathogen protein interaction network and identifying repurposed drug, and 4) *in-silico* prediction of PTM sites in protein sequences.

It has become extremely difficult for researchers to process these lengthy and repeated sequences. One of the key stages to reducing the duplication of these immense resources and analysis of such massive biological sequences is clustering by similarity. For greater values of n, the *n-gram* feature representation, which is typically employed in sequence clustering and classification, produces high dimensional input spaces. However, because there are so many dimensions, it becomes impossible to cluster such massive sequences using present techniques. By harnessing the power of parallel computing with high-performance computing platforms, an effectively designed clustering technique may quickly scale to accommodate large-size sequences.

In Chapter 2, addressing the computing challenges of processing and analyzing large-scale biological data, a two-level parallel DBSCAN clustering for human protein sequences is proposed. Using parallel computer resources, the DBSCAN technique may be used effectively for high-dimensional input spaces and large-scale human sequencing data. The suggested technique was put into practice on a spark cluster, which ultimately aids in the resource and computing facility's parallelization.

The results of the experiments demonstrated the effectiveness of the suggested strategy in speeding up the process and eliminating unnecessary sequences. The suggested technique using the trigram feature (n=3) outperforms *state-of-the-art* methods by increasing the proportion of non-singleton clusters with Domain Correspondence Score=1. According to the data, the clustering results become better with larger values of n, and the speedup ratio gets better as the data size grows.

All living things depend on protein interactions for cellular and biological processes. Following, we provide a quick overview of the state of many interconnected project-related application domains where high throughput parallel architecture is a crucial factor.

On the basis of the existing high-throughput experimentally validated positive and negative interactions, machine learning-based PPI prediction models are built. Except for a select few, few people are aware of the empirically confirmed PPNI databases. The creation of the models in the current techniques utilized the negative data created randomly. This haphazard selection of data samples yields an imprecise PPI predictive model, which can lead to inaccurate prediction. Negative data is just as crucial as positive data in supervised classification, therefore careful selection of negative data makes the classifier resilient towards PPI prediction.

PPIs must be understood in order to comprehend how a protein works. Combining the ontological connection between any two protein pairs will yield an estimate of their interaction affinity. GO-based methods offer a more accurate and reliable PPI evaluation framework. In Chapter 3, a unique method for determining the interaction affinity between protein pairs has been proposed by using a hybrid approach of GO graph. The method may take into account the three ontological links in order to characterize the protein binding affinities into a fuzzy score that lies between [0,1]. The larger the interaction affinity, the higher the score. The PPI interactions are represented in this scheme in a weighted form, making it a useful input object for pathway analysis, illness identification, medication development, etc.

Large-scale PPI analysis at the proteome level is labor and time-intensive, and sometimes computing becomes unsolvable for an organism like a human. For instance, the set of 19,000 evaluated human proteins may interact in 180,000,000 different ways. Similar to this, there are 14 billion potential interactions for the 170000 unreviewed human proteins. It has been suggested to use a distributed environment and a GO-based graph-theoretic technique to map all of these large-scale interactions at fuzzy semantic space.

Finally, classifying the protein interactions qualitatively based on affinity scores is one of the main focuses. It can be seen that the *Fuzzy*PPI may be effective for both high-quality negative data selection and high-quality positive interactions with extremely low FPR. The work's efficacy has also been confirmed by comparison with the relevant *state-of-the-art* methods

PPI prediction has been a popular issue in proteomic research for years. Different PPI-based approaches provide crucial data for future investigation of infection pathways between various species. Pathogenic microorganisms can influence host systems to harness host capabilities and evade host immune responses thanks to PPIs between the virus and host proteins. Therefore, for the creation of novel and more potent treatments, a thorough knowledge of infection processes via PPIs is essential.

In Chapter 4, the SARS-CoV-Human PPIN network and the spreader nodes at both level-1 and level-2 using the SIS model have been used. When calculating the protein interaction affinity score to identify the level-1 human spreaders of nCoV, these spreader nodes are taken into account. To make this model more powerful and meaningful, GO annotations and PPIN attributes have also been taken into account. The suggested model's identification of the chosen human spreader nodes as prospective protein targets for the COVID-19 FDA-approved medications has been seen as the study has progressed. The proposed model evaluates high-quality positive interactions in human-nCoV PPIN. A critical threshold has been identified that gives a balanced false-positive rate. Finally, it has been suggested that the created computational model successfully and very specifically detects Human-nCoV PPIs. The interactions between nCoV and humans are deduced from SARS-CoV, another pandemic starter that has a great deal of genetic similarity with nCoV. It has also been acknowledged that the SIS model was used to validate the spreadability index up to level-2 of the human spreader proteins. The number of proteins grows with each level change in human contact networks because of the high network density. As a result, the spreadability score confirmed by the SIS model may be used by the proposed approach to identify human spreader proteins at level-2.

In chapter 5, it has been tried to find potential treatments for the level-1 and level-2 spreader proteins by analyzing the human-nCoV PPIN. The analysis names the FDA-approved medication Fostamatinib/R406 as the most promising medication with the best potential to target the COVID-19 spreading proteins. The research depends on the assertion that SARS-CoV2/nCoV shares 89% of its genetic makeup with SARS-CoV. Based on this, the human-nCoV PPIN was created, and the SIS model and fuzzy thresholding were used to determine the spreader nodes of the PPIN. Additionally, a two-way analytic consensus technique has been used to evaluate medications based on the overlap of spreader proteins and drug-protein targets. In the analysis of the potential treatments for COVID-19 spreading proteins, the consensus ratings for Fostamatinib/R406 are the highest. Additionally, Fostamatinib/R406 produces positive results in molecular docking with the COVID-19 protein structures that are currently accessible.

In, chapter 6, a complete human-coronavirus interactome has been proposed. In the family Coronaviridae, a coronavirus is a member. A new coronavirus called SARS-CoV2 replicates by interacting with the host proteins. Therefore, identifying viral and host PPIs might aid researchers in better understanding the manner in which viruses spread illness and help them discover potential COVID-19 medicines. It impacts not just people but other animals and birds. In addition to severe acute and chronic respiratory diseases, multiple organ failure, and eventually human mortality, the coronavirus also often causes the common cold, cough, etc.

As vaccine and medicine development can take years, drug repurposing is a potent method that provides new therapeutic alternatives by exploring additional uses for already-approved pharmaceuticals. The conventional conservative drug development method, which is limited to ”one drug, one target” paradigms, does not consider or evaluate the probability of various pharmacological indications or off-target effects. All level-1 Coronavirus human proteins are mapped with the relevant drugs via DrugBank upon the creation of the coronavirus-human PPIN. It is noted that fostamatinib has the highest frequency of occurrence in the entire PPIN and has a significant overlap of target proteins in the human-coronavirus PPIN with the highest DCS of 181 when compared to the other human protein-associated medications. As already discussed and proposed by [352], fostamatinib has the highest DCS score for level-1 and level-2 spreader proteins, Thus the drug of concern shifted to the one with the next highest score, copper. The purpose of Copper is to examine the effects of a highly specialized medication called "Hinokitiol Copper Chelate" on massive amounts of 2019-nCoV Spike Glycoprotein with a single receptor binding domain. Supplemental zinc is also essential in the fight against several coronavirus species. For the healthy functioning of the human immune system, zinc is crucial for maintaining natural tissue barriers like the respiratory epithelium, which block the entry of pathogens. Promethazine, an antipsychotic medicine with clathrin-mediated endocytosis that has been used to treat COVID-19 due to its close genetic resemblance to SARS-CoV-2 and SARS-CoV, is one of the most effective medications for SARS-CoV and MERS-CoV.

S-palmitoylation is a post-translational covalent modification of the side chain of cysteine thiol by palmitic acid. S-palmitoylation is involved in various human disorders and is essential in a number of biological activities. In Chapter 7, an RF-classifier-based prediction strategy has been proposed to predict palmitoylated cysteine sites on synaptic proteins from three categories of mouse data *viz.*male, female, and sex-independent. A heuristic method for choosing the best collection of physicochemical characteristics from the AA-Index dataset has been developed. It makes use of the KB, GA, and UN of KB and GA-based features. A rigorous feature selection technique and the best model for predicting the S-palmitoylation alteration sites in a certain sub-sequence window were used to create the computational model. Three separate feature sets—have been used to assess the proposed model using five-fold cross-validation, and model results have been compared with *state-of-the-art* techniques. On the basis of the feature-specific best models from their cross-validated models, a consensus strategy is then developed. When compared to *state-of-the-art*, the consensus model performed much better.

Scope for future study

While designing the two-level parallel DBSCAN algorithm using Apache spark framework, bi-gram, and tri-gram features have been used to show the speed-up efficiency and cluster quality. In future, the method can be extended to operate on higher dimensions with a higher value of n, where n$\geq$3, and even large datasets at the cross-organism level.

In designing large-scale interactions for human proteome at fuzzy semantic space, the work has been done using reviewed human protein having GO annotations. The work can be extended for both reviewed and unreviewed proteins and also for multi-organism PPI prediction problems, leading to more than a trillion interactions.

In designing a computational model for human-nCoV protein interaction network, the work highlights Fostamatinib/R406 as one of the potential drugs for SARS-CoV2. This motivates to further do a drug repurposing study on the generated SARS-CoV2-human PPIN.

As the study shows Fostamatinib/R406, an FDA-approved drug, as the most promising drug with the best chances of targeting the COVID-19 spreader proteins, a clinical test is needed as the FDA approves Fostamatinib/R406 in ITP and to determine its efficacy against SARS-CoV-2. According to the reports, Fostamatinib meets the "primary endpoint of Safety in Phase 2 Clinical Trial" conducted in hospitalized patients affected with COVID-19. This Phase 3 trial is highly desirable.

Analyzing the Human-coronavirus family interactome and identifying target proteins for FDA-approved drugs using level-1 spreader nodes, apart from Fostamatinib, Promethazine is also one of the potential drug candidates for coronavirus-related diseases under clinical trials. The study can be extended to identify target proteins using level-2 spreader nodes.

S-Palmitoylation site prediction may further be enhanced by incorporating deep-learning models though the major bottleneck lies with the limitation of adequate training samples. The proposed model for S-Palmitoylation site prediction can be extended by developing a web server and can further be extended for other PTM types.

References

[1] T. Pawson and P. Nash, “Assembly of cell regulatory systems through protein interaction domains,” *science*, vol. 300, no. 5618, pp. 445–452, 2003.

[2] J. Liu, G. K. Sukhova, J.-S. Sun, W.-H. Xu, P. Libby, and G.-P. Shi, “Lysosomal cysteine proteases in atherosclerosis,” *Arteriosclerosis, thrombosis, and vascular biology*, vol. 24, no. 8, pp. 1359–1366, 2004.

[3] M. Skwarczynska and C. Ottmann, “Protein–protein interactions as drug targets,” *Future medicinal chemistry*, vol. 7, no. 16, pp. 2195–2219, 2015.

[4] J. Janin, R. P. Bahadur, and P. Chakrabarti, “Protein–protein interaction and quaternary structure,” *Quarterly reviews of biophysics*, vol. 41, no. 2, pp. 133–180, 2008.

[5] P. L. Kastritis and A. M. Bonvin, “Are scoring functions in protein- protein docking ready to predict interactomes? clues from a novel binding affinity benchmark,” *Journal of proteome research*, vol. 9, no. 5, pp. 2216–2225, 2010.

[6] S. Jones and J. M. Thornton, “Principles of protein-protein interactions.” *Proceedings of the National Academy of Sciences*, vol. 93, no. 1, pp. 13–20, 1996.

[7] R. P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin, “Dissecting subunit interfaces in homodimeric proteins,” *Proteins: Structure, Function, and Bioinformatics*, vol. 53, no. 3, pp. 708–719, 2003.

[8] N. Tuncbag, A. Gursoy, and O. Keskin, “Prediction of protein–protein interactions: unifying evolution and structure at protein interfaces,” *Physical biology*, vol. 8, no. 3, p. 035006, 2011.

[9] A.-L. Barabasi and Z. N. Oltvai, “Network biology: understanding the cell’s functional organization,” *Nature reviews genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[10] M. Caldera, P. Buphamalai, F. Müller, and J. Menche, “Interactome-based approaches to human disease,” *Current Opinion in Systems Biology*, vol. 3, pp. 88–94, 2017.

[11] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca *et al.*, "A proteome-scale map of the human interactome network," *Cell*, vol. 159, no. 5, pp. 1212–1226, 2014.

[12] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter *et al.*, "Structure-based prediction of protein–protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[13] S. Fields and O.-k. Song, "A novel genetic system to detect protein–protein interactions," *Nature*, vol. 340, no. 6230, pp. 245–246, 1989.

[14] A.-C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld *et al.*, "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–636, 2006.

[15] V. S. Rao, K. Srinivas, G. Sujini, and G. Kumar, "Protein-protein interaction detection: methods and analysis," *International journal of proteomics*, vol. 2014, 2014.

[16] N. T. Suresh and S. Ashok, "Comparative strategy for the statistical & network based analysis of biological networks," *Procedia computer science*, vol. 143, pp. 165–180, 2018.

[17] P. Durek and D. Walther, "The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles," *BMC systems biology*, vol. 2, no. 1, pp. 1–20, 2008.

[18] F. S. Pair and T. A. Yacoubian, "14-3-3 proteins: novel pharmacological targets in neurodegenerative diseases," *Trends in pharmacological sciences*, vol. 42, no. 4, pp. 226–238, 2021.

[19] J. Wang, M. Li, Y. Deng, and Y. Pan, "Recent advances in clustering methods for protein interaction networks," *BMC genomics*, vol. 11, no. 3, pp. 1–19, 2010.

[20] L. J. Ko and C. Prives, "p53: puzzle and paradigm." *Genes & development*, vol. 10, no. 9, pp. 1054–1072, 1996.

[21] P. Nordlund and P. Reichard, "Ribonucleotide reductases," *Annu. Rev. Biochem.*, vol. 75, pp. 681–706, 2006.

[22] A. C. Conibear, "Deciphering protein post-translational modifications using chemical biology tools," *Nature Reviews Chemistry*, vol. 4, no. 12, pp. 674–695, 2020.

[23] M. M. Muller, "Post-translational modifications of protein backbones: unique functions, mechanisms, and challenges," *Biochemistry*, vol. 57, no. 2, pp. 177–185, 2018.

[24] M. Tatjewski, M. Kierczak, and D. Plewczynski, "Predicting post-translational modifications from local sequence fragments using machine learning algorithms: Overview and best practices," *Prediction of Protein Secondary Structure*, pp. 275–300, 2017.

[25] M. Leutert, S. W. Entwisle, and J. Villén, "Decoding post-translational modification crosstalk with proteomics," *Molecular & Cellular Proteomics*, vol. 20, 2021.

[26] G. Duan and D. Walther, "The roles of post-translational modifications in the context of protein interaction networks," *PLoS computational biology*, vol. 11, no. 2, p. e1004049, 2015.

[27] S. Yerneni, I. K. Khan, Q. Wei, and D. Kihara, "Ias: Interaction specific go term associations for predicting protein-protein interaction networks," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 4, pp. 1247–1258, 2015.

[28] 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, 2018.

[29] R. C. Edgar, "Search and clustering orders of magnitude faster than blast," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.

[30] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[31] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "Uniref: comprehensive and non-redundant uniprot reference clusters," *Bioinformatics*, vol. 23, no. 10, pp. 1282–1288, 2007.

[32] TutorialsPoint. What is dbscan? Website (Accessed 25-04-2023 Access 2023). [Online]. Available: https://www.tutorialspoint.com/what-is-dbscan

[33] "Uniprot: the universal protein knowledgebase in 2021," *Nucleic acids research*, vol. 49, no. D1, pp. D480–D489, 2021.

[34] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "Dip: the database of interacting proteins," *Nucleic acids research*, vol. 28, no. 1, pp. 289–291, 2000.

[35] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam *et al.*, "The biogrid interaction database: 2019 update," *Nucleic acids research*, vol. 47, no. D1, pp. D529–D541, 2019.

[36] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz *et al.*, "The intact molecular interaction database in 2012," *Nucleic acids research*, vol. 40, no. D1, pp. D841–D846, 2012.

[37] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork *et al.*, "String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic acids research*, vol. 47, no. D1, pp. D607–D613, 2019.

[38] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The protein data bank: a computer-based archival file for macromolecular structures," *Journal of molecular biology*, vol. 112, no. 3, pp. 535–542, 1977.

[39] C. N. Magnan and P. Baldi, "Sspro/accpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity," *Bioinformatics*, vol. 30, no. 18, pp. 2592–2597, 2014.

[40] R. Singh, J. Xu, and B. Berger, "Pairwise global alignment of protein interaction networks by matching neighborhood topology," in *Annual international conference on research in computational molecular biology.* Springer, 2007, pp. 16–31.

[41] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, “Isorankn: spectral methods for global alignment of multiple protein networks,” *Bioinformatics*, vol. 25, no. 12, pp. i253–i258, 2009.

[42] R. A. Laskowski, J. D. Watson, and J. M. Thornton, “Profunc: a server for predicting protein function from 3d structure,” *Nucleic acids research*, vol. 33, no. suppl_2, pp. W89–W93, 2005.

[43] ——, “Protein function prediction using local 3d templates,” *Journal of molecular biology*, vol. 351, no. 3, pp. 614–626, 2005.

[44] H. Fang and J. Gough, “Dcgo: database of domain-centric ontologies on functions, phenotypes, diseases and more,” *Nucleic acids research*, vol. 41, no. D1, pp. D536–D544, 2013.

[45] T. N. Petersen, S. Brunak, G. Von Heijne, and H. Nielsen, “Signalp 4.0: discriminating signal peptides from transmembrane regions,” *Nature methods*, vol. 8, no. 10, pp. 785–786, 2011.

[46] M. Källberg, H. Wang, S. Wang, J. Peng, Z. Wang, H. Lu, and J. Xu, “Template-based protein structure modeling using the raptorx web server,” *Nature protocols*, vol. 7, no. 8, pp. 1511–1522, 2012.

[47] J. Peng and J. Xu, “Raptorx: exploiting structure information for protein alignment by statistical inference,” *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. S10, pp. 161–171, 2011.

[48] N. Furnham, G. L. Holliday, T. A. de Beer, J. O. Jacobsen, W. R. Pearson, and J. M. Thornton, “The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes,” *Nucleic acids research*, vol. 42, no. D1, pp. D485–D489, 2014.

[49] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes *et al.*, “The genemania prediction server: biological network integration for gene prioritization and predicting gene function,” *Nucleic acids research*, vol. 38, no. suppl_2, pp. W214–W220, 2010.

[50] Y. Ofran, V. Mysore, and B. Rost, “Prediction of dna-binding residues from sequence,” *Bioinformatics*, vol. 23, no. 13, pp. i347–i353, 2007.

[51] L. Wang, C. Huang, M. Q. Yang, and J. Y. Yang, "Bindn+ for accurate prediction of dna and rna-binding residues from protein sequence features," *BMC Systems Biology*, vol. 4, pp. 1–9, 2010.

[52] D. Del Vecchio, "A control theoretic framework for modular analysis and design of biomolecular networks," *Annual Reviews in Control*, vol. 37, no. 2, pp. 333–345, 2013.

[53] D. Park, R. Singh, M. Baym, C.-S. Liao, and B. Berger, "Isobase: a database of functionally related proteins across ppi networks," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D295–D300, 2010.

[54] F. Ay, M. Kellis, and T. Kahveci, "Submap: aligning metabolic pathways with subnetwork mappings," *Journal of computational biology*, vol. 18, no. 3, pp. 219–235, 2011.

[55] 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.

[56] A. K. Halder, P. Dutta, M. Kundu, S. Basu, and M. Nasipuri, "Review of computational methods for virus–host protein interaction prediction: a case study on novel ebola–human interactions," *Briefings in functional genomics*, vol. 17, no. 6, pp. 381–391, 2018.

[57] R. Dolin, J. E. Bennett, G. L. Mandell *et al.*, "Mandell, douglas, and bennett's principles and practice of infectious diseases," in *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, 2005, pp. 1727–3662.

[58] H. Zhou, J. Jin, and L. Wong, "Progress in computational studies of host–pathogen interactions," *Journal of bioinformatics and computational biology*, vol. 11, no. 02, p. 1230001, 2013.

[59] H. Zhou, S. Gao, N. N. Nguyen, M. Fan, J. Jin, B. Liu, L. Zhao, G. Xiong, M. Tan, S. Li *et al.*, "Stringent homology-based prediction of h. sapiens-m. tuberculosis h37rv protein-protein interactions," *Biology direct*, vol. 9, no. 1, pp. 1–30, 2014.

[60] S. M. Jones, H. Feldmann, U. Ströher, J. B. Geisbert, L. Fernando, A. Grolla, H.-D. Klenk, N. J. Sullivan, V. E. Volchkov, E. A. Fritz *et al.*, "Live attenuated recombinant vaccine protects nonhuman primates against ebola and marburg viruses," *Nature medicine*, vol. 11, no. 7, pp. 786–790, 2005.

[61] A. Tuffs, "Experimental vaccine may have saved hamburg scientist from ebola fever," 2009.

[62] D. S. Chertow, C. Kleine, J. K. Edwards, R. Scaini, R. Giuliani, and A. Sprecher, "Ebola virus disease in west africa—clinical manifestations and management," *New England Journal of Medicine*, vol. 371, no. 22, pp. 2054–2057, 2014.

[63] D. Knipe, P. Howley, D. Griffin, R. Lamb, M. Martin, B. Roizman, and S. Straus, *Fields Virology, Volumes 1 and 2.* Lippincott Williams & Wilkins: Philadelphia, PA, USA, 2013.

[64] Y. Huang, L. Xu, Y. Sun, and G. J. Nabel, "The assembly of ebola virus nucleocapsid requires virion-associated proteins 35 and 24 and posttranslational modification of nucleoprotein," *Molecular cell*, vol. 10, no. 2, pp. 307–316, 2002.

[65] M. D. Dyer, T. Murali, and B. W. Sobral, "Supervised learning and prediction of physical interactions between human and hiv proteins," *Infection, Genetics and Evolution*, vol. 11, no. 5, pp. 917–923, 2011.

[66] P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, and D. Plewczynski, "Ppi_svm: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables," *Cellular and Molecular Biology Letters*, vol. 16, no. 2, pp. 264–278, 2011.

[67] O. Krishnadev and N. Srinivasan, "Prediction of protein–protein interactions between human host and a pathogen and its application to three pathogenic bacteria," *International journal of biological macromolecules*, vol. 48, no. 4, pp. 613–619, 2011.

[68] Z. Itzhaki, "Domain-domain interactions underlying herpesvirus-human protein-protein interaction networks," *PloS one*, vol. 6, no. 7, p. e21724, 2011.

[69] M. Tyagi, K. Hashimoto, B. A. Shoemaker, S. Wuchty, and A. R. Panchenko, "Large-scale mapping of human protein interactome using structural complexes," *EMBO reports*, vol. 13, no. 3, pp. 266–271, 2012.

[70] E. A. Franzosa and Y. Xia, "Structural principles within the human-virus protein-protein interaction network," *Proceedings of the National Academy of Sciences*, vol. 108, no. 26, pp. 10 538–10 543, 2011.

[71] L.-L. Zheng, C. Li, J. Ping, Y. Zhou, Y. Li, P. Hao *et al.*, "The domain landscape of virus-host interactomes," *BioMed research international*, vol. 2014, 2014.

[72] A. Becerra, V. A. Bucheli, and P. A. Moreno, “Prediction of virus-host protein-protein interactions mediated by short linear motifs,” *BMC bioinformatics*, vol. 18, pp. 1–11, 2017.

[73] A. Segura-Cabrera, C. A. García-Pérez, X. Guo, and M. A. Rodríguez-Pérez, “A viral-human interactome based on structural motif-domain interactions captures the human infectome,” *PloS one*, vol. 8, no. 8, p. e71526, 2013.

[74] N. Kharrat, S. Belmabrouk, R. Abdelhedi, R. Benmarzoug, M. Assidi, M. H. Al Qahtani, and A. Rebai, “Screening for clusters of charge in human virus proteomes,” *BMC genomics*, vol. 17, pp. 11–19, 2016.

[75] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, “A novel biclustering approach to association rule mining for predicting hiv-1–human protein interactions,” *PLoS One*, vol. 7, no. 4, p. e32289, 2012.

[76] A. Mukhopadhyay, S. Ray, and U. Maulik, “Incorporating the type and direction information in predicting novel regulatory interactions between hiv-1 and human proteins using a biclustering approach,” *BMC bioinformatics*, vol. 15, pp. 1–22, 2014.

[77] G. Abdel-Azim, “New hierarchical clustering algorithm for protein sequences based on hellinger distance,” *Appl Math*, vol. 10, no. 4, pp. 1541–9, 2016.

[78] K.-A. Spencer, M. Dee, P. Britton, and J. A. Hiscox, “Role of phosphorylation clusters in the biology of the coronavirus infectious bronchitis virus nucleocapsid protein,” *Virology*, vol. 370, no. 2, pp. 373–381, 2008.

[79] S. Mei and H. Zhu, “Adaboost based multi-instance transfer learning for predicting proteome-wide interactions between salmonella and human proteins,” *PloS one*, vol. 9, no. 10, p. e110488, 2014.

[80] R. Mariano and S. Wuchty, “Structure-based prediction of host–pathogen protein interactions,” *Current opinion in structural biology*, vol. 44, pp. 119–124, 2017.

[81] G. Cui, C. Fang, and K. Han, “Prediction of protein-protein interactions between viruses and human by an svm model,” in *BMC bioinformatics*, vol. 13, no. 7. BioMed Central, 2012, pp. 1–10.

[82] B. Kim, S. Alguwaizani, X. Zhou, D.-S. Huang, B. Park, and K. Han, “An improved method for predicting interactions between virus and human proteins,”

Journal of bioinformatics and computational biology, vol. 15, no. 01, p. 1650024, 2017.

[83] S. Mei, E. K. Flemington, and K. Zhang, "A computational framework for distinguishing direct versus indirect interactions in human functional protein–protein interaction networks," *Integrative Biology*, vol. 9, no. 7, pp. 595–606, 2017.

[84] U. Ogmen, O. Keskin, A. S. Aytuna, R. Nussinov, and A. Gursoy, "Prism: protein interactions by structural matching," *Nucleic acids research*, vol. 33, no. suppl_2, pp. W331–W336, 2005.

[85] C. Winter, A. Henschel, W. K. Kim, and M. Schroeder, "Scoppi: a structural classification of protein–protein interfaces," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D310–D314, 2006.

[86] J. Teyra, M. Paszkowski-Rogacz, G. Anders, and M. T. Pisabarro, "Scowlp classification: structural comparison and analysis of protein binding regions," *BMC bioinformatics*, vol. 9, no. 1, pp. 1–11, 2008.

[87] O. Keskin, R. Nussinov, and A. Gursoy, "Prism: protein-protein interaction prediction by structural matching," *Functional Proteomics: Methods and Protocols*, pp. 505–521, 2008.

[88] K. Bohn-Wippert, E. N. Tevonian, M. R. Megaridis, and R. D. Dar, "Similarity in viral and host promoters couples viral reactivation with host cell migration," *Nature communications*, vol. 8, no. 1, p. 15006, 2017.

[89] G. Yu and Q.-Y. He, "Functional similarity analysis of human virus-encoded mirnas," *Journal of clinical bioinformatics*, vol. 1, pp. 1–7, 2011.

[90] S.-B. Zhang and Q.-R. Tang, "Protein–protein interaction inference based on semantic similarity of gene ontology terms," *Journal of theoretical biology*, vol. 401, pp. 30–37, 2016.

[91] N. Ikram, M. A. Qadir, and M. T. Afzal, "Investigating correlation between protein sequence similarity and semantic similarity using gene ontology annotations," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 3, pp. 905–912, 2017.

[92] M. Audagnotto and M. Dal Peraro, "Protein post-translational modifications: In silico prediction tools and molecular modeling," *Computational and structural biotechnology journal*, vol. 15, pp. 307–319, 2017.

[93] C. Walsh, *Posttranslational modification of proteins: expanding nature's inventory.* Roberts and Company Publishers, 2006.

[94] C. T. Walsh, S. Garneau-Tsodikova, and G. J. Gatto Jr, "Protein posttranslational modifications: the chemistry of proteome diversifications," *Angewandte Chemie International Edition*, vol. 44, no. 45, pp. 7342–7372, 2005.

[95] M. Fairbank, K. Huang, A. El-Husseini, and I. R. Nabi, "Ring finger palmitoylation of the endoplasmic reticulum gp78 e3 ubiquitin ligase," *FEBS letters*, vol. 586, no. 16, pp. 2488–2493, 2012.

[96] O. Rocks, A. Peyker, M. Kahms, P. J. Verveer, C. Koerner, M. Lumbierres, J. Kuhlmann, H. Waldmann, A. Wittinghofer, and P. I. Bastiaens, "An acylation cycle regulates localization and activity of palmitoylated ras isoforms," *Science*, vol. 307, no. 5716, pp. 1746–1752, 2005.

[97] A. Maeda, K. Okano, P. S.-H. Park, J. Lem, R. K. Crouch, T. Maeda, and K. Palczewski, "Palmitoylation stabilizes unliganded rod opsin," *Proceedings of the National Academy of Sciences*, vol. 107, no. 18, pp. 8428–8433, 2010.

[98] T. Hunter, "Tyrosine phosphorylation: thirty years and counting," *Current opinion in cell biology*, vol. 21, no. 2, pp. 140–146, 2009.

[99] A. S. Venne, F. A. Solari, F. Faden, T. Paretti, N. Dissmeyer, and R. P. Zahedi, "An improved workflow for quantitative n-terminal charge-based fractional diagonal chromatography (chafradic) to study proteolytic events in arabidopsis thaliana," *Proteomics*, vol. 15, no. 14, pp. 2458–2469, 2015.

[100] R. T. Premont, J. Inglese, and R. J. Lefkowitz, "Protein kinases that phosphorylate activated g protein-coupled receptors," *The FASEB Journal*, vol. 9, no. 2, pp. 175–182, 1995.

[101] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[102] Y. Xu, X. Wang, Y. Wang, Y. Tian, X. Shao, L.-Y. Wu, and N. Deng, "Prediction of posttranslational modification sites from amino acid sequences with kernel methods," *Journal of theoretical biology*, vol. 344, pp. 78–87, 2014.

[103] C. Kraft, F. Herzog, C. Gieffers, K. Mechtler, A. Hagting, J. Pines, and J.-M. Peters, "Mitotic regulation of the human anaphase-promoting complex by phosphorylation," *The EMBO journal*, vol. 22, no. 24, pp. 6598–6609, 2003.

[104] L. Rychlewski, M. Kschischo, L. Dong, M. Schutkowski, and U. Reimer, “Target specificity analysis of the abl kinase using peptide microarray data,” *Journal of molecular biology*, vol. 336, no. 2, pp. 307–311, 2004.

[105] Z. A. Knight, B. Schilling, R. H. Row, D. M. Kenski, B. W. Gibson, and K. M. Shokat, “Phosphospecific proteolysis for mapping sites of protein phosphorylation,” *Nature biotechnology*, vol. 21, no. 9, pp. 1047–1054, 2003.

[106] K.-C. Chou and Y.-D. Cai, “Using functional domain composition and support vector machines for prediction of protein subcellular location,” *Journal of biological chemistry*, vol. 277, no. 48, pp. 45 765–45 769, 2002.

[107] Y.-D. Cai, G.-P. Zhou, and K.-C. Chou, “Support vector machines for predicting membrane protein types by using functional domain composition,” *Biophysical Journal*, vol. 84, no. 5, pp. 3257–3263, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006349503700502

[108] K.-C. Chou and H.-B. Shen, “Memtype-2l: A web server for predicting membrane proteins and their types by incorporating evolution information through pse-pssm,” *Biochemical and Biophysical Research Communications*, vol. 360, no. 2, pp. 339–345, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006291X07012521

[109] Y.-D. Cai and K.-C. Chou, “Using functional domain composition to predict enzyme family classes,” *Journal of proteome research*, vol. 4, no. 1, pp. 109–111, 2005.

[110] K.-C. Chou, “Structural bioinformatics and its impact to biomedical science,” *Current medicinal chemistry*, vol. 11, no. 16, pp. 2105–2134, 2004.

[111] Y. Xu, J. Ding, L.-Y. Wu, and K.-C. Chou, “isno-pseaac: predict cysteine s-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition,” *PloS one*, vol. 8, no. 2, p. e55844, 2013.

[112] J. H. Kim, J. Lee, B. Oh, K. Kimm, and I. Koh, “Prediction of phosphorylation sites using svms,” *Bioinformatics*, vol. 20, no. 17, pp. 3179–3184, 2004.

[113] Y.-H. Wong, T.-Y. Lee, H.-K. Liang, C.-M. Huang, T.-Y. Wang, Y.-H. Yang, C.-H. Chu, H.-D. Huang, M.-T. Ko, and J.-K. Hwang, “Kinasephos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences

and coupling patterns," *Nucleic acids research*, vol. 35, no. suppl_2, pp. W588–W594, 2007.

[114] J. Gao, J. J. Thelen, A. K. Dunker, and D. Xu, "Musite, a tool for global prediction of general and kinase-specific phosphorylation sites," *Molecular & Cellular Proteomics*, vol. 9, no. 12, pp. 2586–2600, 2010.

[115] W.-C. Chang, T.-Y. Lee, D.-M. Shien, J. B.-K. Hsu, J.-T. Horng, P.-C. Hsu, T.-Y. Wang, H.-D. Huang, and R.-L. Pan, "Incorporating support vector machine for identifying protein tyrosine sulfation sites," *Journal of computational chemistry*, vol. 30, no. 15, pp. 2526–2537, 2009.

[116] X. Zhao, W. Zhang, X. Xu, Z. Ma, and M. Yin, "Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs," 2012.

[117] P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. Martin, G. Ausiello, B. Brannetti, A. Costantini *et al.*, "Elm server: a new resource for investigating short functional sites in modular eukaryotic proteins," *Nucleic acids research*, vol. 31, no. 13, pp. 3625–3630, 2003.

[118] C. J. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher, "Prosite: a documented database using patterns and profiles as motif descriptors," *Briefings in bioinformatics*, vol. 3, no. 3, pp. 265–274, 2002.

[119] M. B. Yaffe, G. G. Leparc, J. Lai, T. Obata, S. Volinia, and L. C. Cantley, "A motif-based profile scanning approach for genome-wide prediction of signaling pathways," *Nature biotechnology*, vol. 19, no. 4, pp. 348–353, 2001.

[120] D. Plewczynski, A. Tkacz, L. S. Wyrwicz, and L. Rychlewski, "Automotif server: prediction of single residue post-translational modifications in proteins," *Bioinformatics*, vol. 21, no. 10, pp. 2525–2527, 2005.

[121] S. Basu and D. Plewczynski, "Ams 3.0: prediction of post-translational modifications," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–15, 2010.

[122] D. Plewczynski, S. Basu, and I. Saha, "Ams 4.0: consensus prediction of post-translational modifications in protein sequences," *Amino Acids*, vol. 43, pp. 573–582, 2012.

[123] Y. Cai and Y. Sun, "Esprit-tree: hierarchical clustering analysis of millions of 16s rrna pyrosequences in quasilinear computational time," *Nucleic acids research*, vol. 39, no. 14, pp. e95–e95, 2011.

[124] M. Vidal, M. E. Cusick, and A.-L. Barabási, “Interactome networks and human disease,” *Cell*, vol. 144, no. 6, pp. 986–998, 2011.

[125] K. Karagoz, T. Sevimoglu, and K. Y. Arga, “Integration of multiple biological features yields high confidence human protein interactome,” *Journal of theoretical biology*, vol. 403, pp. 85–96, 2016.

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

[127] 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.

[128] P. Dutta, S. Basu, and M. Kundu, “Assessment of semantic similarity between proteins using information content and topological properties of the gene ontology graph,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 3, pp. 839–849, 2017.

[129] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, “Global protein function prediction from protein-protein interaction networks,” *Nature biotechnology*, vol. 21, no. 6, pp. 697–700, 2003.

[130] Y.-R. Cho, Y. Xin, and G. Speegle, “P-finder: Reconstruction of signaling networks from protein-protein interactions and go annotations,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 12, no. 2, pp. 309–321, 2014.

[131] G. Bebek and J. Yang, “Pathfinder: mining signal transduction pathway segments from protein-protein interaction networks,” *BMC bioinformatics*, vol. 8, no. 1, p. 335, 2007.

[132] A. K. Halder, M. Denkiewicz, K. Sengupta, S. Basu, and D. Plewczynski, “Aggregated network centrality shows non-random structure of genomic and proteomic networks,” *Methods*, 2019.

[133] D. Dasagrandhi, A. S. K. Ravindran, A. Muthuswamy, and K. Jayachandran, “Construction and analysis of protein-protein interaction network: Role in identification of key signaling molecules involved in a disease pathway,” in *Computer Applications in Drug Discovery and Development.* IGI Global, 2019, pp. 204–220.

[134] J. H. Moon, S. Lim, K. Jo, S. Lee, S. Seo, and S. Kim, "Pintnet: construction of condition-specific pathway interaction network by computing shortest paths on weighted ppi," *BMC systems biology*, vol. 11, no. 2, p. 15, 2017.

[135] M. D. Dyer, T. Murali, and B. W. Sobral, "Computational prediction of host-pathogen protein–protein interactions," *Bioinformatics*, vol. 23, no. 13, pp. i159–i166, 2007.

[136] P. Dutta, A. K. Halder, S. Basu, and M. Kundu, "A survey on ebola genome and current trends in computational research on the ebola virus," *Briefings in Functional Genomics*, vol. 17, no. 6, pp. 374–380, 2018.

[137] S. Saha, K. Sengupta, P. Chatterjee, S. Basu, and M. Nasipuri, "Analysis of protein targets in pathogen–host interaction in infectious diseases: a case study on plasmodium falciparum and homo sapiens interaction network," *Briefings in functional genomics*, vol. 17, no. 6, pp. 441–450, 2018.

[138] 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*, 2020.

[139] S. Kassani, P. Kassasni, M. Wesolowski, K. Schneider, and R. Deters, "Automatic detection of coronavirus disease (covid-19) in x-ray and ct images: a machine learning-based approach.(2020)," *arXiv preprint arXiv:2004.10641*, 2004.

[140] E. De Wit, N. Van Doremalen, D. Falzarano, and V. J. Munster, "Sars and mers: recent insights into emerging coronaviruses," *Nature reviews microbiology*, vol. 14, no. 8, pp. 523–534, 2016.

[141] A. Goncearenco, M. Li, F. L. Simonetti, B. A. Shoemaker, and A. R. Panchenko, "Exploring protein-protein interactions as drug targets for anti-cancer therapy with in silico workflows," *Proteomics for Drug Discovery: Methods and Protocols*, pp. 221–236, 2017.

[142] P. Chène, "Drugs targeting protein–protein interactions," *ChemMedChem: Chemistry Enabling Drug Discovery*, vol. 1, no. 4, pp. 400–411, 2006.

[143] S. Yang, C. Fu, X. Lian, X. Dong, and Z. Zhang, "Understanding human-virus protein-protein interactions using a human protein complex-based analysis framework. msystems," *Am Soc Microbiol*, vol. 4, no. 2, 2019.

[144] Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston, "Semi-supervised multi-task learning for predicting interactions between hiv-1 and human proteins," *Bioinformatics*, vol. 26, no. 18, pp. i645–i652, 2010.

[145] T. Kazmirchuk, K. Dick, D. J. Burnside, B. Barnes, H. Moteshareie, M. Hajikarimlou, K. Omidi, D. Ahmed, A. Low, C. Lettl *et al.*, "Designing anti-zika virus peptides derived from predicted human-zika virus protein-protein interactions," *Computational biology and chemistry*, vol. 71, pp. 180–187, 2017.

[146] V. A. Golubeva, T. C. Nepomuceno, G. de Gregoriis, R. D. Mesquita, X. Li, S. Dash, P. P. Garcez, G. Suarez-Kurtz, V. Izumi, J. Koomen *et al.*, "Network of interactions between zika virus non-structural proteins and human host proteins," *Cells*, vol. 9, no. 1, p. 153, 2020.

[147] M. Utomo, Y. Whulanza, F. Lestari, A. Erryani, and I. Kartika, "In aip conference proceedings," 2020.

[148] R. Nadeau, S. Shahryari Fard, A. Scheer, E. Hashimoto-Roth, D. Nygard, I. Abramchuk, Y.-E. Chung, S. A. Bennett, and M. Lavallée-Adam, "Computational identification of human biological processes and protein sequence motifs putatively targeted by sars-cov-2 proteins using protein–protein interaction networks," *Journal of proteome research*, vol. 19, no. 11, pp. 4553–4566, 2020.

[149] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, "Detection of covid-19 infection from routine blood exams with machine learning: a feasibility study," *Journal of medical systems*, vol. 44, pp. 1–12, 2020.

[150] A. Chakraborty, S. Mitra, M. Bhattacharjee, D. De, and A. J. Pal, "Determining human-coronavirus protein-protein interaction using machine intelligence," *Medicine in Novel Technology and Devices*, vol. 18, p. 100228, 2023.

[151] Y. Cai, J. Wang, and L. Deng, "Sdn2go: An integrated deep learning model for protein function prediction," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00391

[152] S. Saha, P. Chatterjee, S. Basu, M. Nasipuri, and D. Plewczynski, "Funpred 3.0: improved protein function prediction using protein interaction network," *PeerJ*, vol. 7, p. e6830, 2019.

[153] S. Saha, A. Prasad, P. Chatterjee, S. Basu, and M. Nasipuri, “Protein function prediction from dynamic protein interaction network using gene expression data,” *Journal of bioinformatics and computational biology*, vol. 17, no. 04, p. 1950025, 2019.

[154] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, “Lethality and centrality in protein networks,” *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[155] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, “High-betweenness proteins in the yeast protein interaction network,” *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, p. 96, 2005.

[156] M. Li, J. Wang, X. Chen, H. Wang, and Y. Pan, “A local average connectivity-based method for identifying essential proteins from the network level,” *Computational biology and chemistry*, vol. 35, no. 3, pp. 143–150, 2011.

[157] J. Zhong, C. Tang, W. Peng, M. Xie, Y. Sun, Q. Tang, Q. Xiao, and J. Yang, “A novel essential protein identification method based on ppi networks and gene expression data,” *BMC bioinformatics*, vol. 22, no. 1, pp. 1–21, 2021.

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

[159] G. Sabidussi, “The centrality index of a graph,” *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[160] A. Zumla, J. F. Chan, E. I. Azhar, D. S. Hui, and K.-Y. Yuen, “Coronaviruses—drug discovery and therapeutic options,” *Nature reviews Drug discovery*, vol. 15, no. 5, pp. 327–347, 2016.

[161] C. I. Paules, H. D. Marston, and A. S. Fauci, “Coronavirus infections—more than just the common cold,” *Jama*, vol. 323, no. 8, pp. 707–708, 2020.

[162] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei *et al.*, “Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study,” *The lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[163] R. K. Guy, R. S. DiPaola, F. Romanelli, and R. E. Dutch, “Rapid repurposing of drugs for covid-19,” *Science*, vol. 368, no. 6493, pp. 829–830, 2020. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.abb9332

[164] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney *et al.*, "A sars-cov-2 protein interaction map reveals targets for drug repurposing," *Nature*, vol. 583, no. 7816, pp. 459–468, 2020.

[165] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang *et al.*, "Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor," *nature*, vol. 581, no. 7807, pp. 215–220, 2020.

[166] J. Shang, G. Ye, K. Shi, Y. Wan, C. Luo, H. Aihara, Q. Geng, A. Auerbach, and F. Li, "Structural basis of receptor recognition by sars-cov-2," *Nature*, vol. 581, no. 7807, pp. 221–224, 2020.

[167] B. J. McConkey, V. Sobolev, and M. Edelman, "The performance of current methods in ligand–protein docking," *Current Science*, pp. 845–856, 2002.

[168] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, "A geometric approach to macromolecule-ligand interactions," *Journal of molecular biology*, vol. 161, no. 2, pp. 269–288, 1982.

[169] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, "Docking and scoring in virtual screening for drug discovery: methods and applications," *Nature reviews Drug discovery*, vol. 3, no. 11, pp. 935–949, 2004.

[170] M. Lazniewski, D. Dermawan, S. Hidayat, M. Muchtaridi, W. K. Dawson, and D. Plewczynski, "Drug repurposing for identification of potential spike inhibitors for sars-cov-2 using molecular docking and molecular dynamics simulations," *Methods*, vol. 203, pp. 498–510, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1046202322000329

[171] S. E. Omer, T. M. Ibrahim, O. A. Krar, A. M. Ali, A. A. Makki, W. Ibraheem, and A. A. Alzain, "Drug repurposing for sars-cov-2 main protease: Molecular docking and molecular dynamics investigations," *Biochemistry and biophysics reports*, vol. 29, p. 101225, 2022.

[172] F. Ahmed, J. W. Lee, A. Samantasinghar, Y. S. Kim, K. H. Kim, I. S. Kang, F. H. Memon, J. H. Lim, and K. H. Choi, "Speropredictor: An integrated machine learning and molecular docking-based drug repurposing framework with use case of covid-19," *Frontiers in public health*, vol. 10, p. 902123, 2022.

[173] V. Chandel, S. Raj, B. Rathi, and D. Kumar, “In silico identification of potent covid-19 main protease inhibitors from fda approved antiviral compounds and active phytochemicals through molecular docking: a drug repurposing approach,” 2020.

[174] Y. Kumar, H. Singh, and C. N. Patel, “In silico prediction of potential inhibitors for the main protease of sars-cov-2 using molecular docking and dynamics simulation based drug-repurposing,” *Journal of Infection and Public Health*, vol. 13, no. 9, pp. 1210–1223, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876034120305268

[175] M. Lai, “Coronaviridae,” *Fields virology*, pp. 1305–1318, 2007.

[176] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu *et al.*, “Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding,” *The lancet*, vol. 395, no. 10224, pp. 565–574, 2020.

[177] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang *et al.*, “A pneumonia outbreak associated with a new coronavirus of probable bat origin,” *nature*, vol. 579, no. 7798, pp. 270–273, 2020.

[178] M. Letko, A. Marzi, and V. Munster, “Functional assessment of cell entry and receptor usage for sars-cov-2 and other lineage b betacoronaviruses,” *Nature microbiology*, vol. 5, no. 4, pp. 562–569, 2020.

[179] Y. Cai, J. Wang, and L. Deng, “Sdn2go: an integrated deep learning model for protein function prediction. front bioeng biotechnol. 2020; 8: 391,” 2020.

[180] S. Saha, A. Prasad, P. Chatterjee, S. Basu, and M. Nasipuri, “Protein function prediction from protein–protein interaction network using gene ontology based neighborhood analysis and physico-chemical features,” *Journal of Bioinformatics and Computational Biology*, vol. 16, no. 06, p. 1850025, 2018.

[181] J. M. Anthonisse, “The rush in a directed graph,” *Stichting Mathematisch Centrum. Mathematische Besliskunde*, no. BN 9/71, 1971.

[182] Y. Guan, B. Zheng, Y. He, X. Liu, Z. Zhuang, C. Cheung, S. Luo, P. H. Li, L. Zhang, Y. Guan *et al.*, “Isolation and characterization of viruses related to the sars coronavirus from animals in southern china,” *Science*, vol. 302, no. 5643, pp. 276–278, 2003.

[183] H.-D. Song, C.-C. Tu, G.-W. Zhang, S.-Y. Wang, K. Zheng, L.-C. Lei, Q.-X. Chen, Y.-W. Gao, H.-Q. Zhou, H. Xiang *et al.*, "Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human," *Proceedings of the National Academy of Sciences*, vol. 102, no. 7, pp. 2430–2435, 2005.

[184] W. Li, C. Zhang, J. Sui, J. H. Kuhn, M. J. Moore, S. Luo, S.-K. Wong, I.-C. Huang, K. Xu, N. Vasilieva *et al.*, "Receptor and viral determinants of sars-coronavirus adaptation to human ace2," *The EMBO journal*, vol. 24, no. 8, pp. 1634–1643, 2005.

[185] Z.-y. Yang, H. C. Werner, W.-p. Kong, K. Leung, E. Traggiai, A. Lanzavecchia, and G. J. Nabel, "Evasion of antibody neutralization in emerging severe acute respiratory syndrome coronaviruses," *Proceedings of the National Academy of Sciences*, vol. 102, no. 3, pp. 797–801, 2005.

[186] G. Lu and D. Liu, "Sars-like virus in the middle east: a truly bat-related coronavirus causing human diseases," *Protein & cell*, vol. 3, no. 11, p. 803, 2012.

[187] S. K. Lau, K. S. Li, A. K. Tsang, C. S. Lam, S. Ahmed, H. Chen, K.-H. Chan, P. C. Woo, and K.-Y. Yuen, "Genetic characterization of betacoronavirus lineage c viruses in bats reveals marked sequence divergence in the spike protein of pipistrellus bat coronavirus hku5 in japanese pipistrelle: implications for the origin of the novel middle east respiratory syndrome coronavirus," *Journal of virology*, vol. 87, no. 15, pp. 8638–8650, 2013.

[188] P. C. Woo, S. K. Lau, K. S. Li, R. W. Poon, B. H. Wong, H.-w. Tsoi, B. C. Yip, Y. Huang, K.-h. Chan, and K.-y. Yuen, "Molecular diversity of coronaviruses in bats," *Virology*, vol. 351, no. 1, pp. 180–187, 2006.

[189] A. S. Cockrell, B. L. Yount, T. Scobey, K. Jensen, M. Douglas, A. Beall, X.-C. Tang, W. A. Marasco, M. T. Heise, and R. S. Baric, "A mouse model for mers coronavirus-induced acute respiratory distress syndrome," *Nature microbiology*, vol. 2, no. 2, pp. 1–11, 2016.

[190] J. F.-W. Chan, Y. Yao, M.-L. Yeung, W. Deng, L. Bao, L. Jia, F. Li, C. Xiao, H. Gao, P. Yu *et al.*, "Treatment with lopinavir/ritonavir or interferon-β1b improves outcome of mers-cov infection in a nonhuman primate model of common marmoset," *The Journal of infectious diseases*, vol. 212, no. 12, pp. 1904–1913, 2015.

[191] E. De Wit, A. L. Rasmussen, D. Falzarano, T. Bushmaker, F. Feldmann, D. L. Brining, E. R. Fischer, C. Martellaro, A. Okumura, J. Chang *et al.*, "Middle east respiratory syndrome coronavirus (mers-cov) causes transient lower respiratory tract infection in rhesus macaques," *Proceedings of the National Academy of Sciences*, vol. 110, no. 41, pp. 16 598–16 603, 2013.

[192] D. Falzarano, E. de Wit, F. Feldmann, A. L. Rasmussen, A. Okumura, X. Peng, M. J. Thomas, N. van Doremalen, E. Haddock, L. Nagy *et al.*, "Infection with mers-cov causes lethal pneumonia in the common marmoset," *PLoS pathogens*, vol. 10, no. 8, p. e1004250, 2014.

[193] M. Yeste-Velasco, M. E. Linder, and Y.-J. Lu, "Protein s-palmitoylation and cancer," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1856, no. 1, pp. 107–120, 2015.

[194] A. Buszka, A. Pytyś, D. Colvin, J. Włodarczyk, and T. Wójtowicz, "S-palmitoylation of synaptic proteins in neuronal plasticity in normal and pathological brains," *Cells*, vol. 12, no. 3, p. 387, 2023.

[195] G. M. Thomas and T. Hayashi, "Smarter neuronal signaling complexes from existing components: how regulatory modifications were acquired during animal evolution: evolution of palmitoylation-dependent regulation of ampa-type ionotropic glutamate receptors," *Bioessays*, vol. 35, no. 11, pp. 929–939, 2013.

[196] A. K. Beery and I. Zucker, "Sex bias in neuroscience and biomedical research," *Neuroscience Biobehavioral Reviews*, vol. 35, no. 3, pp. 565–572, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763410001156

[197] N. C. Tronson, "Focus on females: a less biased approach for studying strategies and mechanisms of memory," *Current Opinion in Behavioral Sciences*, vol. 23, pp. 92–97, 2018, sex and Gender. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352154617302000

[198] K. Mizuno, A. Antunes-Martins, L. Ris, M. Peters, E. Godaux, and K. Giese, "Calcium/calmodulin kinase kinase β has a male-specific role in memory formation," *Neuroscience*, vol. 145, no. 2, pp. 393–402, 2007.

[199] E. M. Waters, L. I. Thompson, P. Patel, A. D. Gonzales, H. Z. Ye, E. J. Filardo, D. J. Clegg, J. Gorecka, K. T. Akama, B. S. McEwen *et al.*, "G-protein-coupled estrogen receptor 1 is anatomically positioned to modulate synaptic plasticity in

the mouse hippocampus," *Journal of Neuroscience*, vol. 35, no. 6, pp. 2384–2397, 2015.

[200] J. Nunez and M. M. McCarthy, "Resting intracellular calcium concentration, depolarizing gamma-aminobutyric acid and possible role of local estradiol synthesis in the developing male and female hippocampus," *Neuroscience*, vol. 158, no. 2, pp. 623–634, 2009.

[201] P. Monfort, B. Gomez-Gimenez, M. Llansola, and V. Felipo, "Gender differences in spatial learning, synaptic activity, and long-term potentiation in the hippocampus in rats: molecular mechanisms," *ACS chemical neuroscience*, vol. 6, no. 8, pp. 1420–1427, 2015.

[202] W. Wang, A. A. Le, B. Hou, J. C. Lauterborn, C. D. Cox, E. R. Levin, G. Lynch, and C. M. Gall, "Memory-related synaptic plasticity is sexually dimorphic in rodent hippocampus," *Journal of Neuroscience*, vol. 38, no. 37, pp. 7935–7951, 2018.

[203] J. M. Andreano and L. Cahill, "Sex influences on the neurobiology of learning and memory," *Learning & memory*, vol. 16, no. 4, pp. 248–266, 2009.

[204] S. Hamann, "Sex differences in the responses of the human amygdala," *The Neuroscientist*, vol. 11, no. 4, pp. 288–293, 2005.

[205] M. M. Wickens, D. A. Bangasser, and L. A. Briand, "Sex differences in psychiatric disease: a focus on the glutamate system," *Frontiers in molecular neuroscience*, vol. 11, p. 197, 2018.

[206] M. Altemus, N. Sarvaiya, and C. N. Epperson, "Sex differences in anxiety and depression clinical perspectives," *Frontiers in neuroendocrinology*, vol. 35, no. 3, pp. 320–330, 2014.

[207] C. Ecker, D. S. Andrews, C. M. Gudbrandsen, A. F. Marquand, C. E. Ginestet, E. M. Daly, C. M. Murphy, M.-C. Lai, M. V. Lombardo, A. N. Ruigrok *et al.*, "Association between the probability of autism spectrum disorder and normative sex-related phenotypic diversity in brain structure," *Jama Psychiatry*, vol. 74, no. 4, pp. 329–338, 2017.

[208] M. Zareba-Kozioł, I. Figiel, A. Bartkowiak-Kaczmarek, and J. Włodarczyk, "Insights into protein s-palmitoylation in synaptic plasticity and neurological disorders: potential and limitations of methods for detection and analysis," *Frontiers in Molecular Neuroscience*, vol. 11, p. 175, 2018.

[209] A. Pedram, M. Razandi, R. J. Deschenes, and E. R. Levin, "Dhhc-7 and-21 are palmitoylacyltransferases for sex steroid receptors," *Molecular biology of the cell*, vol. 23, no. 1, pp. 188–199, 2012.

[210] M. Fukata, Y. Fukata, H. Adesnik, R. A. Nicoll, and D. S. Bredt, "Identification of psd-95 palmitoylating enzymes," *Neuron*, vol. 44, no. 6, pp. 987–996, 2004.

[211] J. Greaves, O. A. Gorleku, C. Salaun, and L. H. Chamberlain, "Palmitoylation of the snap25 protein family: specificity and regulation by dhhc palmitoyl transferases," *Journal of Biological Chemistry*, vol. 285, no. 32, pp. 24 629–24 638, 2010.

[212] E. Ponimaskin, G. Dityateva, M. O. Ruonala, M. Fukata, Y. Fukata, F. Kobe, F. S. Wouters, M. Delling, D. S. Bredt, M. Schachner *et al.*, "Fibroblast growth factor-regulated palmitoylation of the neural cell adhesion molecule determines neuronal morphogenesis," *Journal of Neuroscience*, vol. 28, no. 36, pp. 8897–8907, 2008.

[213] F. Zhou, Y. Xue, X. Yao, and Y. Xu, "Css-palm: palmitoylation site prediction with a clustering and scoring strategy (css)," *Bioinformatics*, vol. 22, no. 7, pp. 894–896, 2006.

[214] Y. Xue, H. Chen, C. Jin, Z. Sun, and X. Yao, "Nba-palm: prediction of palmitoylation site implemented in naive bayes algorithm," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–10, 2006.

[215] Z. R. Yang, "Predicting palmitoylation sites using a regularised bio-basis function neural network," in *Bioinformatics Research and Applications: Third International Symposium, ISBRA 2007, Atlanta, GA, USA, May 7-10, 2007. Proceedings 3*. Springer, 2007, pp. 406–417.

[216] X.-B. Wang, L.-Y. Wu, Y.-C. Wang, and N.-Y. Deng, "Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs," *Protein Engineering, Design & Selection*, vol. 22, no. 11, pp. 707–712, 2009.

[217] L.-L. Hu, S.-B. Wan, S. Niu, X.-H. Shi, H.-P. Li, Y.-D. Cai, and K.-C. Chou, "Prediction and analysis of protein palmitoylation sites," *Biochimie*, vol. 93, no. 3, pp. 489–496, 2011.

[218] S.-P. Shi, X.-Y. Sun, J.-D. Qiu, S.-B. Suo, X. Chen, S.-Y. Huang, and R.-P. Liang, "The prediction of palmitoylation site locations using a multiple feature

extraction method," *Journal of Molecular Graphics and Modelling*, vol. 40, pp. 125–130, 2013.

[219] L. Fu, H.-L. Xie, X.-R. Xu, H.-J. Yang, and X.-D. Nie, "Combining random forest with multi-amino acid features to identify protein palmitoylation sites," *Chemometrics and Intelligent Laboratory Systems*, vol. 135, pp. 208–212, 2014.

[220] B. Kumari, R. Kumar, and M. Kumar, "Palmpred: an svm based palmitoylation prediction method using sequence profile information," *PloS one*, vol. 9, no. 2, p. e89246, 2014.

[221] V. Pejaver, W.-L. Hsu, F. Xin, A. K. Dunker, V. N. Uversky, and P. Radivojac, "The structural and functional signatures of proteins that undergo multiple events of post-translational modification," *Protein Science*, vol. 23, no. 8, pp. 1077–1093, 2014.

[222] S. S. Bandyopadhyay, A. K. Halder, P. Chatterjee, M. Nasipuri, and S. Basu, "Hdk-means: Hadoop based parallel k-means clustering for big data," in *2017 IEEE Calcutta Conference (CALCON)*. IEEE, 2017, pp. 452–456.

[223] S. Sekhar Bandyopadhyay, A. Kumar Halder, P. Chatterjee, J. Sroka, M. Nasipuri, and S. Basu, "Analysis of large-scale human protein sequences using an efficient spark-based dbscan algorithm," in *Proceedings of International Conference on Frontiers in Computing and Systems: COMSYS 2020*. Springer, 2020, pp. 601–609.

[224] P. K. Saha, D. Jin, Y. Liu, G. E. Christensen, and C. Chen, "Fuzzy object skeletonization: theory, algorithms, and applications," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 8, pp. 2298–2314, 2017.

[225] S. Kumar, "Covid-19: A drug repurposing and biomarker identification by using comprehensive gene-disease associations through protein-protein interaction network analysis," 2020.

[226] 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.

[227] L. Chin, J. Cox, S. Esmail, M. Franklin, and D. Le, "Covid-19: Finding the right fit identifying potential treatments using a data-driven approach," *Drugbank White Paper*, 2020.

[228] J. Ren, L. Wen, X. Gao, C. Jin, Y. Xue, and X. Yao, "Css-palm 2.0: an updated software for palmitoylation sites prediction," *Protein Engineering, Design & Selection*, vol. 21, no. 11, pp. 639–644, 2008.

[229] D. Wang, Y. Liang, and D. Xu, "Capsule network for protein post-translational modification site prediction," *Bioinformatics*, vol. 35, no. 14, pp. 2386–2394, 2019.

[230] D. Wang, S. Zeng, C. Xu, W. Qiu, Y. Liang, T. Joshi, and D. Xu, "Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction," *Bioinformatics*, vol. 33, no. 24, pp. 3909–3916, 2017.

[231] M. Zareba-Kozioł, A. Bartkowiak-Kaczmarek, I. Figiel, A. Krzystyniak, T. Wojtowicz, M. Bijata, and J. Wlodarczyk, "Stress-induced changes in the s-palmitoylation and s-nitrosylation of synaptic proteins*[s]," *Molecular & Cellular Proteomics*, vol. 18, no. 10, pp. 1916–1938, 2019.

[232] M. Zareba-Kozioł, A. Bartkowiak-Kaczmarek, M. Roszkowska, K. Bijata, I. Figiel, A. K. Halder, P. Kamińska, F. E. Müller, S. Basu, W. Zhang *et al.*, "S-palmitoylation of synaptic proteins as a novel mechanism underlying sex-dependent differences in neuronal plasticity," *International journal of molecular sciences*, vol. 22, no. 12, p. 6253, 2021.

[233] N. Gorinski, D. Wojciechowski, D. Guseva, D. A. Galil, F. E. Mueller, A. Wirth, S. Thiemann, A. Zeug, S. Schmidt, M. Zareba-Kozioł *et al.*, "Dhhc7-mediated palmitoylation of the accessory protein barttin critically regulates the functions of clc-k chloride channels," *Journal of Biological Chemistry*, vol. 295, no. 18, pp. 5970–5983, 2020.

[234] S. Writer, "Big growth forecasted for big data," *Datanami*, January 2022. [Online]. Available: https://www.datanami.com/2022/01/11/big-growth-forecasted-for-big-data/

[235] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th symposium on mass storage systems and technologies (MSST).* Ieee, 2010, pp. 1–10.

[236] A. K. Koundinya, K. Sharma, K. Kumar, K. U. Shanbag *et al.*, "Map/reduce deisgn and implementation of apriori alogirthm for handling voluminous datasets," *arXiv preprint arXiv:1212.4692*, 2012.

[237] H. Hu, Y. Wen, T.-S. Chua, and X. Li, “Toward scalable systems for big data analytics: A technology tutorial,” *IEEE access*, vol. 2, pp. 652–687, 2014.

[238] Y. Qi and L. Jie, “Research of cloud storage security technology based on hdfs,” *Computer Engineering and Design*, vol. 34, no. 8, pp. 2700–2705, 2013.

[239] G. Luo, X. Luo, T. F. Gooch, L. Tian, and K. Qin, “A parallel dbscan algorithm based on spark,” in *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)*. IEEE, 2016, pp. 548–553.

[240] U. Consortium *et al.*, “Uniprot: the universal protein knowledgebase,” *Nucleic acids research*, vol. 46, no. 5, p. 2699, 2018.

[241] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund *et al.*, “The pfam protein families database,” *Nucleic acids research*, vol. 38, no. suppl_1, pp. D211–D222, 2010.

[242] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen *et al.*, “A human protein-protein interaction network: a resource for annotating the proteome,” *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

[243] T. Nepusz, H. Yu, and A. Paccanaro, “Detecting overlapping protein complexes in protein-protein interaction networks,” *Nature methods*, vol. 9, no. 5, p. 471, 2012.

[244] M. AY, K.-I. Goh, M. E. Cusick, A.-L. Barabasi, M. Vidal *et al.*, “Drug–target network,” *Nature biotechnology*, vol. 25, no. 10, pp. 1119–1127, 2007.

[245] H. Ruffner, A. Bauer, and T. Bouwmeester, “Human protein–protein interaction networks and the value for drug discovery,” *Drug discovery today*, vol. 12, no. 17-18, pp. 709–716, 2007.

[246] Y.-C. Chen, S. V. Rajagopala, T. Stellberger, and P. Uetz, “Exhaustive benchmarking of the yeast two-hybrid system,” *Nature methods*, vol. 7, no. 9, pp. 667–668, 2010.

[247] S. Fields, “High-throughput two-hybrid analysis: The promise and the peril,” *The FEBS journal*, vol. 272, no. 21, pp. 5391–5399, 2005.

[248] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat *et al.*, “Functional organization of the yeast proteome by systematic analysis of protein complexes,” *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[249] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier *et al.*, “Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry,” *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[250] M. Bellucci, F. Agostini, M. Masin, and G. G. Tartaglia, “Predicting protein associations with long noncoding rnas,” *Nature methods*, vol. 8, no. 6, p. 444, 2011.

[251] 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.

[252] V. Perovic, N. Sumonja, L. A. Marsh, S. Radovanovic, M. Vukicevic, S. G. Roberts, and N. Veljkovic, “Idppi: Protein-protein interaction analyses of human intrinsically disordered proteins,” *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[253] K. K. Wan, J. Park, and J. K. Suh, “Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair,” *Genome Informatics*, vol. 13, pp. 42–50, 2002.

[254] S. Bandyopadhyay and K. Mallick, “A new feature vector based on gene ontology terms for protein-protein interaction prediction,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 4, pp. 762–770, 2016.

[255] S. R. Maetschke, M. Simonsen, M. J. Davis, and M. A. Ragan, “Gene ontology-driven inference of protein–protein interactions using inducers,” *Bioinformatics*, vol. 28, no. 1, pp. 69–75, 2012.

[256] A. Ben-Hur and W. S. Noble, “Kernel methods for predicting protein–protein interactions,” *Bioinformatics*, vol. 21, no. suppl_1, pp. i38–i46, 2005.

[257] P. Blohm, G. Frishman, P. Smialowski, F. Goebels, B. Wachinger, A. Ruepp, and D. Frishman, “Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis,” *Nucleic acids research*, vol. 42, no. D1, pp. D396–D400, 2014.

[258] 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.

[259] U. Consortium, “Uniprot: a hub for protein information,” *Nucleic acids research*, vol. 43, no. D1, pp. D204–D212, 2015.

[260] C. Zhao and Z. Wang, “Gogo: an improved algorithm to measure the semantic similarity between gene ontology terms,” *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[261] G. O. Consortium, “The gene ontology resource: 20 years and still going strong,” *Nucleic acids research*, vol. 47, no. D1, pp. D330–D338, 2019.

[262] I. Kuznetsova, A. Lugmayr, S. J. Siira, O. Rackham, and A. Filipovska, “Cirgo: an alternative circular way of visualising gene ontology terms,” *BMC bioinformatics*, vol. 20, no. 1, pp. 1–7, 2019.

[263] J. Zhu, Q. Zhao, E. Katsevich, and C. Sabatti, “Exploratory gene ontology analysis with interactive visualization,” *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[264] H. Hassan and S. Shanak, “Gotrapper: a tool to navigate through branches of gene ontology hierarchy,” *BMC bioinformatics*, vol. 20, no. 1, pp. 1–6, 2019.

[265] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia, “Spark SQL: relational data processing in spark,” in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, T. K. Sellis, S. B. Davidson, and Z. G. Ives, Eds. ACM, 2015, pp. 1383–1394. [Online]. Available: https://doi.org/10.1145/2723372.2742797

[266] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, and I. Stoica, “Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing,” in *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2012, San Jose, CA, USA, April 25-27, 2012*, S. D. Gribble and D. Katabi, Eds. USENIX Association, 2012, pp. 15–28. [Online]. Available: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia

[267] V. Pekar and S. Staab, "Taxonomy learning-factoring the structure of a taxonomy into a semantic classification decision," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[268] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 1994, pp. 133–138.

[269] H. Yu, L. Gao, K. Tu, and Z. Guo, "Broadly predicting specific gene functions with expression similarity and taxonomy similarity," *Gene*, vol. 352, pp. 75–81, 2005.

[270] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLoS computational biology*, vol. 5, no. 7, 2009.

[271] M. G. Ahsaee, M. Naghibzadeh, and S. E. Y. Naeini, "Semantic similarity assessment of words using weighted wordnet," *International Journal of Machine Learning and Cybernetics*, vol. 5, no. 3, pp. 479–490, 2014.

[272] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of artificial intelligence research*, vol. 11, pp. 95–130, 1999.

[273] D. Lin *et al.*, "An information-theoretic definition of similarity." in *Icml*, vol. 98, no. 1998, 1998, pp. 296–304.

[274] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.

[275] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer, "A new measure for functional similarity of gene products based on gene ontology," *BMC bioinformatics*, vol. 7, no. 1, p. 302, 2006.

[276] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[277] X. Wu, E. Pang, K. Lin, and Z.-M. Pei, "Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge-and ic-based hybrid method," *PloS one*, vol. 8, no. 5, 2013.

[278] G. K. Mazandu and N. J. Mulder, "Information content-based gene ontology semantic similarity approaches: toward a unified framework theory," *BioMed research international*, vol. 2013, 2013.

[279] F. M. Couto and M. J. Silva, "Disjunctive shared information between ontology concepts: application to gene ontology," *Journal of biomedical semantics*, vol. 2, no. 1, p. 5, 2011.

[280] P. H. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic similarity analysis of protein data: assessment with biological features and issues," *Briefings in bioinformatics*, vol. 13, no. 5, pp. 569–585, 2012.

[281] G. K. Mazandu, E. R. Chimusa, M. Mbiyavanga, and N. J. Mulder, "A-dago-fun: an adaptable gene ontology semantic similarity-based functional analysis tool," *Bioinformatics*, vol. 32, no. 3, pp. 477–479, 2016.

[282] A. Nagar and H. Al-Mubaid, "A hybrid semantic similarity measure for gene ontology based on offspring and path length," in *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2015, pp. 1–7.

[283] A. Bairoch and R. Apweiler, "The swiss-prot protein sequence database and its supplement trembl in 2000," *Nucleic acids research*, vol. 28, no. 1, pp. 45–48, 2000.

[284] C. O'Donovan, M. J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, and R. Apweiler, "High-quality protein knowledge resource: Swiss-prot and trembl," *Briefings in bioinformatics*, vol. 3, no. 3, pp. 275–284, 2002.

[285] C. H. Wu, L.-S. L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek *et al.*, "The protein information resource," *Nucleic acids research*, vol. 31, no. 1, pp. 345–347, 2003.

[286] S. Pundir, M. Magrane, M. J. Martin, C. O'Donovan, and U. Consortium, "Searching and navigating uniprot databases," *Current protocols in bioinformatics*, vol. 50, no. 1, pp. 1–27, 2015.

[287] M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. A. Andrade-Navarro, "Hippie: Integrating protein interaction networks with experiment based quality scores," *PloS one*, vol. 7, no. 2, 2012.

[288] K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux *et al.*, “A reference map of the human binary protein interactome,” *Nature*, pp. 1–7, 2020.

[289] H.-S. Park and C.-H. Jun, “A simple and fast algorithm for k-medoids clustering,” *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3336–3341, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741740800081X

[290] J. Beel, B. Gipp, S. Langer, and C. Breitinger, “Research-paper recommender systems: a literature survey,” *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00799-015-0156-0

[291] M. Armbrust, T. Das, A. Davidson, A. Ghodsi, A. Or, J. Rosen, I. Stoica, P. Wendell, R. Xin, and M. Zaharia, “Scaling spark in the real world: Performance and usability,” *Proc. VLDB Endow.*, vol. 8, no. 12, p. 1840–1843, Aug. 2015. [Online]. Available: https://doi.org/10.14778/2824032.2824080

[292] J. Rosen and R. Xin, “Packaging - google chrome,” https://databricks.com/blog/2015/04/28/project-tungsten-bringing-spark-closer-to-bare-metal.html.

[293] Y. Ding, J. Tang, and F. Guo, “Predicting protein-protein interactions via multivariate mutual information of protein sequences,” *BMC bioinformatics*, vol. 17, no. 1, p. 398, 2016.

[294] S. Van Dongen, *A new cluster algorithm for graphs.* Citeseer, 1998.

[295] M. Uhlen, C. Zhang, S. Lee, E. Sjöstedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. Liu, F. Edfors *et al.*, “A pathology atlas of the human cancer transcriptome,” *Science*, vol. 357, no. 6352, 2017.

[296] S. Van Dongen, “A stochastic uncoupling process for graphs,” in *NATIONAL RESEARCH INSTITUTE FOR MATHEMATICS AND COMPUTER SCIENCE IN THE.* Citeseer, 2000.

[297] S. Ruan, “Likelihood of survival of coronavirus disease 2019,” *The Lancet Infectious Diseases*, vol. 20, no. 6, pp. 630–631, 2020.

[298] World Health Organization, “Middle East respiratory syndrome coronavirus (MERS-CoV),” https://www.who.int/emergencies/mers-cov/en/, 2020, [Accessed March 10, 2023].

[299] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, “A novel coronavirus outbreak of global health concern,” *The lancet*, vol. 395, no. 10223, pp. 470–473, 2020.

[300] World Health Organization, “Coronavirus disease (covid-19) outbreak,” https://www.who.int/emergencies/diseases/novel-coronavirus-2019, 2020, [Accessed February 28, 2023].

[301] Centers for Disease Control and Prevention, “World map,” https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/world-map.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Flocations-confirmed-cases.html, 2020, [Accessed February 28, 2023].

[302] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu *et al.*, “Clinical features of patients infected with 2019 novel coronavirus in wuhan, china,” *The lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[303] D. L. Heymann, “Data sharing and outbreaks: best practice exemplified,” *The Lancet*, vol. 395, no. 10223, pp. 469–470, 2020.

[304] Y. Chen, Q. Liu, and D. Guo, “Emerging coronaviruses: genome structure, replication, and pathogenesis,” *Journal of medical virology*, vol. 92, no. 4, pp. 418–423, 2020.

[305] X. He, L. Kuang, Z. Chen, Y. Tan, and L. Wang, “Method for identifying essential proteins by key features of proteins in a novel protein-domain network,” *Frontiers in Genetics*, vol. 12, p. 708162, 2021.

[306] S. N. Basak, A. K. Biswas, S. Saha, P. Chatterjee, S. Basu, and M. Nasipuri, “Target protein function prediction by identification of essential proteins in protein-protein interaction network,” in *Computational Intelligence, Communications, and Business Analytics: Second International Conference, CICBA 2018, Kalyani, India, July 27–28, 2018, Revised Selected Papers, Part II 2*. Springer, 2019, pp. 219–231.

[307] S. Saha, A. Prasad, P. Chatterjee, S. Basu, and M. Nasipuri, “Modified fpred-apriori: improving function prediction of target proteins from essential neighbours by finding their association with relevant functional groups using apriori algorithm,” *International Journal of Advanced Intelligence Paradigms*, vol. 19, no. 1, pp. 61–83, 2021.

[308] E. Yeger-Lotem and R. Sharan, “Human protein interaction networks across tissues and diseases,” *Frontiers in genetics*, vol. 6, p. 257, 2015.

[309] M. G. Kann, “Protein interactions and disease: computational approaches to uncover the etiology of diseases,” *Briefings in bioinformatics*, vol. 8, no. 5, pp. 333–346, 2007.

[310] T. Ideker and R. Sharan, “Protein networks in disease,” *Genome research*, vol. 18, no. 4, pp. 644–652, 2008.

[311] Center for Infectious Disease Research and Policy, “China releases genetic data on new coronavirus, now deadly,” https://www.cidrap.umn.edu/news-perspective/2020/01/china-releases-genetic-data-new-coronavirus-now-deadly, 2020, [Accessed March 10, 2023].

[312] J. F.-W. Chan, K.-H. Kok, Z. Zhu, H. Chu, K. K.-W. To, S. Yuan, and K.-Y. Yuen, “Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting wuhan,” *Emerging microbes & infections*, vol. 9, no. 1, pp. 221–236, 2020.

[313] S. Pfefferle, J. Schöpf, M. Kögl, C. C. Friedel, M. A. Müller, J. Carbajo-Lozoya, T. Stellberger, E. von Dall’Armi, P. Herzog, S. Kallies *et al.*, “The sars-coronavirus-host interactome: identification of cyclophilins as target for pan-coronavirus inhibitors,” *PLoS pathogens*, vol. 7, no. 10, p. e1002331, 2011.

[314] A. von Brunn, C. Teepe, J. C. Simpson, R. Pepperkok, C. C. Friedel, R. Zimmer, R. Roberts, R. Baric, and J. Haas, “Analysis of intraviral protein-protein interactions of the sars coronavirus orfeome,” *PloS one*, vol. 2, no. 5, p. e459, 2007.

[315] T. S. Fung and D. X. Liu, “Human coronavirus: host-pathogen interaction,” *Annual review of microbiology*, vol. 73, pp. 529–557, 2019.

[316] G. O. Consortium, “The gene ontology (go) database and informatics resource,” *Nucleic acids research*, vol. 32, no. suppl_1, pp. D258–D261, 2004.

[317] N. T. Bailey *et al.*, *The mathematical theory of infectious diseases and its applications.* Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

[318] M. Agrawal, M. Zitnik, and J. Leskovec, "Large-scale analysis of disease pathways in the human interactome," in *PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium.* World Scientific, 2018, pp. 111–122.

[319] Stanford Network Analysis Project (SNAP), "BioSNAP: Network datasets: Human protein-protein interaction network," https://snap.stanford.edu/biodata/datasets/10000/10000-PP-Pathways.html, 2021, [Accessed March 10, 2023].

[320] UniProt Consortium, "COVID-19 UniProtKB," https://covid-19.uniprot.org/, 2021, [Accessed March 10, 2023].

[321] C. Harrison, "Coronavirus puts drug repurposing on the fast track," *Nature biotechnology*, vol. 38, no. 4, pp. 379–381, 2020.

[322] B. Cao, Y. Wang, D. Wen, W. Liu, J. Wang, G. Fan, L. Ruan, B. Song, Y. Cai, M. Wei *et al.*, "A trial of lopinavir–ritonavir in adults hospitalized with severe covid-19," *New England journal of medicine*, 2020.

[323] P. Gautret, J. Lagier, P. Parola, V. Hoang, L. Meddeb, M. Mailhe, B. Doudier, J. Courjon, V. Giordanengo, V. Vieira *et al.*, "March 2020. hydroxychloroquine and azithromycin as a treatment of covid-19: results of an open-label non-randomized clinical trial," *Int J Antimicrob Agents https://doi. org/10.1016/j. ijantimicag*, 2020.

[324] E. De Wit, F. Feldmann, J. Cronin, R. Jordan, A. Okumura, T. Thomas, D. Scott, T. Cihlar, and H. Feldmann, "Prophylactic and therapeutic remdesivir (gs-5734) treatment in the rhesus macaque model of mers-cov infection," *Proceedings of the National Academy of Sciences*, vol. 117, no. 12, pp. 6771–6776, 2020.

[325] Gilead Sciences, "Emergency access to remdesivir outside of clinical trials," https://www.gilead.com/purpose/advancing-global-health/covid-19/emergency-access-to-remdesivir-outside-of-clinical-trials, 2021, [Accessed March 10, 2023].

[326] Gilead Sciences, Inc., "Remdesivir Clinical Trials," https://www.gilead.com/purpose/advancing-global-health/covid-19/remdesivir-clinical-trials, 2021, [Accessed March 10, 2023].

[327] United Press International, "China approves antiviral favilavir to treat coronavirus," https://www.upi.com/Health_News/2020/02/17/China-approves-antiviral-favilavir-to-treat-coronavirus/5291581953892/, 2020, [Accessed March 10, 2023].

[328] Focus Taiwan, "Taiwan synthesizes anti-viral drug favilavir for COVID-19 patients," https://focustaiwan.tw/sci-tech/202003020012, 2020, [Accessed March 10, 2023].

[329] "Efficacy and safety of darunavir and cobicistat for treatment of covid-19 - full text view - clinicaltrials.gov," https://clinicaltrials.gov/ct2/show/NCT04252274, 2021, accessed: 10-03-2023.

[330] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "Drugbank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D901–D906, 2008.

[331] S. Wang and F. Wu, "Detecting overlapping protein complexes in ppi networks based on robustness," *Proteome science*, vol. 11, pp. 1–8, 2013.

[332] N. Samadi and A. Bouyer, "Identifying influential spreaders based on edge ratio and neighborhood diversity measures in complex networks," *Computing*, vol. 101, pp. 1147–1175, 2019.

[333] World Health Organization, "Statement on the meeting of the international health regulations (2005) emergency committee regarding the outbreak of novel coronavirus 2019 (n-cov) on 23 january 2020," https://www.who.int/news-room/detail/23-01-2020-statement-on-the-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov), 2020, [Accessed March 10, 2023].

[334] Y. Tang, M. Li, J. Wang, Y. Pan, and F.-X. Wu, "Cytonca: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks," *Biosystems*, vol. 127, pp. 67–72, 2015.

[335] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 343–344.

[336] ——, “Measuring semantic similarity between gene ontology terms,” *Data & knowledge engineering*, vol. 61, no. 1, pp. 137–152, 2007.

[337] P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy,” *arXiv preprint cmp-lg/9511007*, 1995.

[338] S. Jain and G. D. Bader, “An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology,” *BMC bioinformatics*, vol. 11, no. 1, pp. 1–14, 2010.

[339] D. Botstein, J. M. Cherry, M. Ashburner, C. A. Ball, J. A. Blake, H. Butler, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, “Gene ontology: tool for the unification of biology,” *Nat genet*, vol. 25, no. 1, pp. 25–9, 2000.

[340] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

[341] C. T. Arena, “Trial shows covid-19 patients recover with gilead’s remdesivir,” https://www.clinicaltrialsarena.com/news/niaid-trial-remdesivir-covid-19-data/, 2021, (Accessed 28-02-2021).

[342] “Advanced search - drugbank,” https://www.drugbank.ca/unearth/advanced/drugs, 2021, accessed: 10-03-2023.

[343] “Drugbank,” https://www.drugbank.ca/, 2021, accessed: 10-03-2023.

[344] R. Systems, “ACE-2 is shown to be the entry receptor for SARS-CoV-2,” https://www.rndsystems.com/resources/articles/ace-2-sars-receptor-identified, 2021, accessed: 10-03-2023.

[345] A. B. Patel and A. Verma, “Covid-19 and angiotensin-converting enzyme inhibitors and angiotensin receptor blockers: what is the evidence?” *Jama*, vol. 323, no. 18, pp. 1769–1770, 2020.

[346] A. A. T. Naqvi, K. Fatima, T. Mohammad, U. Fatima, I. K. Singh, A. Singh, S. M. Atif, G. Hariprasad, G. M. Hasan, and M. I. Hassan, “Insights into sars-cov-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach,” *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1866, no. 10, p. 165878, 2020.

[347] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche *et al.*, “Sars-cov-2 cell entry

depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor," *cell*, vol. 181, no. 2, pp. 271–280, 2020.

[348] R. Yadav, J. K. Chaudhary, N. Jain, P. K. Chaudhary, S. Khanra, P. Dhamija, A. Sharma, A. Kumar, and S. Handu, "Role of structural and non-structural proteins and therapeutic targets of sars-cov-2 for covid-19," *Cells*, vol. 10, no. 4, p. 821, 2021.

[349] S. Matsuyama, N. Nagata, K. Shirato, M. Kawase, M. Takeda, and F. Taguchi, "Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease tmprss2," *Journal of virology*, vol. 84, no. 24, pp. 12 658–12 664, 2010.

[350] P. Zhou, X. Yang, X. Wang, B. Hu, L. Zhang, W. Zhang, H. Si, Y. Zhu, B. Li, C. Huang *et al.*, "Regulation of inflammatory cytokines and inhibition of t and b lymphocytes x," *Wang, XS Zheng, K. Zhao, QJ Chen, F. Deng, LL Liu, B. Yan, FX Zhan, YY Wang, GF Xiao, ZL Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature*, vol. 579, pp. 270–273, 2020.

[351] ClinicalTrials.gov, "Clinical trial of favipiravir tablets combine with chloroquine phosphate in the treatment of novel coronavirus pneumonia - full text view," https://clinicaltrials.gov/ct2/show/NCT04319900, 2021, [Accessed 10-03-2021].

[352] S. Saha, A. K. Halder, S. S. Bandyopadhyay, P. Chatterjee, M. Nasipuri, D. Bose, and S. Basu, "Drug repurposing for covid-19 using computational screening: Is fostamatinib/r406 a potential candidate?" *Methods*, vol. 203, pp. 564–574, 2022.

[353] WHO, "Modes of transmission of virus causing covid-19: implications for ipc precaution recommendations," https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution\recommendations, 2021, accessed 28-02-2021.

[354] G. Li and E. De Clercq, "Therapeutic options for the 2019 novel coronavirus (2019-ncov)," *Nature reviews Drug discovery*, vol. 19, no. 3, pp. 149–150, 2020.

[355] World Health Organization, "Coronavirus disease (COVID-19) dashboard," 2021, (Accessed 17-05-2021 Access 2021). [Online]. Available: https://covid19.who.int/

[356] M. I. I. Rabby, "Current drugs with potential for treatment of covid-19: A literature review: Drugs for the treatment process of covid-19," *Journal of pharmacy & pharmaceutical sciences*, vol. 23, pp. 58–64, 2020.

[357] Centers for Disease Control and Prevention (CDC), "Different COVID-19 vaccines," https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html, 2021, [Accessed on 10-03-2023].

[358] "Recommendation for an emergency use listing of on COVISHIELD™ submitted by SIIPL," 2021. [Online]. Available: https://extranet.who.int/pqweb/sites/default/files/documents/COVISHIELD_TAG_REPORT_EULvaccine.pdf

[359] B. Biotech, "Covaxin," https://www.bharatbiotech.com/covaxin.html, 2021, accessed 17-05-2021.

[360] ClinicalTrials.gov, "Covid-19 studies from the world health organization database," https://clinicaltrials.gov/ct2/who_table, 2021, [Accessed on 28-02-2021].

[361] M. Adhami, B. Sadeghi, A. Rezapour, A. A. Haghdoost, and H. MotieGhader, "Repurposing novel therapeutic candidate drugs for coronavirus disease-19 based on protein-protein interaction network analysis," *BMC biotechnology*, vol. 21, pp. 1–11, 2021.

[362] DrugBank, "Fostamatinib - DrugBank," https://www.drugbank.ca/drugs/DB12010, 2020, (Accessed 26-08-2020 Access 2020).

[363] T. Jacob, C. Van den Broeke, and H. W. Favoreel, "Viral serine/threonine protein kinases," *Journal of virology*, vol. 85, no. 3, pp. 1158–1173, 2011.

[364] N. Author, "Drug approval package: TAVALISSE (fostamatinib disodium hexahydrate)," https://www.accessdata.fda.gov/drugsatfda_docs/nda/2018/209299Orig1s000TOC.cfm, 2021, [Online; accessed 28-February-2021].

[365] FDA, "Fda approves fostamatinib tablets for itp," 2021. [Online]. Available: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-fostamatinib-tablets-itp

[366] K. McKeage and K. A. Lyseng-Williamson, "Fostamatinib in chronic immune thrombocytopenia: a profile of its use in the usa," *Drugs & Therapy Perspectives*, vol. 34, pp. 451–456, 2018.

[367] G. Lippi, M. Plebani, and B. M. Henry, “Thrombocytopenia is associated with severe coronavirus disease 2019 (covid-19) infections: a meta-analysis,” *Clinica chimica acta*, vol. 506, pp. 145–148, 2020.

[368] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng *et al.*, “Structure of mpro from sars-cov-2 and discovery of its inhibitors,” *Nature*, vol. 582, no. 7811, pp. 289–293, 2020.

[369] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao, and H. Yang, “Sars-cov-2 3cl protease (3clpro) apo structure (space group c21),” RCSB Protein Data Bank, 2020, [Online; accessed 17-May-2021].

[370] J. Osipiuk, S.-A. Azizi, S. Dvorkin, M. Endres, R. Jedrzejczak, K. A. Jones, S. Kang, R. S. Kathayat, Y. Kim, V. G. Lisnyak, S. L. Maki, V. Nicolaescu, C. A. Taylor, C. Tesar, N. Thanki, A. C. Thwin, V. L. Woods Jr., M. Wu, B. Zhang, Z. Zhou, G. Randall, K. Michalska, and A. Joachimiak, “The crystal structure of papain-like protease of sars-cov-2,” *RCSB Protein Data Bank*, 2020. [Online]. Available: http://www.rcsb.org/structure/6W9C

[371] Y. Gao, L. Yan, Y. Huang, F. Liu, Y. Zhao, L. Cao, T. Wang, Q. Sun, Z. Ming, L. Zhang *et al.*, “Structure of the rna-dependent rna polymerase from covid-19 virus,” *Science*, vol. 368, no. 6492, pp. 779–782, 2020.

[372] A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler, “Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein,” *Cell*, vol. 181, no. 2, pp. 281–292, 2020.

[373] U. Food and D. Administration, “Emergency use authorization,” https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization, 2021, accessed on April 3, 2023.

[374] M. S. Xydakis, P. Dehgani-Mobaraki, E. H. Holbrook, U. W. Geisthoff, C. Bauer, C. Hautefort, P. Herman, G. T. Manley, D. M. Lyon, and C. Hopkins, “Smell and taste dysfunction in patients with covid-19,” *The Lancet Infectious Diseases*, vol. 20, no. 9, pp. 1015–1016, 2020.

[375] C. Menni, C. H. Sudre, C. J. Steves, S. Ourselin, and T. D. Spector, "Quantifying additional covid-19 symptoms will save lives," *The Lancet*, vol. 395, no. 10241, pp. e107–e108, 2020.

[376] C. Menni, A. M. Valdes, M. B. Freidin, C. H. Sudre, L. H. Nguyen, D. A. Drew, S. Ganesh, T. Varsavsky, M. J. Cardoso, J. S. El-Sayed Moustafa *et al.*, "Real-time tracking of self-reported symptoms to predict potential covid-19," *Nature medicine*, vol. 26, no. 7, pp. 1037–1040, 2020.

[377] J. Piñero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong, "Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, 2015.

[378] C. J. Mattingly, M. C. Rosenstein, G. T. Colby, J. Forrest Jr, and J. Boyer, "The comparative toxicogenomics database (ctd): a resource for comparative toxicological studies," *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, vol. 305, no. 9, pp. 689–692, 2006.

[379] E. Ciaglia, C. Vecchione, and A. A. Puca, "Covid-19 infection and circulating ace2 levels: protective role in women and children," *Frontiers in pediatrics*, vol. 8, p. 206, 2020.

[380] H. Xiao, L. H. Xu, Y. Yamada, and D. X. Liu, "Coronavirus spike protein inhibits host cell translation by interaction with eif3f," *PLoS one*, vol. 3, no. 1, p. e1494, 2008.

[381] Y. Ma-Lauer, J. Carbajo-Lozoya, M. Y. Hein, M. A. Müller, W. Deng, J. Lei, B. Meyer, Y. Kusov, B. Von Brunn, D. R. Bairad *et al.*, "p53 down-regulates sars coronavirus replication and is targeted by the sars-unique domain and plpro via e3 ubiquitin ligase rchy1," *Proceedings of the National Academy of Sciences*, vol. 113, no. 35, pp. E5192–E5201, 2016.

[382] L. Wendt, J. Brandt, B. S. Bodmer, S. Reiche, M. L. Schmidt, S. Traeger, and T. Hoenen, "The ebola virus nucleoprotein recruits the nuclear rna export factor nxf1 into inclusion bodies to facilitate viral protein expression," *Cells*, vol. 9, no. 1, p. 187, 2020.

[383] C. Luo, H. Luo, S. Zheng, C. Gui, L. Yue, C. Yu, T. Sun, P. He, J. Chen, J. Shen *et al.*, "Nucleocapsid protein of sars coronavirus tightly binds to human

cyclophilin a," *Biochemical and biophysical research communications*, vol. 321, no. 3, pp. 557–565, 2004.

[384] P. Nigro, G. Pompilio, and M. Capogrossi, "Cyclophilin a: a key player for human disease," *Cell death & disease*, vol. 4, no. 10, pp. e888–e888, 2013.

[385] Q. Chai, V. Jovasevic, V. Malikov, Y. Sabo, S. Morham, D. Walsh, and M. H. Naghavi, "Hiv-1 counteracts an innate restriction by amyloid precursor protein resulting in neurodegeneration," *Nature communications*, vol. 8, no. 1, p. 1522, 2017.

[386] Z. Chen, A. A. Kolokoltsov, J. Wang, S. Adhikary, M. Lorinczi, L. A. Elferink, and R. A. Davey, "Grb2 interaction with the ecotropic murine leukemia virus receptor, mcat-1, controls virus entry and is stimulated by virus binding," *Journal of virology*, vol. 86, no. 3, pp. 1421–1432, 2012.

[387] T. Venkataraman and M. B. Frieman, "The role of epidermal growth factor receptor (egfr) signaling in sars coronavirus-induced pulmonary fibrosis," *Antiviral research*, vol. 143, pp. 142–150, 2017.

[388] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, "Molecular docking: a powerful approach for structure-based drug discovery," *Current computer-aided drug design*, vol. 7, no. 2, pp. 146–157, 2011.

[389] R. Thomsen and M. H. Christensen, "Moldock: a new technique for high-accuracy molecular docking," *Journal of medicinal chemistry*, vol. 49, no. 11, pp. 3315–3321, 2006.

[390] MolBioTools, "MULTIPLE LIST COMPARATOR - A free online tool to find list overlaps and draw Venn diagrams," http://molbiotools.com/listcompare.html, 2021, [Accessed: April 6, 2023].

[391] V. J. Stella, W. Charman, and V. H. Naringrekar, "Prodrugs: Do they have advantages in clinical practice?" *Drugs*, vol. 29, pp. 455–473, 1985.

[392] M. Baluom, E. B. Grossbard, T. Mant, and D. T. Lau, "Pharmacokinetics of fostamatinib, a spleen tyrosine kinase (syk) inhibitor, in healthy human subjects following single and multiple oral dosing in three phase i studies," *British journal of clinical pharmacology*, vol. 76, no. 1, pp. 78–88, 2013.

[393] W. Zhao, "Negative regulation of tbk1-mediated antiviral immunity," *FEBS letters*, vol. 587, no. 6, pp. 542–548, 2013.

[394] R. Jung, S. Radko, and P. Pelka, “The dual nature of nek9 in adenovirus replication,” *Journal of Virology*, vol. 90, no. 4, pp. 1931–1943, 2016.

[395] L. Mifflin, D. Ofengeim, and J. Yuan, “Receptor-interacting protein kinase 1 (ripk1) as a therapeutic target,” *Nature reviews Drug discovery*, vol. 19, no. 8, pp. 553–571, 2020.

[396] K. K. Singh, G. Chaubey, J. Y. Chen, and P. Suravajhala, “Decoding sars-cov-2 hijacking of host mitochondria in covid-19 pathogenesis,” *American Journal of Physiology-Cell Physiology*, 2020.

[397] The Human Protein Atlas, “CSNK2A2,” https://www.proteinatlas.org/ENSG00000070770-CSNK2A2, 2021, [Online; accessed 10-April-2023].

[398] ZFIN, “Gene: mark1,” https://zfin.org/ZDB-GENE-070626-2, 2021, [Online; accessed 10-April-2023].

[399] Open Targets Platform, “89 diseases associated with MARK3,” https://www.targetvalidation.org/target/ENSG00000075413/associations, 2021, [Online; accessed 10-April-2023].

[400] V. Malikov and M. H. Naghavi, “Localized phosphorylation of a kinesin-1 adaptor by a capsid-associated kinase regulates hiv-1 motility and uncoating,” *Cell reports*, vol. 20, no. 12, pp. 2792–2799, 2017.

[401] R. E. Turnham and J. D. Scott, “Protein kinase a catalytic subunit isoform prkaca; history, function and physiology,” *Gene*, vol. 577, no. 2, pp. 101–108, 2016.

[402] K.-i. Fujita, “Food-drug interactions via human cytochrome p450 3a (cyp3a),” *Drug metabolism and drug interactions*, vol. 20, no. 4, pp. 195–218, 2004.

[403] Rigel Pharmaceuticals, Inc., “Positive Topline Data Shows Fostamatinib Meets Primary Endpoint of Safety in Phase 2 Clinical Trial in Hospitalized Patients with COVID-19,” https://www.rigel.com/investors/news-events/press-releases/detail/312/positive-topline-data-shows-fostamatinib-meets-primary, 2021, [Online; accessed 10-April-2023].

[404] J. Guarner, “Three emerging coronaviruses in two decades: the story of sars, mers, and now covid-19,” pp. 420–421, 2020.

[405] K. Sengupta, S. Saha, A. K. Halder, P. Chatterjee, M. Nasipuri, S. Basu, and D. Plewczynski, "Pfp-go: Integrating protein sequence, domain and protein-protein interaction information for protein function prediction using ranked go terms," *Frontiers in Genetics*, vol. 13, 2022.

[406] S. Saha, P. Chatterjee, A. K. Halder, M. Nasipuri, S. Basu, and D. Plewczynski, "Ml-dtd: Machine learning-based drug target discovery for the potential treatment of covid-19," *Vaccines*, vol. 10, no. 10, p. 1643, 2022.

[407] A. Banik, S. Podder, S. Saha, P. Chatterjee, A. K. Halder, M. Nasipuri, S. Basu, and D. Plewczynski, "Rule-based pruning and in silico identification of essential proteins in yeast ppin," *Cells*, vol. 11, no. 17, p. 2648, 2022.

[408] S. Saha, P. Chatterjee, S. Basu, M. Kundu, and M. Nasipuri, "Funpred-1: protein function prediction from a protein interaction network using neighborhood analysis," *Cellular and Molecular Biology Letters*, vol. 19, no. 4, pp. 675–691, 2014.

[409] A. Prasad, S. Saha, P. Chatterjee, S. Basu, and M. Nasipuri, "Protein function prediction from protein interaction network using bottom-up l2l apriori algorithm," in *Computational Intelligence, Communications, and Business Analytics: First International Conference, CICBA 2017, Kolkata, India, March 24–25, 2017, Revised Selected Papers, Part II.* Springer, 2017, pp. 3–16.

[410] L. Schnirring, "China releases genetic data on new coronavirus, now deadly," *Center for Infectious Disease Research and Policy (CIDRAP). See https://www. cidrap. umn. edu/news-perspective/2020/01/china-releases-genetic-data-new-coronavirus-nowdeadly (accessed 15 September 2021)*, 2020.

[411] H. Xu, M. Wang, Z. Zhang, X. Zou, Y. Gao, X. Liu, E. Lu, B. Pan, S. Wu, and S. Yu, "An epidemiologic investigation on infection with severe acute respiratory syndrome coronavirus in wild animals traders in guangzhou," *Zhonghua yu fang yi xue za zhi [Chinese Journal of Preventive Medicine]*, vol. 38, no. 2, pp. 81–83, 2004.

[412] M. A. Marra, S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan *et al.*, "The genome sequence of the sars-associated coronavirus," *Science*, vol. 300, no. 5624, pp. 1399–1404, 2003.

[413] P. A. Rota, M. S. Oberste, S. S. Monroe, W. A. Nix, R. Campagnoli, J. P. Icenogle, S. Peñaranda, B. Bankamp, K. Maher, M.-h. Chen *et al.*, "Characterization of a novel coronavirus associated with severe acute respiratory syndrome," *science*, vol. 300, no. 5624, pp. 1394–1399, 2003.

[414] N. Zhong, B. Zheng, Y. Li, L. Poon, Z. Xie, K. Chan, P. Li, S. Tan, Q. Chang, J. Xie *et al.*, "Epidemiology and cause of severe acute respiratory syndrome (sars) in guangdong, people's republic of china, in february, 2003," *The Lancet*, vol. 362, no. 9393, pp. 1353–1358, 2003.

[415] Z. Shi and Z. Hu, "A review of studies on animal reservoirs of the sars coronavirus," *Virus research*, vol. 133, no. 1, pp. 74–87, 2008.

[416] I. M. Mackay and K. E. Arden, "Mers coronavirus: diagnostics, epidemiology and transmission," *Virology journal*, vol. 12, no. 1, pp. 1–21, 2015.

[417] A. M. Zaki, S. Van Boheemen, T. M. Bestebroer, A. D. Osterhaus, and R. A. Fouchier, "Isolation of a novel coronavirus from a man with pneumonia in saudi arabia," *New England Journal of Medicine*, vol. 367, no. 19, pp. 1814–1820, 2012.

[418] E. I. Azhar, D. S. Hui, Z. A. Memish, C. Drosten, and A. Zumla, "The middle east respiratory syndrome (mers)," *Infectious Disease Clinics*, vol. 33, no. 4, pp. 891–905, 2019.

[419] S. Grabherr, B. Ludewig, and N. B. Pikor, "Insights into coronavirus immunity taught by the murine coronavirus," *European Journal of Immunology*, vol. 51, no. 5, pp. 1062–1070, 2021.

[420] S. R. Weiss and S. Navas-Martin, "Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus," *Microbiology and molecular biology reviews*, vol. 69, no. 4, pp. 635–664, 2005.

[421] S. J. Bender and S. R. Weiss, "Pathogenesis of murine coronavirus in the central nervous system," *Journal of Neuroimmune Pharmacology*, vol. 5, no. 3, pp. 336–354, 2010.

[422] J. L. Leibowitz, R. Srinivasa, S. T. Williamson, M. M. Chua, M. Liu, S. Wu, H. Kang, X.-Z. Ma, J. Zhang, I. Shalev *et al.*, "Genetic determinants of mouse hepatitis virus strain 1 pneumovirulence," *Journal of virology*, vol. 84, no. 18, pp. 9278–9291, 2010.

[423] A. E. Gorbalenya, E. J. Snijder, and W. J. Spaan, “Severe acute respiratory syndrome coronavirus phylogeny: toward consensus,” *Journal of virology*, vol. 78, no. 15, pp. 7863–7866, 2004.

[424] A. N. Vlasova and L. J. Saif, “Bovine coronavirus and the associated diseases,” *Frontiers in Veterinary Science*, vol. 8, p. 643220, 2021.

[425] X. Zhang, W. Herbst, K. Kousoulas, and J. Storz, “Biological and genetic characterization of a hemagglutinating coronavirus isolated from a diarrhoeic child,” *Journal of medical virology*, vol. 44, no. 2, pp. 152–161, 1994.

[426] K. P. Alekseev, A. N. Vlasova, K. Jung, M. Hasoksuz, X. Zhang, R. Halpin, S. Wang, E. Ghedin, D. Spiro, and L. J. Saif, “Bovine-like coronaviruses isolated from four species of captive wild ruminants are homologous to bovine coronaviruses, based on complete genomic sequences,” *Journal of Virology*, vol. 82, no. 24, pp. 12 422–12 431, 2008.

[427] S. K. Lau, P. Lee, A. K. Tsang, C. C. Yip, H. Tse, R. A. Lee, L.-Y. So, Y.-L. Lau, K.-H. Chan, P. C. Woo *et al.*, “Molecular epidemiology of human coronavirus oc43 reveals evolution of different genotypes over time and recent emergence of a novel genotype due to natural recombination,” *Journal of virology*, vol. 85, no. 21, pp. 11 325–11 337, 2011.

[428] S. K. Lau, P. C. Woo, C. C. Yip, R. Y. Fan, Y. Huang, M. Wang, R. Guo, C. S. Lam, A. K. Tsang, K. K. Lai *et al.*, “Isolation and characterization of a novel betacoronavirus subgroup a coronavirus, rabbit coronavirus hku14, from domestic rabbits,” *Journal of virology*, vol. 86, no. 10, pp. 5481–5496, 2012.

[429] W. Li, Z. Shi, M. Yu, W. Ren, C. Smith, J. H. Epstein, H. Wang, G. Crameri, Z. Hu, H. Zhang *et al.*, “Bats are natural reservoirs of sars-like coronaviruses,” *Science*, vol. 310, no. 5748, pp. 676–679, 2005.

[430] S. K. Lau, P. C. Woo, K. S. Li, Y. Huang, H.-W. Tsoi, B. H. Wong, S. S. Wong, S.-Y. Leung, K.-H. Chan, and K.-Y. Yuen, “Severe acute respiratory syndrome coronavirus-like virus in chinese horseshoe bats,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 39, pp. 14 040–14 045, 2005.

[431] J. Yuan, C.-C. Hon, Y. Li, D. Wang, G. Xu, H. Zhang, P. Zhou, L. L. Poon, T. T.-Y. Lam, F. C.-C. Leung *et al.*, “Intraspecies diversity of sars-like coronaviruses in rhinolophus sinicus and its implications for the origin of sars coronaviruses in humans,” *Journal of general virology*, vol. 91, no. 4, pp. 1058–1062, 2010.

[432] W. Ren, W. Li, M. Yu, P. Hao, Y. Zhang, P. Zhou, S. Zhang, G. Zhao, Y. Zhong, S. Wang *et al.*, “Full-length genome sequences of two sars-like coronaviruses in horseshoe bats and genetic variation analysis,” *Journal of General Virology*, vol. 87, no. 11, pp. 3355–3359, 2006.

[433] P.-L. Quan, C. Firth, C. Street, J. A. Henriquez, A. Petrosov, A. Tashmukhamedova, S. K. Hutchison, M. Egholm, M. O. Osinubi, M. Niezgoda *et al.*, “Identification of a severe acute respiratory syndrome coronavirus-like virus in a leaf-nosed bat in nigeria,” *MBio*, vol. 1, no. 4, pp. e00 208–10, 2010.

[434] S. K. P. Lau, K. S. M. Li, A. K. L. Tsang, C. S. F. Lam, S. Ahmed, H. Chen, K.-H. Chan, P. C. Y. Woo, and K.-Y. Yuen, “Genetic characterization of betacoronavirus lineage c viruses in bats reveals marked sequence divergence in the spike protein of pipistrellus bat coronavirus hku5 in japanese pipistrelle: Implications for the origin of the novel middle east respiratory syndrome coronavirus,” *Journal of Virology*, vol. 87, no. 15, pp. 8638–8650, 2013. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/JVI.01055-13

[435] V. S. Raj, H. Mou, S. L. Smits, D. H. Dekkers, M. A. Müller, R. Dijkman, D. Muth, J. A. Demmers, A. Zaki, R. A. Fouchier *et al.*, “Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-emc,” *Nature*, vol. 495, no. 7440, pp. 251–254, 2013.

[436] X. Song, L. Li, P. K. Srimani, S. Y. Philip, and J. Z. Wang, “Measure the semantic similarity of go terms using aggregate information content,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 11, no. 3, pp. 468–476, 2013.

[437] S. Benabderrahmane, M. Smail-Tabbone, O. Poch, A. Napoli, and M.-D. Devignes, “Intelligo: a new vector-based semantic similarity measure including annotation origin,” *BMC bioinformatics*, vol. 11, no. 1, pp. 1–16, 2010.

[438] A. K. Halder, P. Dutta, M. Kundu, M. Nasipuri, and S. Basu, “Prediction of thyroid cancer genes using an ensemble of post translational modification, semantic and structural similarity based clustering results,” in *Pattern Recognition and Machine Intelligence: 7th International Conference, PReMI 2017, Kolkata, India, December 5-8, 2017, Proceedings 7*. Springer, 2017, pp. 418–423.

[439] M. Zitnik, R. Sosic, and J. Leskovec, “Biosnap datasets: Stanford biomedical network dataset collection,” *Note: http://snap. stanford. edu/biodata Cited by*, vol. 5, no. 1, 2018.

[440] M. Hasöksüz, S. Kilic, and F. Sarac, "Coronaviruses and sars-cov-2," *Turkish journal of medical sciences*, vol. 50, no. 9, pp. 549–556, 2020.

[441] B. Hu, H. Guo, P. Zhou, and Z.-L. Shi, "Characteristics of sars-cov-2 and covid-19," *Nature Reviews Microbiology*, vol. 19, no. 3, pp. 141–154, 2021.

[442] N. Decaro and A. Lorusso, "Novel human coronavirus (sars-cov-2): A lesson from animal coronaviruses," *Veterinary microbiology*, vol. 244, p. 108693, 2020.

[443] Y. Yan, L. Chang, and L. Wang, "Laboratory testing of sars-cov, mers-cov, and sars-cov-2 (2019-ncov): Current status, challenges, and countermeasures," *Reviews in medical virology*, vol. 30, no. 3, p. e2106, 2020.

[444] A. Balboni, M. Battilani, and S. Prosperi, "The sars-like coronaviruses: the role of bats and evolutionary relationships with sars coronavirus." *Microbiologica-Quarterly Journal of Microbiological Sciences*, vol. 35, no. 1, p. 1, 2012.

[445] M. Buonocore, C. Marino, M. Grimaldi, A. Santoro, M. Firoznezhad, O. Paciello, F. Prisco, and A. M. D'Ursi, "New putative animal reservoirs of sars-cov-2 in italian fauna: A bioinformatic approach," *Heliyon*, vol. 6, no. 11, p. e05430, 2020.

[446] P. C. Woo, S. K. Lau, K. S. Li, A. K. Tsang, and K.-Y. Yuen, "Genetic relatedness of the novel human group c betacoronavirus to tylonycteris bat coronavirus hku4 and pipistrellus bat coronavirus hku5," *Emerging microbes & infections*, vol. 1, no. 1, pp. 1–5, 2012.

[447] Q. Wang, J. Qi, Y. Yuan, Y. Xuan, P. Han, Y. Wan, W. Ji, Y. Li, Y. Wu, J. Wang *et al.*, "Bat origins of mers-cov supported by bat coronavirus hku4 usage of human receptor cd26," *Cell host & microbe*, vol. 16, no. 3, pp. 328–337, 2014.

[448] A. S. Abdel-Moneim, "Middle east respiratory syndrome coronavirus (mers-cov): evidence and speculations," *Archives of virology*, vol. 159, no. 7, pp. 1575–1584, 2014.

[449] M. Cotten, S. J. Watson, P. Kellam, A. A. Al-Rabeeah, H. Q. Makhdoom, A. Assiri, J. A. Al-Tawfiq, R. F. Alhakeem, H. Madani, F. A. AlRabiah *et al.*, "Transmission and evolution of the middle east respiratory syndrome coronavirus in saudi arabia: a descriptive genomic study," *The Lancet*, vol. 382, no. 9909, pp. 1993–2002, 2013.

[450] D. F. Kohn and C. B. Clifford, "Biology and diseases of rats," *Laboratory animal medicine*, p. 121, 2002.

[451] R. T. So, D. K. Chu, E. Miguel, R. A. Perera, J. O. Oladipo, O. Fassi-Fihri, G. Aylet, R. L. Ko, Z. Zhou, M.-S. Cheng *et al.*, "Diversity of dromedary camel coronavirus hku23 in african camels revealed multiple recombination events among closely related betacoronaviruses of the subgenus embecovirus," *Journal of virology*, vol. 93, no. 23, pp. e01 236–19, 2019.

[452] S. Kyuwa and Y. Sugiura, "Role of cytotoxic t lymphocytes and interferon-γ in coronavirus infection: Lessons from murine coronavirus infections in mice," *Journal of Veterinary Medical Science*, vol. 82, no. 10, pp. 1410–1414, 2020.

[453] P. J. Macphee, V. J. Dindzans, L.-S. Fung, and G. A. Levy, "Acute and chronic changes in the microcirculation of the liver in inbred strains of mice following infection with mouse hepatitis virus type 3," *Hepatology*, vol. 5, no. 4, pp. 649–660, 1985.

[454] R. W. Koerner, M. Majjouti, M. A. A. Alcazar, and E. Mahabir, "Of mice and men: the coronavirus mhv and mouse models as a translational approach to understand sars-cov-2," *Viruses*, vol. 12, no. 8, p. 880, 2020.

[455] M. Orzechowski, "Alpaca coronavirus sequences producing significant alignments to human betacoronavirus," Ph.D. dissertation, Figshare, 2022.

[456] P. C. Woo, Y. Huang, S. K. Lau, and K.-Y. Yuen, "Coronavirus genomics and bioinformatics analysis," *viruses*, vol. 2, no. 8, pp. 1804–1820, 2010.

[457] F. Li, "Structure, function, and evolution of coronavirus spike proteins," *Annual review of virology*, vol. 3, pp. 237–261, 2016.

[458] A. R. Fehr and S. Perlman, "Coronaviruses: an overview of their replication and pathogenesis," *Coronaviruses: methods and protocols*, pp. 1–23, 2015.

[459] A. de Mira Fernandes, P. E. Brandao, M. dos Santos Lima, M. de Souza Nunes Martins, T. G. da Silva, V. da Silva Cardoso Pinto, L. T. De Paula, M. E. S. Vicente, L. H. Okuda, and E. M. Pituco, "Genetic diversity of bcov in brazilian cattle herds," *Veterinary medicine and science*, vol. 4, no. 3, pp. 183–189, 2018.

[460] A. H. Asadi, M. Baghinezhad, and H. Asadi, "Neonatal calf diarrhea induced by rotavirus and coronavirus," *Int. J. Biosci*, vol. 6, pp. 230–236, 2015.

[461] L. J. Saif, “Bovine respiratory coronavirus,” *Veterinary Clinics: Food Animal Practice*, vol. 26, no. 2, pp. 349–364, 2010.

[462] D. Yoo, Y. Pei, N. Christie, and M. Cooper, “Primary structure of the sialodacryoadenitis virus genome: sequence of the structural-protein region and its application for differential diagnosis,” *Clinical Diagnostic Laboratory Immunology*, vol. 7, no. 4, pp. 568–573, 2000.

[463] A. K. Haick, J. P. Rzepka, E. Brandon, O. B. Balemba, and T. A. Miura, “Neutrophils are needed for an effective immune response against pulmonary rat coronavirus infection, but also contribute to pathology,” *The Journal of general virology*, vol. 95, no. Pt 3, p. 578, 2014.

[464] L. M. Bradley, M. F. Douglass, D. Chatterjee, S. Akira, and B. J. Baaten, “Matrix metalloprotease 9 mediates neutrophil migration into the airways in response to influenza virus-induced toll-like receptor signaling,” *PLoS pathogens*, vol. 8, no. 4, p. e1002641, 2012.

[465] L. C. Denlinger, R. L. Sorkness, W.-M. Lee, M. D. Evans, M. J. Wolff, S. K. Mathur, G. M. Crisafi, K. L. Gaworski, T. E. Pappas, R. F. Vrtis *et al.*, “Lower airway rhinovirus burden and the seasonal risk of asthma exacerbation,” *American journal of respiratory and critical care medicine*, vol. 184, no. 9, pp. 1007–1014, 2011.

[466] A. Khanolkar, S. M. Hartwig, B. A. Haag, D. K. Meyerholz, J. T. Harty, and S. M. Varga, “Toll-like receptor 4 deficiency increases disease and mortality after mouse hepatitis virus type 1 infection of susceptible c3h mice,” *Journal of virology*, vol. 83, no. 17, pp. 8946–8956, 2009.

[467] N. Nagata, N. Iwata, H. Hasegawa, S. Fukushi, A. Harashima, Y. Sato, M. Saijo, F. Taguchi, S. Morikawa, and T. Sata, “Mouse-passaged severe acute respiratory syndrome-associated coronavirus leads to lethal pulmonary edema and diffuse alveolar damage in adult but not young mice,” *The American journal of pathology*, vol. 172, no. 6, pp. 1625–1637, 2008.

[468] B. Khorsand, A. Savadi, and M. Naghibzadeh, “Sars-cov-2-human protein-protein interaction network,” *Informatics in medicine unlocked*, vol. 20, p. 100413, 2020.

[469] K. Dick, K. K. Biggar, and J. R. Green, "Computational prediction of the comprehensive sars-cov-2 vs. human interactome to guide the design of therapeutics," *BioRxiv*, pp. 2020–03, 2020.

[470] A. Schoenrock, F. Dehne, J. R. Green, A. Golshani, and S. Pitre, "Mp-pipe: a massively parallel protein-protein interaction prediction engine," in *Proceedings of the international conference on Supercomputing*, 2011, pp. 327–337.

[471] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan *et al.*, "Pipe: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–15, 2006.

[472] S. Pitre, M. Hooshyar, A. Schoenrock, B. Samanfar, M. Jessulat, J. R. Green, F. Dehne, and A. Golshani, "Short co-occurring polypeptide regions can predict global protein interaction maps," *Scientific reports*, vol. 2, no. 1, p. 239, 2012.

[473] M. Naseer, U. Aslam, B. Khalid, and B. Chen, "Green route to synthesize zinc oxide nanoparticles using leaf extracts of cassia fistula and melia azadarach and their antibacterial potential," *Scientific Reports*, vol. 10, no. 1, p. 9055, 2020.

[474] Y. Li and L. Ilie, "Sprint: ultrafast protein-protein interaction prediction of the entire human interactome," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–11, 2017.

[475] P. Sun, J. Guo, R. Winnenburg, and J. Baumbach, "Drug repurposing by integrated literature mining and drug–gene–disease triangulation," *Drug discovery today*, vol. 22, no. 4, pp. 615–619, 2017.

[476] E.-k. Tan and W. Ondo, "Restless legs syndrome: clinical features and treatment," *The American journal of the medical sciences*, vol. 319, no. 6, pp. 397–403, 2000.

[477] A. Andreou, S. Trantza, D. Filippou, N. Sipsas, and S. Tsiodras, "Covid-19: The potential role of copper and n-acetylcysteine (nac) in a combination of candidate antiviral treatments against sars-cov-2," *in vivo*, vol. 34, no. 3 suppl, pp. 1567–1588, 2020.

[478] S. Kumar and M. Choudhary, "Synthesis and characterization of novel copper (ii) complexes as potential drug candidates against sars-cov-2 main protease," *New Journal of Chemistry*, vol. 46, no. 10, pp. 4911–4926, 2022.

[479] I. Wessels, B. Rolles, and L. Rink, “The potential impact of zinc supplementation on covid-19 pathogenesis,” *Frontiers in immunology*, p. 1712, 2020.

[480] M. Chilvers, M. McKean, A. Rutman, B. Myint, M. Silverman, and C. O’Callaghan, “The effects of coronavirus on human nasal ciliated respiratory epithelium,” *European Respiratory Journal*, vol. 18, no. 6, pp. 965–970, 2001.

[481] A. Darma, I. G. M. R. G. Ranuh, W. Merbawani, R. A. Setyoningrum, B. Hidajat, S. N. Hidayati, A. Endaryanto, S. M. Sudarmo *et al.*, “Zinc supplementation effect on the bronchial cilia length, the number of cilia, and the number of intact bronchial cell in zinc deficiency rats,” *The Indonesian Biomedical Journal*, vol. 12, no. 1, pp. 78–84, 2020.

[482] L. Szarpak, M. Pruc, A. Gasecka, M. J. Jaguszewski, T. Michalski, F. W. Peacock, J. Smereka, K. Pytkowska, and K. J. Filipiak, “Should we supplement zinc in covid-19 patients? evidence from meta-analysis,” *Pol. Arch. Intern. Med*, vol. 131, pp. 802–807, 2021.

[483] V. Chinni, J. El-Khoury, M. Perera, R. Bellomo, D. Jones, D. Bolton, J. Ischia, and O. Patel, “Zinc supplementation as an adjunct therapy for covid-19: Challenges and opportunities,” *British journal of clinical pharmacology*, vol. 87, no. 10, pp. 3737–3746, 2021.

[484] G. S. Kumar, A. Vadgaonkar, S. Purunaik, R. Shelatkar, V. G. Vaidya Sr, G. Ganu, A. Vadgaonkar, and S. Joshi, “Efficacy and safety of aspirin, promethazine, and micronutrients for rapid clinical recovery in mild to moderate covid-19 patients: A randomized controlled clinical trial,” *Cureus*, vol. 14, no. 5, 2022.

[485] K. W. Barber and J. Rinehart, “The abcs of ptms,” *Nature chemical biology*, vol. 14, no. 3, pp. 188–192, 2018.

[486] J. Jiang, V. Suppiramaniam, and M. W. Wooten, “Posttranslational modifications and receptor-associated proteins in ampa receptor trafficking and synaptic plasticity,” *Neurosignals*, vol. 15, no. 5, pp. 266–282, 2006.

[487] M. P. Lussier, A. Sanz-Clemente, and K. W. Roche, “Dynamic regulation of n-methyl-d-aspartate (nmda) and α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (ampa) receptors by posttranslational modifications,” *Journal of Biological Chemistry*, vol. 290, no. 48, pp. 28 596–28 603, 2015.

[488] H. Ghosh, L. Auguadri, S. Battaglia, Z. Simone Thirouin, K. Zemoura, S. Messner, M. A. Acuña, H. Wildner, G. E. Yévenes, A. Dieter *et al.*, "Several post-translational modifications act in concert to regulate gephyrin scaffolding and gabaergic transmission," *Nature communications*, vol. 7, no. 1, p. 13365, 2016.

[489] D. Vallejo, J. F. Codocedo, and N. C. Inestrosa, "Posttranslational modifications regulate the postsynaptic localization of psd-95," *Molecular neurobiology*, vol. 54, no. 3, pp. 1759–1776, 2017.

[490] S. A. Bradley and J. R. Steinert, "Nitric oxide-mediated posttranslational modifications: impacts at the synapse," *Oxidative medicine and cellular longevity*, vol. 2016, 2016.

[491] Y. Fukata and M. Fukata, "Protein palmitoylation in neuronal development and synaptic plasticity," *Nature Reviews Neuroscience*, vol. 11, no. 3, pp. 161–175, 2010.

[492] R. Kang, J. Wan, P. Arstikaitis, H. Takahashi, K. Huang, A. O. Bailey, J. X. Thompson, A. F. Roth, R. C. Drisdel, R. Mastro *et al.*, "Neural palmitoyl-proteomics reveals dynamic synaptic palmitoylation," *Nature*, vol. 456, no. 7224, pp. 904–909, 2008.

[493] M. M. Zhang and H. C. Hang, "Protein s-palmitoylation in cellular differentiation," *Biochemical Society Transactions*, vol. 45, no. 1, pp. 275–285, 2017.

[494] M. Fröhlich, B. Dejanovic, H. Kashkar, G. Schwarz, and S. Nussberger, "S-palmitoylation represents a novel mechanism regulating the mitochondrial targeting of bax and initiation of apoptosis," *Cell death & disease*, vol. 5, no. 2, pp. e1057–e1057, 2014.

[495] X. Meckler, J. Roseman, P. Das, H. Cheng, S. Pei, M. Keat, B. Kassarjian, T. E. Golde, A. T. Parent, and G. Thinakaran, "Reduced alzheimer's disease β-amyloid deposition in transgenic mice expressing s-palmitoylation-deficient aph1al and nicastrin," *Journal of Neuroscience*, vol. 30, no. 48, pp. 16 160–16 169, 2010.

[496] A. L. Pinner, J. Tucholski, V. Haroutunian, R. E. McCullumsmith, and J. H. Meador-Woodruff, "Decreased protein s-palmitoylation in dorsolateral prefrontal cortex in schizophrenia," *Schizophrenia research*, vol. 177, no. 1-3, pp. 78–87, 2016.

[497] B. Chen, B. Zheng, M. DeRan, G. K. Jarugumilli, J. Fu, Y. S. Brooks, and X. Wu, "Zdhhc7-mediated s-palmitoylation of scribble regulates cell polarity," *Nature chemical biology*, vol. 12, no. 9, pp. 686–693, 2016.

[498] I. De and S. Sadhukhan, "Emerging roles of dhhc-mediated protein s-palmitoylation in physiological and pathophysiological context," *European journal of cell biology*, vol. 97, no. 5, pp. 319–338, 2018.

[499] J. Greaves and L. H. Chamberlain, "Dhhc palmitoyl transferases: substrate interactions and (patho) physiology," *Trends in biochemical sciences*, vol. 36, no. 5, pp. 245–253, 2011.

[500] K. T. Woodley and M. O. Collins, "Quantitative analysis of protein s-acylation site dynamics using site-specific acyl-biotin exchange (ssabe)," in *Mass Spectrometry of Proteins.* Springer, 2019, pp. 71–82.

[501] L. Breiman, "Random forest, vol. 45," *Mach Learn*, vol. 1, 2001.

[502] S. Kawashima and M. Kanehisa, "Aaindex: amino acid index database," *Nucleic acids research*, vol. 28, no. 1, pp. 374–374, 2000.

[503] F. Tan, X. Fu, Y. Zhang, and A. G. Bourgeois, "A genetic algorithm-based method for feature subset selection," *Soft Computing*, vol. 12, no. 2, pp. 111–120, 2008.

[504] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* MIT press, 1992.

[505] D. Wang, D. Liu, J. Yuchi, F. He, Y. Jiang, S. Cai, J. Li, and D. Xu, "Musitedeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization," *Nucleic Acids Research*, vol. 48, no. W1, pp. W140–W146, 2020.

[506] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D13–D21, 2007.

[507] U. Consortium, "Update on activities at the universal protein resource (uniprot) in 2013," *Nucleic acids research*, vol. 41, no. D1, pp. D43–D47, 2012.

[508] D. Chicco and G. Jurman, “The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation,” *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[509] A. Elangovan, Y. Li, D. E. Pires, M. J. Davis, and K. Verspoor, “Large-scale protein-protein post-translational modification extraction with distant supervision and confidence calibrated biobert,” *BMC bioinformatics*, vol. 23, pp. 1–23, 2022.

PhD thesis

By: Soumyendu Sekhar Bandyopadhyay

As of: Sep 19, 2023 8:05:23 AM
70,030 words - 179 matches - 66 sources

Similarity Index

7%

Mode: Summary Report

sources:

390 words / 1% - Internet from 15-Nov-2022 12:00AM
peerj.com

382 words / 1% - Internet from 14-Oct-2021 12:00AM
peerj.com

305 words / 1% - Internet from 22-Feb-2023 12:00AM
www.researchgate.net

86 words / < 1% match - Internet from 14-Oct-2021 12:00AM
peerj.com

85 words / < 1% match - Internet from 15-Nov-2022 12:00AM
peerj.com

75 words / < 1% match - Internet from 15-Nov-2022 12:00AM
peerj.com

248 words / < 1% match - Crossref
Pritha Dutta, Subhadip Basu, Mahantapas Kundu. "Assessment of semantic similarity between proteins using information content and topological properties of the Gene Ontology graph", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017

140 words / < 1% match - Internet from 27-Oct-2019 12:00AM
academic.oup.com

54 words / < 1% match - Internet from 29-Nov-2017 12:00AM
academic.oup.com

20 words / < 1% match - Internet from 14-Nov-2022 12:00AM
academic.oup.com

183 words / < 1% match - Internet
Zhao, Chenguang. "GOGO: An Improved Algorithm to Measure the Semantic Similarity Between Gene Ontology Terms", The Aquila Digital Community, 2019

146 words / < 1% match - Crossref
Chenguang Zhao, Zheng Wang. "GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms", Scientific Reports, 2018

104 words / < 1% match - Internet from 28-Feb-2023 12:00AM
discovery.ucl.ac.uk

87 words / < 1% match - Crossref
Kaifu Gao, Rui Wang, Jiahui Chen, Limei Cheng et al. "Methodology-Centered Review of Molecular Modeling, Simulation, and Prediction of SARS-CoV-2", Chemical Reviews, 2022

72 words / < 1% match - Internet from 03-Aug-2011 12:00AM
www.daimi.au.dk

67 words / < 1% match - Internet from 14-Dec-2022 12:00AM
www.mdpi.com

65 words / < 1% match - Crossref
Priyank Purohit, Pobitra Borah, Sangeeta Hazarika, Gaurav Joshi, Pran Kishore Deb. "Chapter 4 Computational Modeling in the Development of Antiviral Agents", Springer Science and Business Media LLC, 2023

65 words / < 1% match - Internet from 17-Dec-2021 12:00AM
www.pubfacts.com

61 words / < 1% match - Internet from 15-Dec-2021 12:00AM
www.x-mol.com

60 words / < 1% match - Crossref
Christian Ebere Enyoh, Tochukwu Maduka, Qingyue Wang, Md. Rezwanul Islam. "In Silico Screening of Active Compounds in Garri for the Inhibition of Key Enzymes Linked to Diabetes Mellitus", ACS Food Science & Technology, 2022

56 words / < 1% match - Crossref Posted Content
Kevin Dick, Kyle K Biggar, James R Green. "Computational Prediction of the Comprehensive SARS-CoV-2 vs. Human Interactome to Guide the Design of Therapeutics", Cold Spring Harbor Laboratory, 2020

45 words / < 1% match - Crossref
Aanzil Akram Halsana, Tapas Chakroborty, Anup Kumar Halder, Subhadip Basu. "DensePPI : A Novel Image-based Deep Learning method for Prediction of Protein-Protein Interactions", IEEE Transactions on NanoBioscience, 2023

43 words / < 1% match - Internet from 06-Jan-2023 12:00AM
www.ijamtes.org

42 words / < 1% match - Crossref
Claudio Cavasotto, Juan Di Filippo. "In silico Drug Repurposing for COVID-19: Targeting SARS-CoV-2 Proteins through Docking and Consensus Ranking", Molecular Informatics, 2020

40 words / < 1% match - Crossref
V. Srinivasa Rao, K. Srinivas, G. N. Sujini, G. N. Sunand Kumar. "Protein-Protein Interaction Detection: Methods and Analysis", International Journal of Proteomics, 2014

38 words / < 1% match - Crossref
Qihui Wang, Jianxun Qi, Yuan Yuan, Yifang Xuan et al. "Bat Origins of MERS-CoV Supported by Bat Coronavirus HKU4Usage of Human Receptor CD26", Cell Host & Microbe, 2014

38 words / < 1% match - Internet
Chen, Xiaojing, Qin, Zhen, An, Le, Bhanu, Bir. "Multiperson Tracking by Online Learned Grouping Model With Nonlinear Motion Context", eScholarship, University of California, 2016

37 words / < 1% match - Crossref
"Pattern Recognition and Machine Intelligence", Springer Science and Business Media LLC, 2017

35 words / < 1% match - Crossref
Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger et al. "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor", Cell, 2020

35 words / < 1% match - Crossref
Peng Sun, Jiong Guo, Rainer Winnenburg, Jan Baumbach. "Drug repurposing by integrated literature mining and drug–gene–disease triangulation", Drug Discovery Today, 2017

35 words / < 1% match - Crossref
Yue Deng, Lin Gao. "ppiPre - an R package for predicting protein-protein interactions", 2012 IEEE 6th International Conference on Systems Biology (ISB), 2012

35 words / < 1% match - Internet from 05-Apr-2020 12:00AM
id.123dok.com

35 words / < 1% match - Internet from 26-Dec-2022 12:00AM
www.biorxiv.org

34 words / < 1% match - Crossref
Aurelia Magdalena Pisoschi, Florin Iordache, Loredana Stanca, Iuliana Gajaila et al. "Antioxidant, Anti-inflammatory, and Immunomodulatory Roles of Nonvitamin Antioxidants in Anti-SARS-CoV-2 Therapy", Journal of Medicinal Chemistry, 2022

34 words / < 1% match - Internet from 09-Aug-2022 12:00AM
ouci.dntb.gov.ua

32 words / < 1% match - Internet from 31-Jul-2019 12:00AM
tessera.spandidos-publications.com

31 words / < 1% match - Crossref
"Intelligent Computing Theories and Application", Springer Science and Business Media LLC, 2017

30 words / < 1% match - Internet from 06-Aug-2022 12:00AM
wikimili.com

29 words / < 1% match - Internet from 21-Feb-2022 12:00AM
figshare.com

28 words / < 1% match - Crossref
Arslan Siraj, Tuvshinbayar Chantsalnyam, Hilal Tayara, Kil To Chong.. "RecSNO: prediction of protein S-Nitrosylation sites using a recurrent neural network", IEEE Access, 2021

28 words / < 1% match - Crossref Posted Content
Yekbun Adiguzel. "Coronavirus-associated molecular mimicry through homology to a SARS-CoV-2 peptide could be leading to susceptibility in patients with HLA-A*02:01 and HLA-A*24:02 serotypes", Cold Spring Harbor Laboratory, 2021

28 words / < 1% match - Internet from 24-Aug-2018 12:00AM
renata.borovica-gajic.com

27 words / < 1% match - Crossref
"Machine Learning for Intelligent Multimedia Analytics", Springer Science and Business Media LLC, 2021

27 words / < 1% match - Crossref
Yi-min Mao, Deborah S. Mwakapesa, Yi-can Li, Kai-bin Xu, Yaser A. Nanehkaran, Mao-sheng Zhang. "Assessment of landslide susceptibility using DBSCAN-AHD and LD-EV methods", Journal of Mountain Science, 2021

27 words / < 1% match - Internet from 22-Nov-2022 12:00AM
www.omicsdi.org

25 words / < 1% match - Crossref
Cody J. Bills, Hongjie Xia, John Yun-Chung Chen, Jason Yeung, Birte Kalveram, David Walker, Xuping Xie, Pei-Yong Shi. "Mutations in SARS-CoV-2 variant nsp6 enhance type-I interferon antagonism", Emerging Microbes & Infections, 2023

24 words / < 1% match - Crossref
"Data Science: From Research to Application", Springer Science and Business Media LLC, 2020

23 words / < 1% match - Internet from 21-Jul-2022 12:00AM
hanmi.co.kr

23 words / < 1% match - Internet from 14-Dec-2022 12:00AM
waseda.repo.nii.ac.jp

22 words / < 1% match - Crossref
Chengyuan Liang, Lei Tian, Yuzhi Liu, Nan Hui et al. "A Promising Antiviral Candidate Drug for the COVID-19 Pandemic: A Mini-Review of Remdesivir", European Journal of Medicinal Chemistry, 2020

22 words / < 1% match - Internet from 04-Oct-2022 12:00AM
coek.info

21 words / < 1% match - Crossref
"Discussion on Big Data: TDFS Vs HDFS", International Journal of Recent Technology and Engineering, 2019

21 words / < 1% match - Crossref
"Intelligent Computing Theories and Application", Springer Science and Business Media LLC, 2020

21 words / < 1% match - Crossref
Shuyan Li, Jiazhong Li, Lulu Ning, Shaopeng Wang, Yuzhen Niu, Nengzhi Jin, Xiaojun Yao, Huanxiang Liu, Lili Xi. "In Silico Identification of Protein S-Palmitoylation Sites and Their Involvement in Human Inherited Disease", Journal of Chemical Information and Modeling, 2015

20 words / < 1% match - Crossref
Mehdi Bouhaddou, Danish Memon, Bjoern Meyer, Kris M. White et al. "The Global Phosphorylation Landscape of SARS-CoV-2 Infection", Cell, 2020

20 words / < 1% match - Internet from 20-Nov-2017 12:00AM
etd.lib.metu.edu.tr

20 words / < 1% match - Internet from 16-Nov-2022 12:00AM
tribuneonlineng.com

20 words / < 1% match - Internet from 23-Nov-2017 12:00AM
www.qucosa.de

18 words / < 1% match - Internet from 23-Jan-2023 12:00AM
medium.com

17 words / < 1% match - Crossref
Sanghamitra Bandyopadhyay, Koushik Mallick. "A New Feature Vector Based on Gene Ontology Terms for Protein-Protein Interaction Prediction", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017

15 words / < 1% match - Crossref
Jiedi Sun, Yuxia Li, Jiangtao Wen, Shengnan Yan. "Novel mixing matrix estimation approach in underdetermined blind source separation", Neurocomputing, 2016

14 words / < 1% match - Crossref
"Proceedings of International Conference on Frontiers in Computing and Systems", Springer Science and Business Media LLC, 2023

14 words / < 1% match - Crossref
Anup Kumar Halder, Piyali Chatterjee, Mita Nasipuri, Dariusz Plewczynski, Subhadip Basu. "3gClust: Human Protein Cluster Analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018

14 words / < 1% match - Crossref
Arijit Chakraborty, Sajal Mitra, Mainak Bhattacharjee, Debashis De, Anindya J. Pal. "Determining human-coronavirus protein-protein interaction using machine intelligence", Medicine in Novel Technology and Devices, 2023

14 words / < 1% match - Crossref
Encyclopedia of Algorithms, 2016.

14 words / < 1% match - Crossref
Juliana Felgueiras, Joana Vieira Silva, Margarida Fardilha. "Adding biological meaning to human protein-protein interactions identified by yeast two-hybrid screenings: A guide through bioinformatics tools", Journal of Proteomics, 2018

paper text:

Chapter 1 Intro duction 1.1 Background A paradigm that describes the flow of genetic information across a living organism is called the Central Dogma of Molecular Biology. In its most fundamental form, it may be broken down into three distinct processes: transcription, translation, and replication. Converting a section of Deoxyribonucleic acid (DNA) into messenger Ribonucleic acid (mRNA) is referred to as transcription. The process by which mRNA is converted into the protein it encodes is called translation. The process of making duplicates of DNA is referred to as replication. The genetic material in the majority of species is made up of DNA; however, there are other organisms, such as retroviruses like HIV, that use RNA as their genetic material instead. The creation of DNA from RNA in these kinds of organisms takes place via a process known as reverse transcription. The process of convering a protein from DNA has been described in Figure 1.1 1.1.1 DNA to RNA to Protein DNA is a molecule with a double helix structure made up of chains of nitrogenous bases and a sugar-phosphate backbone. The Base pairing between the nitrogenous bases in DNA molecules where Guanine interacts with Cytosine with three hydrogen bonds and Thymine interacts with Adenine with two hydrogen bonds. The double Figure 1.1: Central Dogma of Molecular Biology. The figure depicts the scheme for the construction of proteins from DNA molecules. Figure 1.2: The structure of a single amino acid where alpha carbon is attached with a carboxylic group-COOH, an amine group (-N H2), and a side chain R which differentiate the chemical properties of different amino acid helix structure aids in self-replication since it is self-complementary. Evolution is the result of flaws in the replicating process. Proteins and RNA are synthesized to carry out instructions from DNA. In

vaccines

Article

Assessment of GO-Based Protein Interaction Affinities in the Large-Scale Human–Coronavirus Family Interactome

Soumyendu Sekhar Bandyopadhyay [1,2], Anup Kumar Halder [3], Sovan Saha [4], Piyali Chatterjee [5], Mita Nasipuri [1] and Subhadip Basu [1,*]

[1] Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India
[2] Department of Computer Science and Engineering, School of Engineering and Technology, Adamas University, Kolkata 700126, India
[3] Faculty of Mathematics and Information Sciences, Warsaw University of Technology, 00-662 Warsaw, Poland
[4] Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Techno Main Salt Lake, Sector V, Kolkata 700091, India
[5] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata 700152, India
* Correspondence: subhadip.basu@jadavpuruniversity.in

Abstract: SARS-CoV-2 is a novel coronavirus that replicates itself via interacting with the host proteins. As a result, identifying virus and host protein-protein interactions could help researchers better understand the virus disease transmission behavior and identify possible COVID-19 drugs. The International Committee on Virus Taxonomy has determined that nCoV is genetically 89% compared to the SARS-CoV epidemic in 2003. This paper focuses on assessing the host–pathogen protein interaction affinity of the coronavirus family, having 44 different variants. In light of these considerations, a GO-semantic scoring function is provided based on Gene Ontology (GO) graphs for determining the binding affinity of any two proteins at the organism level. Based on the availability of the GO annotation of the proteins, 11 viral variants, *viz.*, SARS-CoV-2, SARS, MERS, *Bat coronavirus* HKU3, *Bat coronavirus* Rp3/2004, *Bat coronavirus* HKU5, *Murine coronavirus*, *Bovine coronavirus*, Rat coronavirus, *Bat coronavirus* HKU4, *Bat coronavirus* 133/2005, are considered from 44 viral variants. The fuzzy scoring function of the entire host–pathogen network has been processed with ~180 million potential interactions generated from 19,281 host proteins and around 242 viral proteins. ~4.5 million potential level one host–pathogen interactions are computed based on the estimated interaction affinity threshold. The resulting host–pathogen interactome is also validated with *state-of-the-art* experimental networks. The study has also been extended further toward the drug-repurposing study by analyzing the FDA-listed COVID drugs.

Keywords: COVID-19; SARS-CoV-2; COVID-19 variants; go-semantic score; gene ontology; COVID-19 drugs; protein–protein interaction network

Citation: Bandyopadhyay, S.S.; Halder, A.K.; Saha, S.; Chatterjee, P.; Nasipuri, M.; Basu, S. Assessment of GO-Based Protein Interaction Affinities in the Large-Scale Human–Coronavirus Family Interactome. *Vaccines* **2023**, *11*, 549. https://doi.org/10.3390/vaccines11030549

Academic Editor: Rajan Kumar Pandey

Received: 9 January 2023
Revised: 19 February 2023
Accepted: 23 February 2023
Published: 25 February 2023

1. Introduction

The emerging coronavirus (CoV) pandemic has sparked a flurry of research into the SARS-CoV-2 virus and the COVID-19 disease it causes in people [1]. COVID-19 was identified in Wuhan (Hubei province) [2]. It starts spreading soon to other nations. On 30 January 2020, World Health Organization (WHO) declared this outbreak of nCoV as a global emergency [3]. A coronavirus is a member of the family Coronaviridae.

Along with humans, it also affects mammals and birds. Even though the coronavirus typically causes the common cold, cough, etc., it also causes severe acute, chronic respiratory disease, multiple organ failure, and, ultimately, human mortality. Before SARS-CoV-2, the two primary outbreaks were Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). Southern China was the location of SARS's inception. Its fatality rate was between 14 and 15% [4]. The MERS outbreak was supposed to start in

Saudi Arabia. In the fight against the MERS virus, 858 out of 2494 afflicted cases prevailed. As a result, it produced a substantially higher death rate of 34.4% compared to the SARS.

Regarding biology, the three epidemic-starting viruses, SARS, MERS, and SARS-CoV-2, belong to Coronaviridae's genus Beta coronavirus. Proteins that are both structural and non-structural contribute to the development of SARS-CoV-2. Out of the two, structural proteins such as the spike (S) protein, nucleocapsid (N) protein, membrane (M) protein, and envelope (E) protein play a crucial part in spreading the disease by binding with receptors after entering the human body [5].

The primary factor which needs to be considered while examining the disease transmission process from SARS-CoV-2 to humans is the Protein–Protein Interaction Network (PPIN). It is critical for determining essential proteins and functions [6–19] responsible for various diseases. The primary focus of research has changed from the study of the PPIN underlying various types of human diseases to the study of the PPIN due to the improvement in the availability of human PPIN data [20]. According to the report, SARS-CoV-2 has ~89% similarity with SARS-CoV [21,22]. SARS-CoV, a disease that initially appeared in the Guangdong Province of China in November 2002, spread to 28 regions worldwide in 2003 and resulted in 774 fatalities among the 8096 people with COVID-19 [23–25]. According to phylogenetic analysis, it was assumed that SARS-CoV was different from previously known coronaviruses [26,27]. Even though the etiological agent was discovered and molecular research on the SARS-CoV advanced quite quickly, the mystery surrounding the disease's cause remained unsolved. Data indicated that SARS was an animal-borne disease from the beginning [23,24,28,29]. After the surge of SARS-CoV in 2012, there was another coronavirus surge, Middle East Respiratory Syndrome (MERS), in Jordon. A bat and numerous dromedary camels have been reported to have MERS-CoV sequences (DC). MERS-CoV is an enzootic disease in the Arabian Peninsula, portions of Africa, and the Middle East. It affects camels as its primary reservoir and occasionally, but infrequently, infects humans [30]. MERS-CoV is a member of the Beta coronavirus family. World Health Organization (WHO) confirmed 2220 people with COVID-19 along with 790 deaths for MERS-CoV [31]. There is a 35% fatality rate from MERS. MERS is not specifically treated. MERS-CoV outbreaks in hospitals and homes are brought on by person-to-person transmission [32].

A beta-CoV prevalent in wild mice, the mouse hepatitis virus (MHV) or Murine-CoV is similar to SARS-CoV-2. In-depth research has been done on laboratory MHV strains to understand host antiviral defense systems and coronavirus virulence factors [33]. Murine-CoV contains several strains that induce variable symptoms in the respiratory, digestive, hepatic, and neurological systems [34–36]. The genus of beta-CoVs includes all MHV strains and certain human CoVs (HCoV-OC43, HCoV-HKU1, SARS-CoV, MERS-CoV, and SARS-CoV-2). The tropism and pathogenicity of various MHV strains vary, and research on recombinant MHV variations has uncovered host and viral variables that affect viral propagation or evade immune Identification [37].

The wide variety of mammalian and avian species that coronaviruses have been found to infect and the highly varied disease syndromes they cause are well known. One of the well-known traits of several CoVs is variable tissue tropism, which also allows them to overcome interspecies boundaries easily. Betacoronaviruses, known as bovine CoVs (BCoVs), cause shipping fever, winter dysentery in older cattle, and neonatal calf diarrhea. Interestingly, there have not been any specific genetic or antigenic markers found in BCoVs linked to these unique clinical disorders. BCoVs, on the other hand, are quasispecies that coexists with other CoVs. In addition to cattle, BCoVs and CoVs resembling cattle were found in several domestic and wild ruminant species, dogs, and humans [38]. The pneumoenteric virus known as the bovine coronavirus (BCoV) is a member of the Betacoronavirus 1 genus. Because of several instances of genetic recombination and interspecies transmission, members of the Betacoronavirus 1 species appear to be host-range variants descended from the same parental virus due to their close antigenic and genetic relatedness [39–42].

Two separate teams reported finding SARS-like CoVs (SL-CoVs) in bats in 2005, and they hypothesized that bats were SARS-CoV natural reservoirs [43,44]. Most bat SL-CoVs were discovered in rhinolopus bats, especially Rhinolophus sinicus. They share 87 to 92% of their nucleic acid and 93 to 100% of their amino acid sequences with the SARS-CoV [43–47]. According to a phylogenetic study, MERS-CoV is a member of lineage C of the Betacoronavirus genus. It resembled the pipistrelle bat (Pipistrellus pipistrellus) and lesser bamboo bat (*Tylonycteris pachypus*) most closely, as well as the bat coronaviruses HKU4 and HKU5 [31,48]. The whole genomic sequences of HKU4 and HKU5 and the RNA-dependent RNA polymerase (RdRp) gene show nucleotide identity with MERS-CoV of 50% and 82%, respectively. A recent study established that CD26, also known as dipeptidyl peptidase 4 (DPPIV), is a functional receptor for MERS-CoV. Additionally, it has been demonstrated that this molecule is evolutionarily conserved among mammals and that MERS-CoV can infect a wide variety of mammalian cells (including those from humans, pigs, monkeys, and bats), indicating ease of transmission between hosts [49,50].

A large-scale PPI network of an organism provides valuable clues for understanding cellular and molecular functionalities, and signaling pathways can provide crucial insights into the disease mechanism, etc. Much biological information is available and encoded in different ontologies called Gene Ontology. Semantic similarity is the degree of relatedness between the two biological entities (Gene/Protein) based on GO annotations that provide a quantitative measure of their GO-level relationship [51]. Different combinations of edge-based and node-based semantic similarity measures have been applied over the years from gene ontology graphs [52–63]. These methods have specific shortcomings concerning their designed GO semantic features. Some of them have used topological properties of the GO graph, some have used only the information content (IC) of the most informative common ancestor [52,53,55,56], and some have used DCA [58–60] based approach. To define the interaction affinity of any two proteins from their GO information, this hybrid approach is more effective as it incorporates topological features and average IC-based DCA techniques. Much work [64] has already been done to analyze host–pathogenic interactions [65,66], disease detection [67], and disease-specific multi-omics network analyses [68].

From the above discussion, it is clear that several similar studies based on GO information have been done on host–pathogen interaction networks. However, a complete PPIN must be identified for humans and different coronavirus organisms to detect probable human targets from all perspectives. So, in this study, the interaction affinity between the protein pairs from the different organisms of the coronavirus family and human spreader proteins is calculated using the available ontological information using the proposed in-silico model. Section 2 describes the proposed in-silico model for calculating the interaction affinity of the bait-prey protein pairs in an apache spark-based parallel computational environment. Section 2.2 gives a detailed description of the database used for different coronavirus organisms. The results are discussed in Section 3, which includes host–pathogen protein interactions for the different organisms of the coronavirus family and validation of our proposed in-silico model using the *state-of-the-art* database.

2. Materials and Methods

A GO-based Graph theoretic model is proposed to determine the interaction affinity between the host–pathogen protein pairs for humans and different coronavirus organisms. Currently, 19,281 human proteins have GO annotations, whereas around 242 viral proteins are obtained from a selected organism having GO annotations. Based on the above data, level 1 interactors generates ~4.5 million potential host–pathogen interaction. The variety and veracity issue plays a significant role in such a large-scale dynamic PPI network. Handling large, dynamic, heterogeneous networks using in-silico methods is tedious. Therefore, an Apache Spark-Based analytical study is proposed to compute the interaction affinity in large-scale protein–protein interaction networks using the Gene Ontology (GO) graph.

2.1. GO Graph-Based Scoring for Potential Host–Pathogen Protein Interaction Identification

Combining the similarity scores of the GO terms connected to the proteins will yield an estimate of the semantic similarity between two interacting proteins [52,66,69,70]. The greater the similarity between two GO pairs, the greater the interaction affinity between the proteins. The GO hierarchy's independent directed acyclic graphs (DAGs) represent three distinct features of proteins: cellular component (CC), biological process (BP), and molecular function (CC). Each node represents GO terms, and edges indicate various hierarchical relationships. The two fundamental relations *"is_a"* and *"part of"* GO graphs are considered for semantic score computation. Considering the similarity between all the GO pairs, the semantic similarity of the protein pairs can be estimated. The shortest path length between a pair of terms in a GO graph and the average information content *(IC)* [57] of the disjunctive common ancestors (*DsjCA*) of the respective GO term [52,70] measures the similarity of the pair. Our proposed method based on the GO graph is fuzzy clustered, and the degree of relationship between each GO term and the cluster center determines which GO term is chosen as the cluster center. The cluster centers are then chosen using the GO term proportion measure. The proportion measure of any GO term t is given by

$$\mathrm{PrT}(t) = \frac{|AnC(t)| + |DnC(t)|}{|No|} \tag{1}$$

where *AnC(t)* is the ascendant term for t and *DnC(t)* is the descendent term of t. *No* is the total number of GO terms in ontology O, and *PrT(t)* is the proportion measure of term t. The GO keywords chosen as cluster centers are those for which this proportion metric is higher than a certain threshold. The cluster centers in this study are selected using the proposed threshold values [66,69]. Once the cluster centers have been chosen, the shortest path lengths between each term in the ontology and the cluster centers have been calculated. The membership value of a GO term decreases with the increase in the shortest path length. The membership function of a GO term is given by

$$Mfn_c(t) = \mathrm{e}^{-\frac{-(x-c_i)^2}{2k^2}} \tag{2}$$

where c_i is the i^{th} cluster center, x is the shortest path length, and k is the width of the membership function. If no path from any GO term to a cluster center is found, then the membership of the GO term with respect to that cluster center will be considered 0. Similar membership for any target GO pair indicates very closely related concepts of GO functionality, and widely related membership value represents separated concepts. For any target pair of GO term (t_i,t_j), a weight parameter is introduced to estimate these differences in membership. The weight parameter is thus defined by

$$WT(t_{i,}\ t_j) = 1 - maxD\ (t_i,t_j)$$

where $maxD$(t_i,t_j) represents the maximum difference in membership values of GO pair (t_i,t_j) across all cluster centers of any particular GO graph type(CC/MF/BP).

The information content (*IC*) based information of the disjunctive common ancestor (*DsjCAs*) of any GO graph is more significant in the semantic similarity assessment of two GO terms [60]. *IC* of any GO term t, with respect to a GO graph, g is defined as *ICg(t) = −log(Pr(t))*. The probability *Pr(t)* is the occurrences of term t with respect to the total annotations of GO graph g. The occurrences of term *t* depend on its annotations over the protein corpus. Using the *IC* of the *DsjCA*, the shared information content (*SIC*) is computed for the target GO term pair (t_i,t_j). The SIC is computed as

$$SIC(t_i, t_j) = \frac{\Sigma a \in DsjCA^{IC(a)}}{|DsjCA(t_i, t_j)|} \tag{3}$$

Finally, the semantic similarity between two GO pair t_i and t_j is calculated as

$$SSt_it_j = WT(t_i, t_j) \times SIC(t_i, t_j) \tag{4}$$

When comparing the annotations of the proteins P_i and P_j for each type of GO, the maximum similarity of all possible GO pairs is used to determine the semantic similarity of the protein pair (P_i, P_j) for each GO type (CC, MF, and BP). The average of the CC, MF, and BP-based semantic similarity is used to define the protein pair's interaction affinity (P_i, P_j). Figure 1 refers to the schematic diagram of our proposed model where the host–pathogen interaction affinity between humans and organisms from the coronavirus family is calculated using the GO information, resulting in high-quality interactions for retrieving vulnerable human prey for coronavirus hosts.

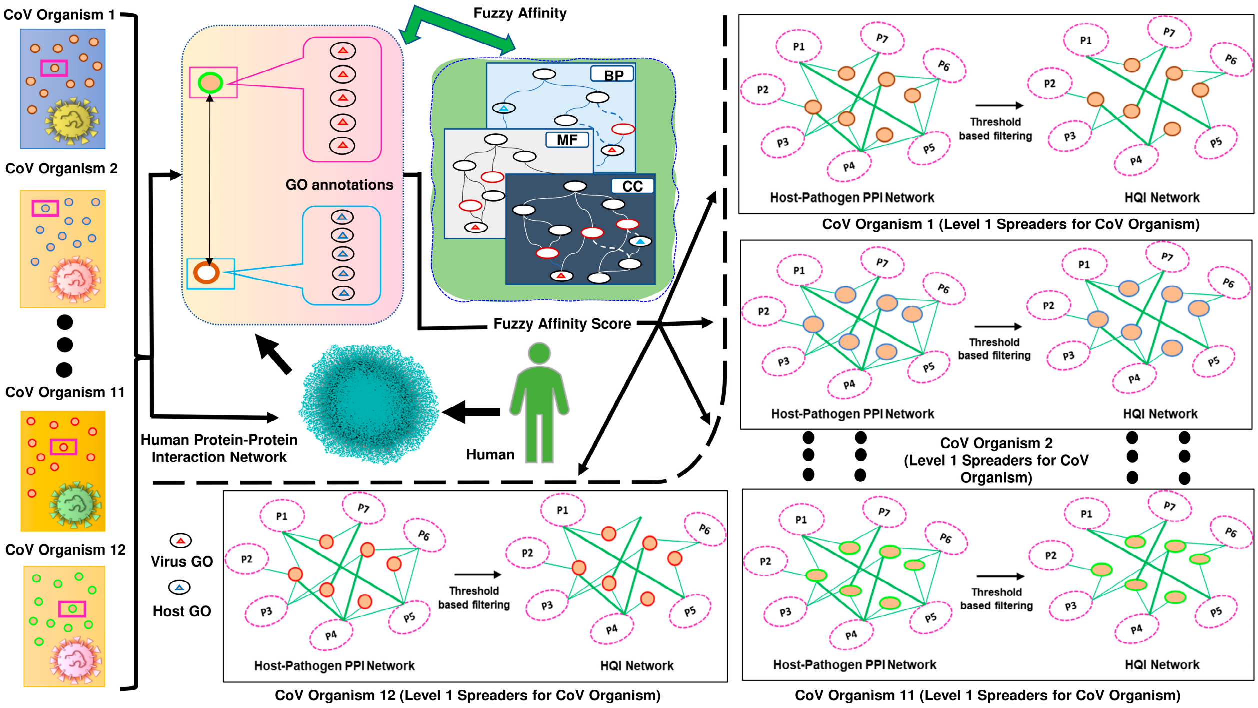

CoV Organism 1
CoV Organism 2
CoV Organism 11
CoV Organism 12
Fuzzy Affinity
BP
MF
CC
GO annotations
Fuzzy Affinity Score
Human Protein-Protein Interaction Network
Human
Virus GO
Host GO
P1
P2
P3
P4
P5
P6
P7
Threshold based filtering
Host-Pathogen PPI Network
HQI Network
CoV Organism 1 (Level 1 Spreaders for CoV Organism)
CoV Organism 2 (Level 1 Spreaders for CoV Organism)
CoV Organism 12 (Level 1 Spreaders for CoV Organism)
CoV Organism 11 (Level 1 Spreaders for CoV Organism)

Figure 1. Schematic diagram of our proposed model. The coronavirus and human proteins' interaction affinities are determined by the model using gene ontology information of the proteins. Three different GO-relationship graphs, CC, MF, and BP, are used to evaluate all GO pair-wise interaction affinities. A protein pair's fuzzy interaction affinity is calculated using the three pair-wise scores of all GO-pair affinities.

2.2. Dataset Preparation

Alpha-, Beta-, Gamma-, and Delta-CoV are the four genera that comprise the enormous family of enveloped positive-strand RNA viruses known as coronaviruses (CoVs). Among all the 44 organisms of coronavirus, here in this work, only 11 organisms have been considered based on the available GO-annotated proteins. The human is considered the host, and the work mainly suggests the affinity of host–pathogen interaction for different coronavirus organisms. Below, a brief description of all selected organisms is given.

2.2.1. Human Protein

All potential interactions between human proteins that have been experimentally verified in humans make up the dataset [71,72]. The proteins in the Human organism are represented by nodes, whereas the edges represent the respective interactions between the organism. The proteins and their GO annotations are collected from UniProt, the protein repository [73]. UniProt contains 20,386 reviewed human proteins, among which 19,283 proteins are associated with GO annotations.

2.2.2. SARS-CoV-2 Proteins

SARS-CoV-2 is a biological member of the Coronaviridae, which belongs to the genus Beta coronavirus. The virus contains four structural proteins, namely envelop(E) protein, membrane(M) protein, nucleocapsid(N) protein, and spike(S) protein, which helps in binding with receptors after entering the human body and has a crucial function in spreading the disease [5]. Here the work is carried out by collecting the dataset of available SARS-CoV-2 protein from UniProtKB. The repository includes 16 reviewed SARS-CoV-2 proteins as of date.

2.2.3. SARS-CoV Proteins

SARS-CoV is a highly pathogenic and zoonotic virus that causes severe respiratory illness, gastrointestinal, neurological, and fatalities among humans [74–76]. The 2002-2003 severe acute respiratory syndrome (SARS) pandemic showed how susceptible humans are to CoV epidemics [77]. However, the dataset is collected from UniProtKB, which holds 15 reviewed SARS-CoV proteins.

2.2.4. MERS-CoV Proteins

MERS-CoV is also a member of Beta-Coronavirus. It is an even more pathogenic and zoonotic virus in comparison to SARS-CoV. MERS-CoV immerged around 2012 in the Arabian Peninsula with very high transmissibility by affecting more than 2000 people [78]. The dataset has been retrieved from UniProtKB, which holds around 10 MERS-CoV proteins.

2.2.5. *Bat coronavirus* HKU3 Proteins

Surveillance research in Hong Kong among non-caged animals from wild regions found that a closely similar bat coronavirus, SARS-related Rhinolophus bat coronavirus HKU3, was the natural animal host [79]. We have retrieved a protein set of *Bat coronavirus* HKU3 from UniProtKB, having 12 proteins.

2.2.6. *Bat coronavirus* RP3/2004 Proteins

With the high geographic spread and species variety, bats represent an order with significant evolutionary success. Bats are the natural reservoirs of several viruses closely related to SARS-CoV [80]. A search for ACE2 sequence similarities in domestic and wild animals in Italy revealed domestic (horses, cats, cattle, and sheep) and wild (European rabbits and grizzly bears) animal species as potential SARS-CoV-2 secondary reservoirs. Molecular docking of these species' ACE2 against the S protein of the *Bat coronavirus* (Bt-CoV/Rp3/2004) suggests that the primary reservoir *Rhinolophus ferrumequinum* may infect secondary reservoirs, domestic and animals living in Italy [81].

2.2.7. *Bat coronavirus* HKU5 Proteins

An enclosed, positive-sense single-stranded RNA mammalian Group 2 Betacoronavirus called bat coronavirus HKU5 (Bat-CoV HKU5) was found in Japanese Pipistrellus in Hong Kong. This coronavirus strain is closely related to the recently discovered novel MERS-CoV, which is to blame for the coronavirus outbreaks linked to the Middle East respiratory illness in 2012 [31,82].

2.2.8. *Bat coronavirus* HKU4 Proteins

Tylonycteris bat coronavirus HKU4 (Bat-CoV HKU4), a member of Betacoronavirus, is an enveloped, single-stranded virus having a genetical similarity with MERS-CoV or HCoV-EMC. The main difference between HCoV-EMC and Bat-CoV HKU4 lies in between the spike protein (S) and envelop (E) protein, where HCoV-EMC have five ORFs instead of four with low amino acid identities to Bat-CoV HKU4 [83]. The human CD26 (hCD26) receptor is engaged explicitly by a receptor binding domain (RBD) in the MERS-CoV envelope-embedded spike protein to start viral entry. Due to the viral spike protein's great sequence identity, we looked into whether or not HKU4 and HKU5 can detect hCD26 for cell entrance. We discovered that HKU4-RBD binds to hCD26, but not HKU5-RBD, and that pseudotyped viruses incorporating HKU4 spike can infect cells by recognizing hCD26. The overall hCD26-binding mechanism of the HKU4-RBD/hCD26 complex was identical to that of the MERS-RBD, according to the structure. However, HKU4-RBD has a lower affinity for receptor binding than MERS-RBD because it is less suited to hCD26 [84].

2.2.9. *Bat coronavirus* 133/2005

The spike (S1) and RNA-dependent RNA polymerase proteins of MERS-CoV were subjected to phylogenetic analysis, which indicated that the virus is linked to bat viruses. Coronavirus surveillance investigations in several populations of bats have shown that they are potential reservoirs for this unique virus [85]. Different phylogenetic studies reveal that MERS-CoV was grouped with the Betacoronavirus genus, particularly near BtCoV/133/2005 and BtCoV HKU4-2, which had the most significant S1 amino acid sequence similarity (60%) with MERS-CoV [86].

2.2.10. *Murine coronavirus*

Murine coronavirus (M-CoV), a member of the Betacoronavirus family having Embacovirus subgenus, is mainly found responsible for infecting rats [87,88]. Enterotropic and Polytropic are the two strains of M-CoV. Mouse hepatitis virus (MHV) strains D, Y, RI, and DVIM are examples of enterotropic strains. In contrast, hepatitis, enteritis, and encephalitis are the leading causes of illness caused by polytropic strains like JHM and A59 [89]. *Murine coronaviruses* come in over 25 distinct strains. These viruses, which spread by the fecal-oral or respiratory routes and infect mice's livers, have been utilized as an animal disease model for hepatitis [90]. The strains MHV-D, MHV-DVIM, MHV-Y, and MHV-RI, which are transmitted in fecal matter, primarily affect the digestive tract. However, they can occasionally affect the spleen, liver, and lymphatic tissue [91].

2.2.11. *Bovine coronavirus*

Bovine coronavirus (BCoV) is a member of Betacoronavirus 1, and it can infect both cattle and humans [92,93]. It is also an enveloped single-stranded RNA virus that enters the host cell by binding itself with the N-acetyl-9-O-acetylneuraminic acid receptor [94,95]. BCov is mainly responsible for causing gastroenteritis in calves resulting in massive economic damage [96]. BCoV consisted of five structural proteins, namely (S) spike glycoprotein; (M) integral membrane protein; (HE) hemagglutinin-esterase glycoprotein; (E) small membrane protein, and (N) nucleocapsid phosphoprotein [97]. A phosphoprotein with a high content of essential amino acids, the N protein joins the genomic RNA directly to create a helicoidal nucleocapsid. The N protein carries out numerous activities related to viral pathogenicity, transcription, and replication. Because it is a highly conserved protein expressed in significant amounts during viral replication, it is frequently employed for molecular diagnosis of BCoV [98].

2.2.12. *Rat coronavirus*

Rat coronavirus (RCoV), subset of *Murine coronavirus*, is also a single stranded RNA virus belonging to Betacoronavirus family which is responsioble for infecting rats [99]. The respiratory disease in adult rats is caused by RCoV in adult rats, which is characterized by an early Polymorphonuclear neutrophils (PMN) response, viral multiplication, inflammatory lung lesions, modest weight loss, and efficient infection resolution [100]. When a virus is present, PMN in the respiratory tract is typically associated with severe disease pathology [101–104].

3. Results

Our developed in-silico model contains the protein interaction affinity between humans and different organisms from the coronavirus family. The in-silico model is validated by identifying the overlapped edges with reference to the *state-of-the-art* datasets. Any computational model must always consider the input and output source, and our suggested model is no exception.

3.1. Identification of Host–Pathogen Protein Interactions for the Different Organisms of the Coronavirus Family

Three different forms of GO hierarchical connection graphs can be used to use the GO information to infer the binding affinity of each pair of interacting proteins (CC, MF, and BP) [64]. Our proposed GO-based in-silico model is applied to find the interaction affinity between the host protein and different organisms of the coronavirus family. Among 44 different organisms of the coronavirus family, based on the availability of the proteins, 11 organisms are considered. Our model is created from the ontological relationship graphs by comparing the affinities of all potential GO pairings that may be annotated from any target protein pair. Finally, the score of interaction affinity of protein pair based on their annotated GO pair-wise interaction is computed within a range of [0, 1]. Table 1 gives a detailed description of the number of proteins available for the respective coronavirus organism and the number of possible host–pathogen interaction networks that can be generated for each organism.

Table 1. Detailed description of proteins and host–pathogen interaction for all organisms from the coronavirus family.

Organism	No. of Proteins	No. of Host–Pathogen Interaction
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	14	205,140
Severe acute respiratory syndrome coronavirus (SARS-CoV)	15	233,411
Bat coronavirus HKU3	12	125,904
Bat coronavirus Rp3/2004	13	125,904
Murine coronavirus	40	425,162
Middle East respiratory syndrome-related coronavirus (MERS-CoV)	10	174,136
Bovine coronavirus	94	688,115
Bat coronavirus HKU5	10	117,090
Rat coronavirus	12	92,508
Bat coronavirus HKU4	10	117,090
Bat coronavirus 133/2005	10	98,494

3.2. Detailed Description of Human–nCoV Protein Interaction Network

The 2019 coronavirus disease pandemic was brought on by the novel coronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2/nCoV). It affected over 12 million people and caused over 560,000 fatalities in 213 nations [105]. To infect a host, the nCoV protein, like other virus proteins, must interact with the host protein and replicate the genome. Detailed descriptions for all types of possible interactions are given in Table 2. At the time of our experiment, UniProt [106] holds around 19,283 human proteins and 16 nCoV proteins (Table 3) having GO annotations. Here, through our proposed in-silico model, we compute all the possible protein interactions between human-nCoV for all the proteins having GO annotations (Table 4). Here 'Total Dataset' refers to the total number of possible interactions generated from the in-silico model. This includes; Human-Human interactions, Human-nCoV interactions, and nCoV-nCoV interactions.

Table 2. Detailed statistics of Human–nCoV protein interactions computed by our proposed model.

Intersection Type	Organism	Proteins	Interactions
All	Total Dataset	19,297	164,701,415
Host–Pathogen	Human–nCoV	19,297	206,516
Pathogen—Pathogen	nCoV–nCoV	14	83
Host–Host	Human–Human	19,283	164,494,816

Table 3. Details of nCoV proteins collected from UniProt [106].

Entry	Entry Name	Gene Names	Protein Names
P0DTD1	R1AB_SARS2	rep 1a–1b	Replicase polyprotein 1ab, pp1ab (ORF1ab polyprotein)
P0DTC1	R1A_SARS2		Replicase polyprotein 1a, pp1a (ORF1a polyprotein)
P0DTC2	SPIKE_SARS2	S 2	Spike glycoprotein, S glycoprotein (E2) (Peplomer protein)
P0DTD8	NS7B_SARS2	7b	ORF7b protein, ORF7b (Accessory protein 7b)
P0DTC6	NS6_SARS2	6	ORF6 protein, ORF6 (Accessory protein 6)
P0DTC8	NS8_SARS2	8	ORF8 protein, ORF8 (Non-structural protein 8, ns8)
P0DTF1	ORF3B_SARS2		Putative ORF3b protein, ORF3b
P0DTC5	VME1_SARS2	M	Membrane protein, M (E1 glycoprotein
P0DTD3	ORF9C_SARS2	9c	Putative ORF9c protein, ORF9c
P0DTC3	AP3A_SARS2	3a	ORF3a protein, ORF3a
P0DTG0	ORF3D_SARS2		Putative ORF3d protein
P0DTG1	ORF3C_SARS2		ORF3c protein, ORF3c (ORF3h protein, ORF3h)
P0DTC7	NS7A_SARS2	7a	ORF7a protein, ORF7a
P0DTD2	ORF9B_SARS2	9b	ORF9b protein, ORF9b
P0DTC9	NCAP_SARS2	N	Nucleoprotein, N (Nucleocapsid protein, NC, Protein N)
P0DTC4	VEMP_SARS2	E 4	Envelope small membrane protein, E, sM protein

Table 4. Details of Human–nCov Interactions at different threshold values.

Interaction Type	Organism	Threshold	Nodes	Edges	Human	nCoV
Host–Pathogen	Human–nCoV	0.2	109	592	10	12
		0.15	245	1174	128	13
		0.1	886	2909	768	13
		0.09	1193	3586	1075	13
		0.08	1754	4619	1636	13
		0.05	7397	16,209	7278	13
		0.02	15,551	74,560	15,431	13
		0.001	18,936	166,382	18,816	14

3.3. Validation through the State-of-the-Art Dataset

Gordon et al. [105] proposed a host–pathogen interaction dataset physically connected with the human cell by cloning, tagging, and expressing 27 out of 29 proteins using affinity-purification mass spectrometry. Up to 14 open-reading frames can be encoded by a 30-kb genome (ORFs). In order to create the 16 non-structural proteins (NSP1-NSP16) that make up the replicase transcriptase complex, ORF1a and ORF1ab encode polyproteins. This produces a dataset of 332 high-confidence host–pathogen protein–protein interaction networks. However, while validating our computational model, we discovered that the protein sequences provided by Gordon et al. do not have any mapping with the corresponding UniProt id. In our situation, we have exclusively focused on the SARS-CoV-2 proteins published on UniProt. We have used a mathematical model to determine the binding affinities of a portion of the evaluated human proteins listed on UniProt. Because SARS-CoV-2 proteins could not be directly mapped into corresponding UniProt accession ids, direct comparison and validation concerning Gordon et al. were impossible. Thus, the nCoV proteins from Gordon et al. were mapped to the corresponding UniProt ids. As our research heavily depends on the underlying GO network of the host–pathogen protein interaction network, those proteins are selected with all three GO annotations. To validate our proposed method, all possible interactions are computed in our proposed computational environment, which gives 57,615 possible interactions, which are their respective fuzzy score from 27 bait and 332 prey. Among these interactions, 129 existing host–pathogen from high confidence dataset proposed by Gordon et al. whose scores are calculated.

Apart from the high-confidence host–pathogen protein interaction network dataset, Gordon et al. also provided a host–pathogen interaction dataset that contains a human-nCoV protein interaction network without any threshold. This mainly contains scoring results of all bait and all prey proteins showing spectral counts of experimental samples. The dataset contains 22,153 interactions, including 27 bait and 2753 host proteins. Our proposed model generates an interaction network with the said protein, which generates all-vs-all interactions. Among those 22,153 interactions, there are 7866 existing host–pathogen interactions whose scores are calculated. Table 5 gives detailed information regarding the host–pathogen interaction for the high-confidence human–nCoV dataset and the generic human–nCoV dataset proposed by Gordon et al.

Table 5. Overall statistics for interaction affinity score of High confidence Human–nCov dataset and all Human–nCov Dataset proposed by Gordon et al. computed by our proposed model.

Dataset	No. of Interactions	No. of Bait	No. of Prey	Total Interaction Score Computed
High Confidence Host–Pathogen PPI	332	27	332	57,615
All Host–Pathogen PPI	22,153	27	2,753	2,156,507

3.3.1. Comparison with Gordon et al.

To validate our computational model, we compare our data set with that proposed by Gordon et al. [107]. To experiment with our proposed computational model, we construct a dataset of human and SARS-CoV-2/nCoV proteins retrieved from the UniProt protein repository, as discussed above. The computation results in fuzzy scoring of the protein pair (*viz.* human–human ppin, human–nCoV ppin, and nCoV–nCoV ppin). The edge-overlapping has shown the validation of our computational model between two datasets at different threshold values set on the fuzzy score. Edge overlapping signifies the common edges present in both datasets. For our experiment, we have kept the fuzzy score threshold ranging from 0.1–0.001. At first, we compare our network with the high-confidence human–nCoV network proposed by Gordon et al. The dataset contains 332 host proteins and 27 viral proteins. Table 6 compares two datasets at different threshold values and produces the intersected nodes and edges between the two datasets, along with the common host and viral proteins.

Table 6. Detailed validation of our model compared to High confidence human–nCoV proposed by Gordon et al.

HQ Data (Gordon et al.)		Our Dataset				
Number of Host	No. of Bait	Threshold	Number of Host	No. of Bait	No. of Intersected Nodes	No. of Intersected Edges
2753	27	0.1	17,875	13	88	149
2753	27	0.09	18,064	13	104	176
2753	27	0.08	18,218	13	128	214
2753	27	0.05	19,838	14	381	626
2753	27	0.02	19,123	14	1129	2513
2753	27	0.001	19,193	14	1817	6634

The high-confidence dataset and the other dataset proposed by Gordon et al., which contains scoring results of all bait and all prey proteins showing spectral counts of experimental samples, are also being compared in the same manner discussed above with varying threshold values imposed on fuzzy interaction affinity score. The threshold ranges from 0.1–0.001. The dataset proposed by Gordon et al. contains 2753 host proteins and 27 viral proteins. Table 7 represents the comparison between the two datasets at different threshold values and produces the intersected nodes and intersected edges between the two datasets.

Table 7. Detailed validation of our model compared to all Human–nCov Datasets proposed by Gordon et al.

HQ Data (Gordon et al.)		Our Dataset				
Number of Host	No. of Bait	Threshold	Number of Host	No. of Bait	No. of Intersected Nodes	No. of Intersected Edges
332	27	0.1	768	13	8	5
332	27	0.09	1075	13	8	5
332	27	0.08	1636	13	8	5
332	27	0.05	7278	13	20	14
332	27	0.02	15,431	13	60	51
332	27	0.001	18,816	14	109	99

3.3.2. Comparison with Dick et al.

Protein-protein Interaction Prediction Engine (PIPE) is a sequence-based PPI prediction approach that looks at sequence windows on each query protein proposed by Dick et al. [108]. The evidence for the putative PPI is strengthened if the two sequence windows have a lot in common with other pairs of proteins that have been found to interact. Normalization is used in a similarity-weighted (SW) scoring system to consider common sequences unrelated to PPIs. A PPI is anticipated, given enough supporting data [109–111]. For understudied species, the Protein-protein Interaction Prediction Engine (PIPE4) iteration has recently been modified [112].

Like PIPE, the SPRINT predictor gathers data from previously reported PPI interactions based on window similarity with the query protein pair to determine its prediction scores [113]. SPRINT uses a spaced seed method to compare the sequences of protein windows, where only certain places in the two windows must match, as determined by the bits of the spaced seeds. Additionally, because proteins are encoded with five bits per amino acid, it is possible to quickly compute protein window similarities and, consequently, forecast scores using very efficient (SIMD) bitwise operations [113].

Here, the two datasets produced by Dick et al. [108] are being compared, and an interaction affinity pair is being generated by using our proposed method. Table 8 shows the details of the comparison with both datasets. The table shows that PIPE4 contains 702 interactions, among which our proposed model identifies 575 interactions, and the score has been generated. On the other hand, the SPRINT dataset contains 510 interactions, among which 413 are identified by our proposed method.

Table 8. Detailed validation of our model compared to all Human–nCov Datasets proposed by Dick et al.

Dataset (Dick et al.)	No. of Interactions	No. of Bait	No. of Prey	Total Interaction Score Computed
PIPE4	702	13	518	575
SPRINT	510	15	368	413

3.4. Vulnerable Host Protein

One of the main focuses of our research is to identify the common vulnerable host proteins at different threshold values. As discussed in Section 3.1, our computational model efficiently computes the interaction affinity and can generate a fuzzy score for any host–pathogen interaction pair for any organism from the corona family. We have experimented with the host–pathogen network for the entire corona family (with the selected organism, as mentioned in Section 2.2) and retrieved the network at different threshold values ranging from 0.1–0.001 at each threshold score, we segregate the network for each covid organism and construct their respective networks. Thus, for each threshold score, we obtained a separate host–pathogen network for each coronavirus organism. So, for each threshold score, some common host protein interacts with all the coronavirus organisms. As the value of the score decreases from a high threshold to a low threshold value, the number of common host proteins increases. These host proteins are the level one spreader nodes.

These spreader nodes are identified by fuzzy thresholding, and these host proteins are vulnerable to the propagation or contamination of the diseases caused by the viral proteins. Table 9 represents the number of vulnerable host proteins at different fuzzy threshold scores. Figures 2 and 3 represent the Venn diagram of the vulnerable host proteins at 0.1 and 0.001 threshold values, respectively. For simplicity and ease of the process, we divide the viral organism into three subsets. SARS-CoV-2, SARS-CoV and MERS-CoV forms one group, all the different organism from BAT-CoV (viz., *Bat coronavirus* HKU3, *Bat coronavirus* Rp3/2004, *Bat coronavirus* HKU5, *Bat coronavirus* HKU4, *Bat coronavirus* 133/2005) forms one group, and Murine-CoV, Bovine-CoV and Rat Coronavirus forms the third group. Then we identified the common host proteins from all three groups separately. Intersected host protein sets from all three groups are identified and again intersected. This results in the common vulnerable host proteins at the specified threshold value. For visualization, we only arbitrarily select a threshold value of 0.1 for constructing the Venn diagram, 0.1 threshold value gives 191 vulnerable host proteins interacting with all selected coronavirus organisms.

Table 9. Number of Vulnerable host proteins identified from the host–pathogen network for all selected coronavirus organisms at a different fuzzy threshold score.

Threshold	No. of Vulnerable Human Proteins
0.001	14,297
0.005	11,208
0.03	3889
0.05	526
0.07	351
0.1	191

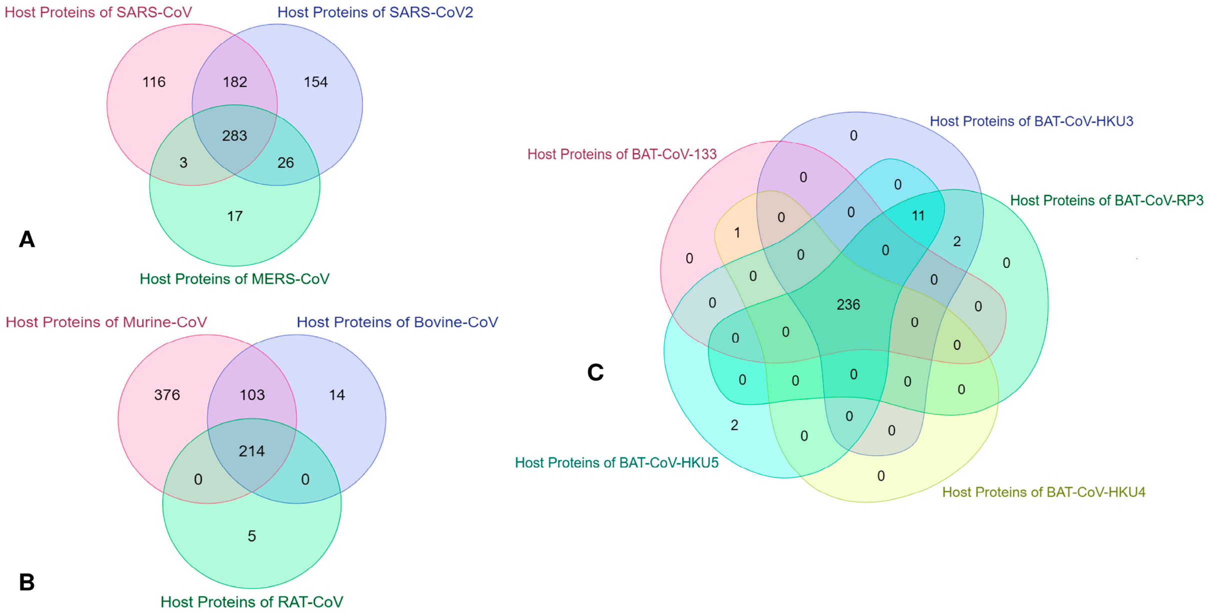


Figure 2. Venn diagram of the number of vulnerable host proteins obtained from host–pathogen interaction for all selected coronavirus organisms at 0.1 fuzzy threshold value. (**A**). The intersection of host protein identified from SARS-CoV-2, SARS-CoV, and MER-CoV. (**B**). Intersected host proteins from Murine-CoV, Bovine-CoV, and Rat Coronavirus. (**C**). Intersected host proteins of different viral organisms of Bat Coronavirus.

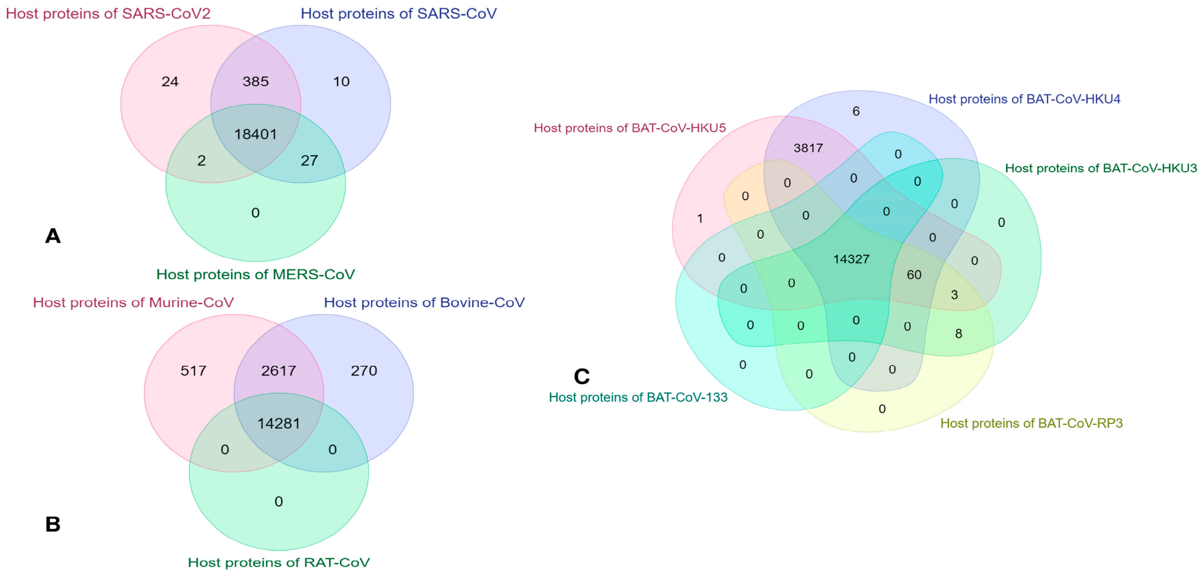


Figure 3. Venn diagram of the number of vulnerable host proteins obtained from host–pathogen interaction for all selected coronavirus organisms at 0.001 fuzzy threshold value. (**A**). Intersection of host protein identified from SARS-CoV-2, SARS-CoV, and MER-CoV. (**B**). The intersected host proteins from Murine-CoV, Bovine-CoV, and Rat Coronavirus. (**C**). Intersected host proteins from different viral organisms of Bat Coronavirus.

3.5. Identification of Potential Candidate FDA Drugs concerning Vulnerable Host-Proteins Using Human–Coronavirus Family Interaction Network Analysis

All level one human proteins of the coronavirus family are mapped with their matching medicines from DrugBank once the coronavirus family–human PIN has been created [114]. DrugBank is an online database that offers extensive information on medicines, drug-protein targets, and drug metabolism [115]. Most in-silico approaches used in drug design, drug docking, and drug interaction prediction use DrugBank as their most frequently used database because of its high-quality annotation.

It has around 60% of FDA-approved medications and 10% of investigational drugs. It has been determined through adequate analysis that some spreader nodes in COVID19-human PPIN are the protein targets of possible COVID-19 FDA-listed medicines [116]: hydroxychloroquine [117], azithromycin [117], lopinavir [118], remdesivir [119,120], etc. Not only the list of drugs for COVID-19, but we have obtained a list of FDA-approved drugs from level 1 vulnerable host proteins for the entire coronavirus family by using Drug Consensus Score algorithm (DCS). The algorithm is defined as the number of times a drug occurs at a specific PPIN level. Each human protein is mapped with the appropriate related medicines in this level 1 PPIN.

The DCS, or frequency of each drug, is therefore calculated. Table 9 represents the top-5 FDA-approved drug at different fuzzy threshold values and the number of vulnerable host proteins at that corresponding threshold value, Drug ID, and corresponding DCS score for each drug. Fostamatinib is thought to be a promising medication for the target nCoV protein in the randomly created COVID-19 human PPI since it has the highest DCS in most cases.

4. Discussion

The number of vulnerable host proteins at different threshold values is represented in Table 10, and the list of the top five drugs, along with their drug-id based on the DCS score, are listed. This leads us to the analysis with the application of the lowest threshold values (i.e., 0.001), based on which the possible repurposed drugs are proposed.

Table 10. Top 5 target drugs with their respective DCS score at different threshold value.

Threshold	Vulnerable Human Proteins	Drug ID	DCS Score	Drug Name
0.001	14,297	DB12010 DB09130 DB14533 DB14487 DB01593	181 47 45 45 45	Fostamatinib Copper Zinc chloride Zinc acetate Zinc
0.005	11,208	DB12010 DB01069 DB01593 DB09130 DB14487	173 45 39 39 39	Fostamatinib Promethazine Zinc Copper Zinc acetate
0.03	3889	DB12010 DB09130 DB04464 DB14487 DB11638	25 6 5 5 5	Fostamatinib Copper N-Formylmethionine Zinc acetate Artenimol
0.05	526	DB12010 DB12267 DB00041 DB00074 DB09130	7 2 2 2 2	Fostamatinib Brigatinib Aldesleukin Basiliximab Copper
0.07	351	DB00041 DB12010 DB11638 DB00004 DB02240	2 2 2 2 1	Aldesleukin Fostamatinib Artenimol Denileukin diftitox Quinacrine mustard
0.1	191	DB12267 DB00111 DB11942 DB08804 DB00047	1 1 1 1 1	Brigatinib Daclizumab Selinexor Nandrolone decanoate Insulin glargine

Drug repurposing is a powerful strategy that gives new therapeutic alternatives by identifying other uses for already-approved medications, as vaccine and drug development can take years [121]. The traditional conservative drug development approach, which is restricted to "one drug, one target" paradigms, does not take into account or assess the off-target effects or the likelihood of numerous drug indications, even though some of them have since been confirmed to exist [122]. Upon the formation of the coronavirus–human PPIN, all level one Coronavirus human proteins are mapped with the appropriate medications via DrugBank [114]. DrugBank is an online database that provides detailed information on pharmaceuticals, drug-protein targets, and drug metabolism. DrugBank is the most often utilized database in practically all in silico approaches used in drug design, drug docking, and drug interaction prediction because of the high-quality annotation in the database. It includes 10% and 60% of FDA-approved and investigational medications [114]. It is observed that the above list of drugs at the threshold value 0.001, listed in Table 9, when compared to the remaining human protein-associated medications, fostamatinib has the highest frequency of occurrence in the entire PPIN and has a sizable overlap of target proteins in the human–coronavirus PPIN with highest Drug Consensus Score of 181. It was already discussed and proposed in [115] that Fostamatinib has the highest DCS score with reference to level one and level two human spreader proteins. Thus, our drug of concern shifted to the one with the next highest score, copper. Copper has an enormous effect in defeating COVID-19, which helps it to dominate with a high DCS score. The study proposed in [120] aims to investigate the effects of a highly specialized drug, "Hinokitiol Copper Chelate", on enormous quantities of 2019-nCoV Spike Glycoprotein with a single receptor binding domain. This investigation offers a superior version of Hinokitiol Copper Chelate for in vitro testing against 2019-nCoV Main Protease. The authors suggest combining copper, NAC, colchicine, NO, and the experimental antivirals remdesivir or EIDD-2801 as a potential treatment for SARS-COV-2 [123]. In-silico docking study of copper complexes with SARS-CoV-2 viruses shows a steady binding with SARS-CoV-2 main protease (M^{pro}) active-site region [124].

Zinc supplements also play a crucial role in combating different organisms of coronavirus. The essentiality of Zinc lies in the preservation of natural tissue barriers such as the respiratory epithelium, preventing pathogen entry for a balanced functioning of the human immune system. The deficiency of Zinc can probably lead to the infection and detrimental progression of COVID-19 [125]. The body's tissue barriers, which contain cilia, mucus, anti-microbial peptides like lysozymes, and interferons, stop infectious organisms from entering. The primary mechanisms for SARS-CoV-2 entering cells are the cellular protease TMPRSS2 and the angiotensin-converting enzyme 2 (ACE2) [126]. People with COVID-19 are accompanied by ciliated epithelium destruction and ciliary dyskinesia, which limit mucociliary clearance [127]. The quantity and length of bronchial cilia increased after Zinc supplementation in Zinc-deficient rats [128].

In COVID-19, Zinc supplementation was hypothesized to reduce mortality. Supplementing with Zinc had no positive effects on how the illness progressed. The Zinc-supplemented group's hospital stay was lengthier. There is no evidence to back up regular Zinc supplementation in COVID-19 [129]. The confounding variables impacting Zinc's bioavailability may be avoided by administering Zinc intravenously, enabling Zinc to fulfill its medicinal potential. If effective, intravenous Zinc might be quickly incorporated into clinical practice due to benefits such as lack of toxicity, cheap cost, and accessibility of supply [130].

Promethazine, an antipsychotic agent showing clathrin-mediated endocytosis, is one most effective drugs for SARS-CoV and MERS-CoV, which has been repurposed for the treatment of COVID-19 as there is almost 89% genetic similarity with SARS-CoV-2 and SARS-CoV [131]. Two pills were offered as an intervention, one with Aspirin and Promethazine and the other with vitamins D3, C, and B3, together with Zinc and selenium supplements [132]. A randomized clinical trial has been conducted to recover mildly to moderate COVID-19 patients.

Based on this validation, further research on the repurposed drug, docking study, and other symptomatic analyses will help to identify the potential drug for the entire coronavirus family. A clinical study on Promethazine and Fostamatinib [115,132] is also in progress. Even though the research is in its early stages, it in some way partially corroborates our findings.

5. Conclusions

Finding spreader nodes in any network of host–pathogen interactions is essential for predicting the course of a disease. However, not every protein in a network of interactions is highly capable of transmitting illness. In this work, we used the host–pathogen protein interaction network between humans and different coronavirus family organisms. Based on the available GO annotations of the proteins, a fuzzy interaction affinity score has been proposed for all the host–pathogen interactions. Our proposed model was validated with the *state-of-the-art* dataset. It has been noticed from this assessment that the chosen human spreader nodes, indicated by our suggested model, emerge as the possible protein targets for the different organisms of coronavirus medications authorized by the FDA, which highlights the significance of this proposed work.

The basic hypothesis of the work is listed as follows: (1) Between SARS-CoV and SARS-CoV-2, there is a genetic overlap of around 89%, which also results in a substantial overlap in spreader proteins between human–SARS-COV and human–SARS-COV2 protein-interaction networks [79]. Moreover, we have considered the viral proteins of 11 different coronavirus organisms based on the available GO notations. (2) A fuzzy scoring approach for finding a protein's interaction affinity with another protein helped build the host–pathogen network. (3) The proposed in-silico can effectively identify the host–pathogen protein–protein interaction network for identifying potential candidate FDA drugs concerning vulnerable host–proteins.

Our proposed in-silico method for identifying host–pathogen protein interaction networks has been validated through different *state-of-the-art* datasets. According to recent research by Gordon et al., who focused on the sequence analysis of SARS-CoV-2 isolates, 332 high-confidence SARS-CoV-2–human protein–protein interactions have been discovered. Using affinity-purification mass spectrometry, they determined the human proteins that were physically linked to each of the 26 of the 29 SARS-CoV-2 proteins after they had been cloned, tagged, and produced in human cells [107]. While validating our work with Gordon et al., we discovered that the SARS-CoV-2 protein sequences employed by Gordon et al. do not exactly correspond to the accessible UniProt accession ids when comparing their foundational work with ours. In our situation, we exclusively focused on the SARS-CoV-2 proteins published on UniProt. We used a mathematical model to analyze the binding affinities of a subset of the human proteins available on UniProt. Because SARS-CoV-2 proteins could not be directly mapped into matching UniProt accession ids, direct comparison and validation concerning Gordon et al. were impossible. However, using the COVID-19 UniProtKB reference database, an attempt has been made to map the UniProt ids of Gordon et al. SARS-CoV-2 proteins [120].

In addition, our approach is not directly deal with the classification problem and does not require prior knowledge of positive and negative interaction. Further, several experiments show that Gordon et al. do not detect all the significant human–nCoV interactions [133,134]. For example, the essential protein for entry into the human host, ACE2 and TMPRSS2, are surprisingly not found in Gordon et al. However, in most of the covid related studies, Gordon et al. are considered one of the gold standards in human–nCoV interactions. When we quantitatively compared our findings with Gordon et al., we primarily focused on estimating TPR (higher is better) and FNR (lower is better) over node and edge overlaps between the two networks using multiple fuzzy thresholds. In this assessment, we observed that the optimal TPR (0.71) and FNR (0.29) are obtained around the fuzzy threshold 0.01 for node intersections while comparing with Gordon et al. Likewise, optimal TPR (0.86) and FNR (0.14) for edge intersection are observed at 0.001.

The target proteins of the possible FDA medications for the coronavirus family coincide with the spreader nodes of the hypothesized human–coronavirus protein interaction network, which may highlight one of the study's major findings. Based on the DCS score applied on vulnerable host proteins identified at different threshold values, we have proposed a list of FDA-approved drugs such as Fostamatinib, Copper, Zinc Acetate, Zinc Chloride, etc. Our previous research has proposed Fostamatinib as a potential drug for COVID-19. This analysis demonstrates that these spreader nodes have biological importance in transmitting illness. Additionally, it spurs us to do medication repurposing research which focuses on the fact that apart from Fostamatinib, Promethazine can also be one of the potential drug candidates for coronavirus-related diseases under clinical trials. In a nutshell, the proposed methodology forms a complete PPIN for humans and different coronavirus organisms and adds much more relevant biological information about existing drugs against SARS-CoV-2 through a drug-repurposing study done with proper assessment and in-depth computational study.

Author Contributions: Conceptualization, S.S.B., A.K.H., S.S. and M.N.; data curation, S.S.B., A.K.H. and S.B.; formal analysis, S.S.B., A.K.H., P.C. and S.B.; investigation, P.C.; methodology, S.S.B., A.K.H. and S.S.; project administration, M.N. and S.B.; Resources, S.S.B., A.K.H., S.S. and M.N.; software, S.S.B., A.K.H. and S.S.; supervision, P.C., M.N. and S.B.; validation, S.S.B. and A.K.H.; visualization, S.S.; writing—original draft, S.S.B., A.K.H. and S.B.; writing—review and editing, S.S., P.C., M.N. and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available at the following GitHub link: https://github.com/SovanSaha/Assessment-of-GO-based-protein-interaction-affinities-in-the-3-large-scale-human-coronavirus-family.git (accessed on 1 February 2023) for free academic use.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guarner, J. Three emerging coronaviruses in two decades: The story of SARS, MERS, and now COVID-19. *Am. J. Clin. Pathol.* **2020**, *153*, 420–421. [CrossRef] [PubMed]
2. Wang, C.; Horby, P.W.; Hayden, F.G.; Gao, G.F. A novel coronavirus outbreak of global health concern. *Lancet* **2020**, *395*, 470–473. [CrossRef] [PubMed]
3. World Health Organization. *Statement on the Second Meeting of the International Health Regulations (2005) Emergency Committee Regarding the Outbreak of Novel Coronavirus (2019-nCoV)*; World Health Organization: Geneva, Switzerland, 2020.
4. Ruan, S. Likelihood of survival of coronavirus disease 2019. *Lancet Infect. Dis.* **2020**, *20*, 630–631. [CrossRef] [PubMed]
5. Chen, Y.; Liu, Q.; Guo, D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. *J. Med. Virol.* **2020**, *92*, 418–423. [CrossRef]
6. Zhong, J.; Tang, C.; Peng, W.; Xie, M.; Sun, Y.; Tang, Q.; Xiao, Q.; Yang, J. A novel essential protein identification method based on PPI networks and gene expression data. *BMC Bioinform.* **2021**, *22*, 248. [CrossRef] [PubMed]
7. He, X.; Kuang, L.; Chen, Z.; Tan, Y.; Wang, L. Method for identifying essential proteins by key features of proteins in a novel protein-domain network. *Front. Genet.* **2021**, *12*, 708162. [CrossRef] [PubMed]
8. Saha, S.; Prasad, A.; Chatterjee, P.; Basu, S.; Nasipuri, M. Modified FPred-Apriori: Improving function prediction of target proteins from essential neighbours by finding their association with relevant functional groups using Apriori algorithm. *Int. J. Adv. Intell. Paradig.* **2021**, *19*, 61–83. [CrossRef]
9. Sengupta, K.; Saha, S.; Halder, A.K.; Chatterjee, P.; Nasipuri, M.; Basu, S.; Plewczynski, D. PFP-GO: Integrating protein sequence, domain and protein-protein interaction information for protein function prediction using ranked GO terms. *Front. Genet.* **2022**, *13*, 969915. [CrossRef] [PubMed]
10. Saha, S.; Chatterjee, P.; Halder, A.K.; Nasipuri, M.; Basu, S.; Plewczynski, D. ML-DTD: Machine Learning-Based Drug Target Discovery for the Potential Treatment of COVID-19. *Vaccines* **2022**, *10*, 1643. [CrossRef]
11. Banik, A.; Podder, S.; Saha, S.; Chatterjee, P.; Halder, A.K.; Nasipuri, M.; Basu, S.; Plewczynski, D. Rule-Based Pruning and In Silico Identification of Essential Proteins in Yeast PPIN. *Cells* **2022**, *11*, 2648. [CrossRef]
12. Saha, S.; Sengupta, K.; Chatterjee, P.; Basu, S.; Nasipuri, M. Analysis of protein targets in pathogen–host interac-tion in infectious diseases: A case study on *Plasmodium falciparum* and *Homo sapiens* interaction network. *Brief. Funct. Genom.* **2018**, *17*, 441–450.
13. Saha, S.; Chatterjee, P.; Nasipuri, M.; Basu, S. Detection of spreader nodes in human-SARS-CoV protein-protein interaction network. *PeerJ* **2021**, *9*, e12117. [CrossRef]
14. Basak, S.N.; Biswas, A.K.; Saha, S.; Chatterjee, P.; Basu, S.; Nasipuri, M. Target Protein Function Prediction by Identification of Essential Proteins in Protein-Protein Interaction Network. In Proceedings of the International Conference on Computational Intelligence, Communications, and Business Analytics, Kalyani, India, 27–28 July 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 219–231.
15. Saha, S.; Chatterjee, P.; Basu, S.; Nasipuri, M.; Plewczynski, D. FunPred 3.0: Improved protein function prediction using protein interaction network. *PeerJ* **2019**, *7*, e6830. [CrossRef] [PubMed]
16. Saha, S.; Chatterjee, P.; Basu, S.; Kundu, M.; Nasipuri, M. FunPred-1: Protein function prediction from a protein interaction network using neighborhood analysis. *Cell. Mol. Biol. Lett.* **2014**, *19*, 675–691. [CrossRef] [PubMed]
17. Prasad, A.; Saha, S.; Chatterjee, P.; Basu, S.; Nasipuri, M. Protein function prediction from protein interaction network using bottom-up L2L apriori algorithm. In Proceedings of the International Conference on Computational Intelligence, Communications, and Business Analytics, Kolkata, India, 24–25 March 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 3–16.
18. Saha, S.; Prasad, A.; Chatterjee, P.; Basu, S.; Nasipuri, M. Protein function prediction from dynamic protein interaction network using gene expression data. *J. Bioinform. Comput. Biol.* **2019**, *17*, 1950025. [CrossRef] [PubMed]
19. Saha, S.; Prasad, A.; Chatterjee, P.; Basu, S.; Nasipuri, M. Protein function prediction from protein–protein interaction network using gene ontology based neighborhood analysis and physico-chemical features. *J. Bioinform. Comput. Biol.* **2018**, *16*, 1850025. [CrossRef] [PubMed]
20. Kann, M.G. Protein interactions and disease: Computational approaches to uncover the etiology of diseases. *Brief. Bioinform.* **2007**, *8*, 333–346. [CrossRef]
21. Schnirring, L. China Releases Genetic Data on New Coronavirus, Now Deadly. Center for Infectious Disease Research and Policy. Available online: https://www.cidrap.umn.edu/covid-19/china-releases-genetic-data-new-coronavirus-now-deadly (accessed on 1 January 2022).
22. Chan, J.F.-W.; Kok, K.-H.; Zhu, Z.; Chu, H.; To, K.K.-W.; Yuan, S.; Yuen, K.-Y. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* **2020**, *9*, 221–236. [CrossRef]

23. Xu, H.F.; Wang, M.; Zhang, Z.B.; Zou, X.Z.; Gao, Y.; Liu, X.N.; Lu, E.J.; Pan, B.Y.; Wu, S.J.; Yu, S.Y. An epidemiologic investigation on infection with severe acute respiratory syndrome coronavirus in wild animals traders in Guangzhou. *Zhonghua Yu Fang Yi Xue Za Zhi* **2004**, *38*, 81–83.
24. Xu, R.-H.; He, J.-F.; Evans, M.R.; Peng, G.-W.; Field, H.E.; Yu, D.-W.; Lee, C.-K.; Luo, H.-M.; Lin, W.-S.; Lin, P.; et al. Epidemiologic clues to SARS origin in China. *Emerg. Infect. Dis.* **2004**, *10*, 1030. [CrossRef]
25. World Health Organization. Summary of Probable SARS Cases with Onset of Illness from 1 November 2002 to 31 July 2003. Available online: http//www.who.int/csr/sars/country/table2004_04_21/en/index.html (accessed on 1 January 2022).
26. Marra, M.A.; Jones, S.J.M.; Astell, C.R.; Holt, R.A.; Brooks-Wilson, A.; Butterfield, Y.S.N.; Khattra, J.; Asano, J.K.; Barber, S.A.; Chan, S.Y.; et al. The genome sequence of the SARS-associated coronavirus. *Science* **2003**, *300*, 1399–1404. [CrossRef]
27. Rota, P.A.; Oberste, M.S.; Monroe, S.S.; Nix, W.A.; Campagnoli, R.; Icenogle, J.P.; Penaranda, S.; Bankamp, B.; Maher, K.; Chen, M.; et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **2003**, *300*, 1394–1399. [CrossRef]
28. Zhong, N.S.; Zheng, B.J.; Li, Y.M.; Poon, L.L.M.; Xie, Z.H.; Chan, K.H.; Li, P.H.; Tan, S.Y.; Chang, Q.; Xie, J.P.; et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet* **2003**, *362*, 1353–1358. [CrossRef]
29. Shi, Z.; Hu, Z. A review of studies on animal reservoirs of the SARS coronavirus. *Virus Res.* **2008**, *133*, 74–87. [CrossRef]
30. Mackay, I.M.; Arden, K.E. MERS coronavirus: Diagnostics, epidemiology and transmission. *Virol. J.* **2015**, *12*, 222. [CrossRef]
31. Zaki, A.M.; Van Boheemen, S.; Bestebroer, T.M.; Osterhaus, A.D.M.E.; Fouchier, R.A.M. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med.* **2012**, *367*, 1814–1820. [CrossRef]
32. Azhar, E.I.; Hui, D.S.C.; Memish, Z.A.; Drosten, C.; Zumla, A. The middle east respiratory syndrome (MERS). *Infect. Dis. Clin.* **2019**, *33*, 891–905. [CrossRef] [PubMed]
33. Grabherr, S.; Ludewig, B.; Pikor, N.B. Insights into coronavirus immunity taught by the murine coronavirus. *Eur. J. Immunol.* **2021**, *51*, 1062–1070. [CrossRef] [PubMed]
34. Weiss, S.R.; Navas-Martin, S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 635–664. [CrossRef] [PubMed]
35. Bender, S.J.; Weiss, S.R. Pathogenesis of murine coronavirus in the central nervous system. *J. Neuroimmune Pharmacol.* **2010**, *5*, 336–354. [CrossRef] [PubMed]
36. Leibowitz, J.L.; Srinivasa, R.; Williamson, S.T.; Chua, M.M.; Liu, M.; Wu, S.; Kang, H.; Ma, X.-Z.; Zhang, J.; Shalev, I.; et al. Genetic determinants of mouse hepatitis virus strain 1 pneumovirulence. *J. Virol.* **2010**, *84*, 9278–9291. [CrossRef] [PubMed]
37. Gorbalenya, A.E.; Snijder, E.J.; Spaan, W.J.M. Severe acute respiratory syndrome coronavirus phylogeny: Toward consensus. *J. Virol.* **2004**, *78*, 7863–7866. [CrossRef] [PubMed]
38. Vlasova, A.N.; Saif, L.J. Bovine coronavirus and the associated diseases. *Front. Vet. Sci.* **2021**, *8*, 643220. [CrossRef]
39. Zhang, X.M.; Herbst, W.; Kousoulas, K.G.; Storz, J. Biological and genetic characterization of a hemagglutinating coronavirus isolated from a diarrhoeic child. *J. Med. Virol.* **1994**, *44*, 152–161. [CrossRef] [PubMed]
40. Alekseev, K.P.; Vlasova, A.N.; Jung, K.; Hasoksuz, M.; Zhang, X.; Halpin, R.; Wang, S.; Ghedin, E.; Spiro, D.; Saif, L.J. Bovine-like coronaviruses isolated from four species of captive wild ruminants are homologous to bovine coronaviruses, based on complete genomic sequences. *J. Virol.* **2008**, *82*, 12422–12431. [CrossRef] [PubMed]
41. Lau, S.K.P.; Lee, P.; Tsang, A.K.L.; Yip, C.C.Y.; Tse, H.; Lee, R.A.; So, L.-Y.; Lau, Y.-L.; Chan, K.-H.; Woo, P.C.Y.; et al. Molecular epidemiology of human coronavirus OC43 reveals evolution of different genotypes over time and recent emergence of a novel genotype due to natural recombination. *J. Virol.* **2011**, *85*, 11325–11337. [CrossRef] [PubMed]
42. Lau, S.K.P.; Woo, P.C.Y.; Yip, C.C.Y.; Fan, R.Y.Y.; Huang, Y.; Wang, M.; Guo, R.; Lam, C.S.F.; Tsang, A.K.L.; Lai, K.K.Y.; et al. Isolation and characterization of a novel Betacoronavirus subgroup A coronavirus, rabbit coronavirus HKU14, from domestic rabbits. *J. Virol.* **2012**, *86*, 5481–5496. [CrossRef] [PubMed]
43. Li, W.; Shi, Z.; Yu, M.; Ren, W.; Smith, C.; Epstein, J.H.; Wang, H.; Crameri, G.; Hu, Z.; Zhang, H.; et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science* **2005**, *310*, 676–679. [CrossRef]
44. Lau, S.K.P.; Woo, P.C.Y.; Li, K.S.M.; Huang, Y.; Tsoi, H.-W.; Wong, B.H.L.; Wong, S.S.Y.; Leung, S.-Y.; Chan, K.-H.; Yuen, K.-Y. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 14040–14045. [CrossRef]
45. Yuan, J.; Hon, C.-C.; Li, Y.; Wang, D.; Xu, G.; Zhang, H.; Zhou, P.; Poon, L.L.M.; Lam, T.T.-Y.; Leung, F.C.-C.; et al. Intraspecies diversity of SARS-like coronaviruses in Rhinolophus sinicus and its implications for the origin of SARS coronaviruses in humans. *J. Gen. Virol.* **2010**, *91*, 1058–1062. [CrossRef]
46. Ren, W.; Li, W.; Yu, M.; Hao, P.; Zhang, Y.; Zhou, P.; Zhang, S.; Zhao, G.; Zhong, Y.; Wang, S.; et al. Full-length genome sequences of two SARS-like coronaviruses in horseshoe bats and genetic variation analysis. *J. Gen. Virol.* **2006**, *87*, 3355–3359. [CrossRef] [PubMed]
47. Quan, P.-L.; Firth, C.; Street, C.; Henriquez, J.A.; Petrosov, A.; Tashmukhamedova, A.; Hutchison, S.K.; Egholm, M.; Osinubi, M.O.V.; Niezgoda, M.; et al. Identification of a severe acute respiratory syndrome coronavirus-like virus in a leaf-nosed bat in Nigeria. *MBio* **2010**, *1*, e00208-10. [CrossRef] [PubMed]

48. Lau, S.K.P.; Li, K.S.M.; Tsang, A.K.L.; Lam, C.S.F.; Ahmed, S.; Chen, H.; Chan, K.-H.; Woo, P.C.Y.; Yuen, K.-Y. Genetic characterization of Betacoronavirus lineage C viruses in bats reveals marked sequence divergence in the spike protein of *pipistrellus* bat coronavirus HKU5 in Japanese pipistrelle: Implications for the origin of the novel Middle East respiratory sy. *J. Virol.* **2013**, *87*, 8638–8650. [CrossRef] [PubMed]
49. Raj, V.S.; Mou, H.; Smits, S.L.; Dekkers, D.H.W.; Müller, M.A.; Dijkman, R.; Muth, D.; Demmers, J.A.A.; Zaki, A.; Fouchier, R.A.M.; et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. *Nature* **2013**, *495*, 251–254. [CrossRef]
50. Lu, G.; Liu, D. SARS-like virus in the Middle East: A truly bat-related coronavirus causing human diseases. *Protein Cell* **2012**, *3*, 803. [CrossRef]
51. Guzzi, P.H.; Mina, M.; Guerra, C.; Cannataro, M. Semantic similarity analysis of protein data: Assessment with biological features and issues. *Brief. Bioinform.* **2011**, *13*, 569–585. [CrossRef]
52. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. *arXiv* **1995**, arXiv:cmp-lg/9511007.
53. Lin, D. An information-theoretic definition of similarity. In Proceedings of the Icml; Citeseer: State College, PA, USA, 1998; Volume 98, pp. 296–304.
54. Song, X.; Li, L.; Srimani, P.K.; Philip, S.Y.; Wang, J.Z. Measure the semantic similarity of GO terms using aggregate information content. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2013**, *11*, 468–476. [CrossRef]
55. Jiang, J.J.; Conrath, D.W. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv* **1997**, arXiv:cmp-lg/9709008.
56. Schlicker, A.; Domingues, F.S.; Rahnenführer, J.; Lengauer, T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinform.* **2006**, *7*, 302. [CrossRef]
57. Shannon, C.E. A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **2001**, *5*, 3–55. [CrossRef]
58. Couto, F.M.; Silva, M.J.; Coutinho, P.M. Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen Germany, 31 October–5 November 2005; ACM: New York, NY, USA, 2005; pp. 343–344.
59. Couto, F.M.; Silva, M.J.; Coutinho, P.M. Measuring semantic similarity between Gene Ontology terms. *Data Knowl. Eng.* **2007**, *61*, 137–152. [CrossRef]
60. Couto, F.M.; Silva, M.J. Disjunctive shared information between ontology concepts: Application to Gene Ontology. *J. Biomed. Semant.* **2011**, *2*, 5. [CrossRef] [PubMed]
61. Pesquita, C.; Faria, D.; Bastos, H.; Ferreira, A.E.N.; Falcão, A.O.; Couto, F.M. Metrics for GO based protein semantic similarity: A systematic evaluation. In *Proceedings of the BMC Bioinformatics*; BioMed Central: London, UK, 2008; Volume 9, p. 4.
62. Benabderrahmane, S.; Smail-Tabbone, M.; Poch, O.; Napoli, A.; Devignes, M.-D. IntelliGO: A new vector-based semantic similarity measure including annotation origin. *BMC Bioinform.* **2010**, *11*, 588. [CrossRef] [PubMed]
63. Wang, J.Z.; Du, Z.; Payattakool, R.; Yu, P.S.; Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **2007**, *23*, 1274–1281. [CrossRef]
64. Dutta, P.; Basu, S.; Kundu, M. Assessment of Semantic Similarity between Proteins Using Information Content and Topological Properties of the Gene Ontology Graph. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *15*, 839–849. [CrossRef]
65. Dutta, P.; Halder, A.K.; Basu, S.; Kundu, M. A survey on Ebola genome and current trends in computational research on the Ebola virus. *Brief. Funct. Genom.* **2017**, *17*, 374–380. [CrossRef]
66. Halder, A.K.; Dutta, P.; Kundu, M.; Basu, S.; Nasipuri, M. Review of computational methods for virus–host protein interaction prediction: A case study on novel Ebola–human interactions. *Brief. Funct. Genom.* **2017**, *17*, 381–391. [CrossRef]
67. Halder, A.K.; Dutta, P.; Kundu, M.; Nasipuri, M.; Basu, S. Prediction of Thyroid Cancer Genes Using an Ensemble of Post Translational Modification, Semantic and Structural Similarity Based Clustering Results. In Proceedings of the International Conference on Pattern Recognition and Machine Intelligence, Kolkata, India, 5–8 December 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 418–423.
68. Halder, A.K.; Denkiewicz, M.; Sengupta, K.; Basu, S.; Plewczynski, D. Aggregated Network Centrality Shows Non-Random Structure of Genomic and Proteomic Networks. *Methods* **2019**, *181–182*, 5–14. [CrossRef]
69. Bailey, N.T.J. *The Mathematical Theory of Infectious Diseases and Its Applications*; Charles Griffin & Company Ltd.: High Wycombe, UK, 1975; ISBN 0852642318.
70. Pesquita, C. Semantic Similarity in the Gene Ontology. *Methods Mol. Biol.* **2017**, *1446*, 161–173. [CrossRef]
71. Agrawal, M.; Zitnik, M.; Leskovec, J. Large-scale analysis of disease pathways in the human interactome. In Proceedings of the Pacific Symposium on Biocomputing 2018, Kohala Coast, HI, USA, 3–7 January 2018; World Scientific: Singapore, 2018; pp. 111–122.
72. Zitnik, M.; Sosic, R.; Maheshwari, S.; Leskovec, J. BioSNAP Datasets: Stanford Biomedical Network Dataset Collection. Available online: https://snap.stanford.edu/biodata/datasets/10015/10015-ChG-TargetDecagon.html (accessed on 1 January 2022).
73. Consortium, U. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699.
74. Hasöksüz, M.; Kilic, S.; Saraç, F. Coronaviruses and SARS-CoV-2. *Turk. J. Med. Sci.* **2020**, *50*, 549–556. [CrossRef] [PubMed]
75. Hu, B.; Guo, H.; Zhou, P.; Shi, Z.-L. Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* **2021**, *19*, 141–154. [CrossRef]
76. Decaro, N.; Lorusso, A. Novel human coronavirus (SARS-CoV-2): A lesson from animal coronaviruses. *Vet. Microbiol.* **2020**, *244*, 108693. [CrossRef] [PubMed]

77. Pfefferle, S.; Schöpf, J.; Kögl, M.; Friedel, C.C.; Müller, M.A.; Carbajo-Lozoya, J.; Stellberger, T.; von Dall'Armi, E.; Herzog, P.; Kallies, S.; et al. The SARS-coronavirus-host interactome: Identification of cyclophilins as target for pan-coronavirus inhibitors. *PLoS Pathog.* **2011**, *7*, e1002331. [CrossRef] [PubMed]
78. Yan, Y.; Chang, L.; Wang, L. Laboratory testing of SARS-CoV, MERS-CoV, and SARS-CoV-2 (2019-nCoV): Current status, challenges, and countermeasures. *Rev. Med. Virol.* **2020**, *30*, e2106. [CrossRef] [PubMed]
79. Naqvi, A.A.T.; Fatima, K.; Mohammad, T.; Fatima, U.; Singh, I.K.; Singh, A.; Atif, S.M.; Hariprasad, G.; Hasan, G.M.; Hassan, M.I. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochim. Biophys. Acta (BBA)-Mol. Basis Dis.* **2020**, *1866*, 165878. [CrossRef]
80. Balboni, A.; Battilani, M.; Prosperi, S. The SARS-like coronaviruses: The role of bats and evolutionary relationships with SARS coronavirus. *Microbiol. J. Microbiol. Sci.* **2012**, *35*, 1.
81. Buonocore, M.; Marino, C.; Grimaldi, M.; Santoro, A.; Firoznezhad, M.; Paciello, O.; Prisco, F.; D'Ursi, A.M. New putative animal reservoirs of SARS-CoV-2 in Italian fauna: A bioinformatic approach. *Heliyon* **2020**, *6*, e05430. [CrossRef]
82. Woo, P.C.Y.; Lau, S.K.P.; Li, K.S.M.; Poon, R.W.S.; Wong, B.H.L.; Tsoi, H.; Yip, B.C.K.; Huang, Y.; Chan, K.; Yuen, K. Molecular diversity of coronaviruses in bats. *Virology* **2006**, *351*, 180–187. [CrossRef]
83. Woo, P.C.Y.; Lau, S.K.P.; Li, K.S.M.; Tsang, A.K.L.; Yuen, K.-Y. Genetic relatedness of the novel human group C betacoronavirus to *Tylonycteris* bat coronavirus HKU4 and *Pipistrellus* bat coronavirus HKU5. *Emerg. Microbes Infect.* **2012**, *1*, 1–5. [CrossRef] [PubMed]
84. Wang, Q.; Qi, J.; Yuan, Y.; Xuan, Y.; Han, P.; Wan, Y.; Ji, W.; Li, Y.; Wu, Y.; Wang, J. Bat origins of MERS-CoV supported by bat coronavirus HKU4 usage of human receptor CD26. *Cell Host Microbe* **2014**, *16*, 328–337. [CrossRef]
85. Abdel-Moneim, A.S. Middle East respiratory syndrome coronavirus (MERS-CoV): Evidence and speculations. *Arch. Virol.* **2014**, *159*, 1575–1584. [CrossRef] [PubMed]
86. Cotten, M.; Watson, S.J.; Kellam, P.; Al-Rabeeah, A.A.; Makhdoom, H.Q.; Assiri, A.; Al-Tawfiq, J.A.; Alhakeem, R.F.; Madani, H.; AlRabiah, F.A.; et al. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: A descriptive genomic study. *Lancet* **2013**, *382*, 1993–2002. [CrossRef] [PubMed]
87. Kohn, D.F.; Clifford, C.B. Biology and diseases of rats. *Lab. Anim. Med.* **2002**, 121–165.
88. So, R.T.Y.; Chu, D.K.W.; Miguel, E.; Perera, R.A.P.M.; Oladipo, J.O.; Fassi-Fihri, O.; Aylet, G.; Ko, R.L.W.; Zhou, Z.; Cheng, M.-S.; et al. Diversity of dromedary camel coronavirus HKU23 in African camels revealed multiple recombination events among closely related betacoronaviruses of the subgenus Embecovirus. *J. Virol.* **2019**, *93*, e01236-19. [CrossRef]
89. Kyuwa, S.; Sugiura, Y. Role of cytotoxic T lymphocytes and interferon-γ in coronavirus infection: Lessons from murine coronavirus infections in mice. *J. Vet. Med. Sci.* **2020**, *82*, 1410–1414. [CrossRef]
90. Macphee, P.J.; Dindzans, V.J.; Fung, L.; Levy, G.A. Acute and chronic changes in the microcirculation of the liver in inbred strains of mice following infection with mouse hepatitis virus type 3. *Hepatology* **1985**, *5*, 649–660. [CrossRef]
91. Körner, R.W.; Majjouti, M.; Alcazar, M.A.A.; Mahabir, E. Of mice and men: The coronavirus MHV and mouse models as a translational approach to understand SARS-CoV-2. *Viruses* **2020**, *12*, 880. [CrossRef]
92. Orzechowski, M. Alpaca Coronavirus Sequences Producing Significant Alignments to Human Betacoronavirus. 2022. Available online: https://oatd.org/oatd/record?record=oai%5C:figshare.com%5C:article%5C%2F16934896 (accessed on 1 January 2022).
93. Woo, P.C.Y.; Huang, Y.; Lau, S.K.P.; Yuen, K.-Y. Coronavirus genomics and bioinformatics analysis. *Viruses* **2010**, *2*, 1804–1820. [CrossRef]
94. Li, F. Structure, function, and evolution of coronavirus spike proteins. *Annu. Rev. Virol.* **2016**, *3*, 237–261. [CrossRef] [PubMed]
95. Fehr, A.R.; Perlman, S. Coronaviruses: An overview of their replication and pathogenesis. In *Coronaviruses: An Overview of Their Replication and Pathogenesis*; Humana Press: New York, NY, USA, 2015; pp. 1–23.
96. de Mira Fernandes, A.; Brandão, P.E.; dos Santos Lima, M.; de Souza Nunes Martins, M.; da Silva, T.G.; da Silva Cardoso Pinto, V.; De Paula, L.T.; Vicente, M.E.S.; Okuda, L.H.; Pituco, E.M. Genetic diversity of BCoV in Brazilian cattle herds. *Vet. Med. Sci.* **2018**, *4*, 183–189. [CrossRef] [PubMed]
97. Asadi, A.H.; Baghinezhad, M.; Asadi, H. Neonatal calf diarrhea induced by rotavirus and coronavirus. *Int. J. Biosci.* **2015**, *6*, 230–236.
98. Saif, L.J. Bovine respiratory coronavirus. *Vet. Clin. Food Anim. Pract.* **2010**, *26*, 349–364. [CrossRef]
99. Yoo, D.; Pei, Y.; Christie, N.; Cooper, M. Primary structure of the sialodacryoadenitis virus genome: Sequence of the structural-protein region and its application for differential diagnosis. *Clin. Diagn. Lab. Immunol.* **2000**, *7*, 568–573. [CrossRef]
100. Haick, A.K.; Rzepka, J.P.; Brandon, E.; Balemba, O.B.; Miura, T.A. Neutrophils are needed for an effective immune response against pulmonary rat coronavirus infection, but also contribute to pathology. *J. Gen. Virol.* **2014**, *95*, 578. [CrossRef]
101. Bradley, L.M.; Douglass, M.F.; Chatterjee, D.; Akira, S.; Baaten, B.J.G. Matrix metalloprotease 9 mediates neutrophil migration into the airways in response to influenza virus-induced toll-like receptor signaling. *PLoS Pathog.* **2012**, *8*, e1002641. [CrossRef]
102. Denlinger, L.C.; Sorkness, R.L.; Lee, W.-M.; Evans, M.D.; Wolff, M.J.; Mathur, S.K.; Crisafi, G.M.; Gaworski, K.L.; Pappas, T.E.; Vrtis, R.F.; et al. Lower airway rhinovirus burden and the seasonal risk of asthma exacerbation. *Am. J. Respir. Crit. Care Med.* **2011**, *184*, 1007–1014. [CrossRef]
103. Khanolkar, A.; Hartwig, S.M.; Haag, B.A.; Meyerholz, D.K.; Harty, J.T.; Varga, S.M. Toll-like receptor 4 deficiency increases disease and mortality after mouse hepatitis virus type 1 infection of susceptible C3H mice. *J. Virol.* **2009**, *83*, 8946–8956. [CrossRef]

104. Nagata, N.; Iwata, N.; Hasegawa, H.; Fukushi, S.; Harashima, A.; Sato, Y.; Saijo, M.; Taguchi, F.; Morikawa, S.; Sata, T. Mouse-passaged severe acute respiratory syndrome-associated coronavirus leads to lethal pulmonary edema and diffuse alveolar damage in adult but not young mice. *Am. J. Pathol.* **2008**, *172*, 1625–1637. [CrossRef]
105. Khorsand, B.; Savadi, A.; Naghibzadeh, M. SARS-CoV-2-human protein-protein interaction network. *Inform. Med. Unlocked* **2020**, *20*, 100413. [CrossRef] [PubMed]
106. Consortium, U. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [CrossRef] [PubMed]
107. Gordon, D.E.; Jang, G.M.; Bouhaddou, M.; Xu, J.; Obernier, K.; White, K.M.; O'Meara, M.J.; Rezelj, V.V.; Guo, J.Z.; Swaney, D.L. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **2020**, *583*, 459–468. [CrossRef]
108. Dick, K.; Biggar, K.K.; Green, J.R. Computational Prediction of the Comprehensive SARS-CoV-2 vs. Human Interactome to Guide the Design of Therapeutics. *bioRxiv* **2020**. [CrossRef]
109. Schoenrock, A.; Dehne, F.; Green, J.R.; Golshani, A.; Pitre, S. Mp-pipe: A massively parallel protein-protein interaction prediction engine. In Proceedings of the international conference on Supercomputing, Tucson, AZ, USA, 31 May–4 June 2011; pp. 327–337.
110. Pitre, S.; Dehne, F.; Chan, A.; Cheetham, J.; Duong, A.; Emili, A.; Gebbia, M.; Greenblatt, J.; Jessulat, M.; Krogan, N. PIPE: A protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinform.* **2006**, *7*, 365. [CrossRef] [PubMed]
111. Pitre, S.; Hooshyar, M.; Schoenrock, A.; Samanfar, B.; Jessulat, M.; Green, J.R.; Dehne, F.; Golshani, A. Short co-occurring polypeptide regions can predict global protein interaction maps. *Sci. Rep.* **2012**, *2*, 239. [CrossRef]
112. Hsu, C.-H.; Hung, Y.; Chu, K.-A.; Chen, C.-F.; Yin, C.-H.; Lee, C.-C. Prognostic nomogram for elderly patients with acute respiratory failure receiving invasive mechanical ventilation: A nationwide population-based cohort study in Taiwan. *Sci. Rep.* **2020**, *10*, 13161. [CrossRef]
113. Li, Y.; Ilie, L. SPRINT: Ultrafast protein-protein interaction prediction of the entire human interactome. *BMC Bioinform.* **2017**, *18*, 485. [CrossRef]
114. Wishart, D.S.; Knox, C.; Guo, A.C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906. [CrossRef]
115. Saha, S.; Halder, A.K.; Bandyopadhyay, S.S.; Chatterjee, P.; Nasipuri, M.; Bose, D.; Basu, S. Drug repurposing for COVID-19 using computational screening: Is Fostamatinib/R406 a potential candidate? *Methods* **2022**, *203*, 564–574. [CrossRef]
116. Chin, L.; Cox, J.; Esmail, S.; Franklin, M.; Le, D. COVID-19: Finding the Right Fit Identifying Potential Treatments Using a Data-Driven Approach. Drugbank White Papper. Available online: https://blog.drugbank.com/data-driven-approaches-to-identify-potential-covid-19-therapies/ (accessed on 1 January 2022).
117. Gautret, P.; Lagier, J.-C.; Parola, P.; Meddeb, L.; Mailhe, M.; Doudier, B.; Courjon, J.; Giordanengo, V.; Vieira, V.E.; Dupont, H.T.; et al. Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial. *Int. J. Antimicrob. Agents* **2020**, *56*, 105949. [CrossRef] [PubMed]
118. Harrison, C. Coronavirus puts drug repurposing on the fast track. *Nat. Biotechnol.* **2020**, *38*, 379–381. [CrossRef] [PubMed]
119. De Wit, E.; Feldmann, F.; Cronin, J.; Jordan, R.; Okumura, A.; Thomas, T.; Scott, D.; Cihlar, T.; Feldmann, H. Prophylactic and therapeutic remdesivir (GS-5734) treatment in the rhesus macaque model of MERS-CoV infection. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 6771–6776. [CrossRef]
120. Saha, S.; Halder, A.K.; Bandyopadhyay, S.S.; Chatterjee, P.; Nasipuri, M.; Basu, S. Computational modeling of human-nCoV protein-protein interaction network. *Methods* **2022**, *203*, 488–497. [CrossRef] [PubMed]
121. Sun, P.; Guo, J.; Winnenburg, R.; Baumbach, J. Drug repurposing by integrated literature mining and drug–gene–disease triangulation. *Drug Discov. Today* **2017**, *22*, 615–619. [CrossRef] [PubMed]
122. Ondo, W. Ropinirole for restless legs syndrome. *Mov. Disord. Off. J. Mov. Disord. Soc.* **1999**, *14*, 138–140. [CrossRef]
123. Andreou, A.; Trantza, S.; Filippou, D.; Sipsas, N.; Tsiodras, S. COVID-19: The potential role of copper and N-acetylcysteine (NAC) in a combination of candidate antiviral treatments against SARS-CoV-2. *In Vivo* **2020**, *34*, 1567–1588. [CrossRef]
124. Kumar, S.; Choudhary, M. Synthesis and characterization of novel copper (II) complexes as potential drug candidates against SARS-CoV-2 main protease. *New J. Chem.* **2022**, *46*, 4911–4926. [CrossRef]
125. Wessels, I.; Rolles, B.; Rink, L. The potential impact of zinc supplementation on COVID-19 pathogenesis. *Front. Immunol.* **2020**, *11*, 1712. [CrossRef]
126. Hoffmann, M.; Kleine-Weber, H.; Krüger, N.; Müller, M.; Drosten, C.; Pöhlmann, S. The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *BioRxiv* **2020**. [CrossRef]
127. Chilvers, M.A.; McKean, M.; Rutman, A.; Myint, B.S.; Silverman, M.; O'Callaghan, C. The effects of coronavirus on human nasal ciliated respiratory epithelium. *Eur. Respir. J.* **2001**, *18*, 965–970. [CrossRef] [PubMed]
128. Darma, A.; Ranuh, I.G.M.R.G.; Merbawani, W.; Setyoningrum, R.A.; Hidajat, B.; Hidayati, S.N.; Endaryanto, A.; Sudarmo, S.M. Zinc supplementation effect on the bronchial cilia length, the number of cilia, and the number of intact bronchial cell in zinc deficiency rats. *Indones. Biomed. J.* **2020**, *12*, 78–84. [CrossRef]
129. Szarpak, L.; Pruc, M.; Gasecka, A.; Jaguszewski, M.J.; Michalski, T.; Peacock, F.W.; Smereka, J.; Pytkowska, K.; Filipiak, K.J. Should we supplement zinc in COVID-19 patients? Evidence from meta-analysis. *Pol. Arch. Intern. Med* **2021**, *131*, 802–807. [CrossRef]
130. Chinni, V.; El-Khoury, J.; Perera, M.; Bellomo, R.; Jones, D.; Bolton, D.; Ischia, J.; Patel, O. Zinc supplementation as an adjunct therapy for COVID-19: Challenges and opportunities. *Br. J. Clin. Pharmacol.* **2021**, *87*, 3737–3746. [CrossRef] [PubMed]

131. Li, G.; De Clercq, E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nat. Rev. Drug Discov.* **2020**, *19*, 149–150. [CrossRef]
132. Kumar, G.S.; Vadgaonkar, A.; Purunaik, S.; Shelatkar, R.; Vaidya Sr, V.G.; Ganu, G.; Vadgaonkar, A.; Joshi, S. Efficacy and Safety of Aspirin, Promethazine, and Micronutrients for Rapid Clinical Recovery in Mild to Moderate COVID-19 Patients: A Randomized Controlled Clinical Trial. *Cureus* **2022**, *14*, e25467.
133. Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, T.; Erichsen, S.; Schiergens, T.S.; Herrler, G.; Wu, N.-H.; Nitsche, A.; et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **2020**, *181*, 271–280.e8. [CrossRef]
134. Shang, J.; Ye, G.; Shi, K.; Wan, Y.; Luo, C.; Aihara, H.; Geng, Q.; Auerbach, A.; Li, F. Structural basis of receptor recognition by SARS-CoV-2. *Nature* **2020**, *581*, 221–224. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

International Journal of *Molecular Sciences*

MDPI

Article

RFCM-PALM: In-Silico Prediction of S-Palmitoylation Sites in the Synaptic Proteins for Male/Female Mouse Data

Soumyendu Sekhar Bandyopadhyay [1,2,†], Anup Kumar Halder [1,3,†], Monika Zaręba-Kozioł [4], Anna Bartkowiak-Kaczmarek [4], Aviinandaan Dutta [1], Piyali Chatterjee [5], Mita Nasipuri [1], Tomasz Wójtowicz [4], Jakub Wlodarczyk [4,*] and Subhadip Basu [1,*]

[1] Department of Computer Science and Engineering, Jadvapur University, Kolkata 700032, India; soumyabane@gmail.com (S.S.B.); anup21.halder@gmail.com (A.K.H.); aviinandaandutta@gmail.com (A.D.); mitanasipuri@gmail.com (M.N.)

[2] Department of Computer Science and Engineering, School of Engineering and Technology, Adamas University, Barasat, Kolkata 700126, India

[3] Department of Computer Science and Engineering, University of Engineering & Management, Kolkata 700156, India

[4] The Nencki Institute of Experimental Biology, Polish Academy of Sciences, 3 Pasteur Street, 02-093 Warsaw, Poland; m.zareba-koziol@nencki.edu.pl (M.Z.-K.); a.bartkowiak@nencki.edu.pl (A.B.-K.); t.wojtowicz@nencki.edu.pl (T.W.)

[5] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata 700152, India; piyali.gini@gmail.com

* Correspondence: j.wlodarczyk@nencki.edu.pl (J.W.); bsubhadip@gmail.com (S.B.)

† Equal Contribution, both shared the first authorship.

Citation: Bandyopadhyay, S.S.; Halder, A.K.; Zaręba-Kozioł, M.; Bartkowiak-Kaczmarek, A.; Dutta, A.; Chatterjee, P.; Nasipuri, M.; Wójtowicz, T.; Wlodarczyk, J.; Basu, S. RFCM-PALM: In-Silico Prediction of S-Palmitoylation Sites in the Synaptic Proteins for Male/Female Mouse Data. *Int. J. Mol. Sci.* **2021**, *22*, 9901. https://doi.org/10.3390/ijms22189901

Academic Editor: Yuriy L. Orlov

Received: 19 August 2021
Accepted: 9 September 2021
Published: 14 September 2021

Abstract: S-palmitoylation is a reversible covalent post-translational modification of cysteine thiol side chain by palmitic acid. S-palmitoylation plays a critical role in a variety of biological processes and is engaged in several human diseases. Therefore, identifying specific sites of this modification is crucial for understanding their functional consequences in physiology and pathology. We present a random forest (RF) classifier-based consensus strategy (RFCM-PALM) for predicting the palmitoylated cysteine sites on synaptic proteins from male/female mouse data. To design the prediction model, we have introduced a heuristic strategy for selection of the optimum set of physicochemical features from the AAIndex dataset using (a) K-Best (KB) features, (b) genetic algorithm (GA), and (c) a union (UN) of KB and GA based features. Furthermore, decisions from best-trained models of the KB, GA, and UN-based classifiers are combined by designing a three-star quality consensus strategy to further refine and enhance the scores of the individual models. The experiment is carried out on three categorized synaptic protein datasets of a male mouse, female mouse, and combined (male + female), whereas in each group, weighted data is used as training, and knock-out is used as the hold-out set for performance evaluation and comparison. RFCM-PALM shows ~80% area under curve (AUC) score in all three categories of datasets and achieve 10% average accuracy (male—15%, female—15%, and combined—7%) improvements on the hold-out set compared to the *state-of-the-art* approaches. To summarize, our method with efficient feature selection and novel consensus strategy shows significant performance gains in the prediction of S-palmitoylation sites in mouse datasets.

Keywords: S-palmitoylation; post-translational modifications; feature selection; genetic algorithm; random-forest; consensus; knock-out; amino acid index; propensity; synaptic protein

1. Introduction

Brain functions strictly depend on precise regulation of structural and functional synaptic integrity. Among the mechanisms governing synaptic protein functions, post-translational modifications (PTM) [1,2] play a pivotal role. PTMs may influence synaptic protein activity and turnover, localization at the synapse, and signaling cascades [3–6].

One of the PTMs is protein S-palmitoylation (S-PALM) involving covalent attachment of palmitic acid (C16:0) to cysteine residue(s) via a thioester bond. Recent studies showed that S-palmitoylation can modulate protein localization, stability, activities, and trafficking and play an essential role in various biological processes, including synaptic plasticity [7,8], cell signaling, cellular differentiation [9], and apoptosis [10].

Unlike other fatty acid modifications, S-palmitoylation is a reversible process, tightly regulated by two groups of enzymes: palmitoyl acyltransferases (PATs, palmitoylating enzymes) and palmitoyl thioesterases (depalmitoylating enzyme). It is widely accepted that repeated cycles of palmitoylation/depalmitoylation are critically involved in regulating multiple protein functions. The molecular mechanisms that lie behind site-specific protein S-palmitoylation remain largely unknown. Several human diseases are often associated with the atypical activity of PATs together with changes in the pattern of S-palmitoylation. S-PALM has been implicated in a wide range of human disease states such as cancer [11], Alzheimer's disease [12], Parkinson's disease, cardiovascular disease, schizophrenia [13], or major depressive disorder MDD [14]. Therefore, identifying substrates that undergo S-PALM and specific sites of these modifications may provide candidates for targeted therapy.

Twenty-three PATs have been identified in mammalian cells, which mediate the majority of protein S-palmitoylation. One of the known PATs is a zinc finger DHHC domain-containing protein 7 (Zdhhc7, abbreviated ZDHHC7). This enzyme palmitoylates various synaptic proteins involved in the regulation of cellular polarity and proliferation [15,16]. Moreover, Zdhhc7 is responsible for S-palmitoylation of sex steroid receptors such as estrogen and progesterone receptors [16–18]. Importantly, *Zdhhc7*$^{-/-}$ mice developed symptoms characteristic of human Bartter syndrome (BS) type IV because ZDHHC7 protein may affect ClC-K-barttin channel activation [19]. Thus, targeting ZDHHC7 activity may offer a potential therapeutic strategy in certain brain pathophysiological states. Most recently, using the mass spectrometry approach, we have identified sex-dependent differences in the S-PALM of synaptic proteins potentially involved in the regulation of membrane excitability and synaptic transmission as well as in the signaling of proteins involved in the structural plasticity of dendritic spines in the mice brain [18]. Our data showed for the first time sex-dependent action of ZDHHC7 acyltransferase. Furthermore, we revealed that different S-PALM proteins control the same biological processes in male and female synapses [18,19].

Several methods have been developed for the identification of S-palmitoylation target proteins. However, site-specific identification of S-palmitoylation is less studied. Large-scale identification of S-palmitoylation sites mainly relies on mass spectrometry-based methods such as PANIMoni developed in our lab [20] or PALMPiscs or ssABE [21]. These methods have been successfully used to identify a large number of S-palmitoylated proteins in different species, such as rats, mice, or humans. For instance, PANIMoni has been used to describe endogenous S-palmitoylation and S-nitrosylation of proteins in the rat brain excitatory synapses at the level of specific single cysteine in a mouse model of depression [20]. In recent years, results of large-scale proteome databases obtained with PANIMoni, PALMPiscs, or ssABE methods were used to develop tools to predict sites of specific S-palmitoylation in other biological complexes. Several machine learning-based algorithms [22–25] have been developed for predictions of S-palmitoylation sites such as; NBA-PALM [26] and CSS-PALM [25], but their accuracy is uncertain. Therefore, with the growing number of publicly available large-scale proteome databases of the brain and somatic tissues, there is a need for the development of reliable and accurate computational tools to process them.

Considering the growing recognition for the importance of post-translational modifications of proteins in cell physiology, this study aims to develop a computational tool for predicting S-palmitoylation sites using proteomic data obtained by the mass spectrometry-based method PANIMoni [20]. Most recently, we have successfully used this approach to create a detailed ZDHHC subtype-specific and sex-mouse S-palmitoylome [18,19]. Here, we have used this protein database for validation of the computational tool described in this study.

Our tool is focused on a random forest (RF) [27] classifier-based consensus strategy, which can predict the palmitoylated cysteine sites on synaptic proteins of the male/female mouse dataset. Different heuristic selection strategies have been applied on the physicochemical features from the AAIndex feature database [28] along with position-specific amino acid (AA) propensity information, which eventually generates three different sets of features: (a) K-Best (KB) features, (b) genetic algorithm (GA) based features [29], and (c) a union (UN) of K-Best and GA based features. The experiment has been carried out on three categorized synaptic protein datasets originally described in our previous publications [18,19]; *viz.*, male, female, and combined (male + female). In each experimental group, the weighted data is used as the training set, and the knock-out is used as the hold-out test set for performance evaluation and comparison. A novel RF-driven consensus strategy with efficient feature selection shows significant performance in predicting S-palmitoylation sites in mouse data.

2. Results

Our method, RFCM-PALM, predicts the S-palmitoylation sites from the primary sequence information of synaptic proteins. In the mouse model experiments, three categories of data, *viz.*, Male, Female, and Combined, and three different feature sets, *viz.*, KB, GA, UN, along with the RF classifier, have been used. The rationale behind the choice of the RF classifier is elaborated in the Supplementary Section S1 and Table S1. Features are extracted from the sequence motifs of variable length, and detailed experiments are conducted to select the optimum length of such sequence motif. A summary of these experiments is discussed in Section 4.3, and detailed results are reported in the Supplementary Table S2. Finally, the proposed approach presents a three-star consensus model for the final classification task. The efficacy of PTM prediction depends heavily on selecting appropriate feature sets, the choice of the classifier, and the underlying evaluation strategy. In this work, GA-based features show better the area under the curve (AUC) score for male, female, and combined datasets. The UN features show promising performances for the female dataset with higher accuracy, whereas KB and GA features achieve the highest accuracy in male and combined datasets, respectively. Finally, we present a three-star consensus approach for the final classification task. The consensus model significantly improved the performance compared to individual feature-specific models. We have further compared the proposed consensus-based approach, RFCM-PALM, with two *state-of-the-art* methods.

2.1. Performance Evaluation

The performance of the proposed model has been evaluated with five-fold cross-validation on three different feature sets (namely KB, GA, UN) using a RF classifier. Five-fold cross-validation has been introduced to estimate the model's strength on all three categories of datasets (Male (M), Female (F), and Combined (M + F)), and the performances are reported in Table 1. The individual fold-wise performances on all three datasets are reported in Supplementary Table S3. In all three datasets, the GA-based feature outperforms the rest two in AUC score. However, in our proposed method, for fold-wise testing, the GA-based feature shows a ~79% AUC score for both male and combined datasets, and 80% AUC on the female dataset, surpassing the other two features. For female data, the UN-based feature outperforms KB and GA-based features, having an accuracy score of 71.9% and F1 of 71.3% (see Table 1). The AUC and AUPRC curves from training models are shown in Figure 1.

The knock-out data has been used as the hold-out test set from three categories of data (Male, Female, and Combined) individually. In the knock-out hold-out test set, the GA-based feature shows better performance for all the datasets than other features with an AUC score of ~66.4% in males, 68.6% in females, and 62.5% in combined datasets (please see Table 2). Moreover, GA has higher accuracy in all hold-out test data except the males set, where the KB-based model achieves 62% accuracy. Furthermore, we have introduced a consensus strategy for the final classification of S-PALM on the hold-out test set. Initially, the best models are extracted from the cross-validation strategy for each feature set on the three categories of data set independently.

Thus, the three best models are identified for classification from each data set (Male/Female/Combined). Finally, three consensus-based classifications are obtained for the final classification. The 1*-consensus (1*Con), 2*-consensus (2*Con), and 3*-consensus (3*Con) represent 1, 2, and 3 model confidence scores, respectively. The detailed consensus mechanism is shown in Figure 2, and the results are depicted in Table 2. The 2*Con (2 model confidence) has significantly improved performance compared to the corresponding individual models. Consensus-based performance with different categories of data for hold-out test sets is shown in Table 2.

Table 1. Performance evaluation of S-palmitoylation prediction from 5-fold cross-validation using three different sets of features on three data types, Male, Female, and Combined.

Type	Feature	5-Fold Cross-Validation					
		AvgAUC	MaxAUC	Precision	Recall	Accuracy	F1
Male	KB	0.785 ± 0.013	0.801	0.732 ± 0.02	0.666 ± 0.02	0.711 ± 0.01	0.697 ± 0.016
	GA	0.790 ± 0.013	0.812	0.726 ± 0.02	0.675 ± 0.02	0.710 ± 0.02	0.700 ± 0.02
	UN	0.786 ± 0.013	0.798	0.726 ± 0.02	0.662 ± 0.01	0.706 ± 0.01	0.693 ± 0.01
Female	KB	0.796 ± 0.02	0.82	0.715 ± 0.02	0.701 ± 0.02	0.708 ± 0.02	0.706 ± 0.02
	GA	0.801 ± 0.018	0.827	0.732 ± 0.02	0.69 ± 0.04	0.718 ± 0.02	0.709 ± 0.02
	UN	0799 ± 0.018	0.821	0.729 ± 0.02	0.698 ± 0.03	0.719 ± 0.02	0.713 ± 0.02
Combined	KB	0.791 ± 0.02	0.830	0.718 ± 0.04	0.689 ± 0.02	0.708 ± 0.03	0.703 ± 0.03
	GA	0.795 ± 0.02	0.830	0.733 ± 0.03	0.684 ± 0.03	0.717 ± 0.02	0.707 ± 0.03
	UN	0.793 ± 0.02	0.820	0.734 ± 0.02	0.670 ± 0.01	0.714 ± 0.02	0.701 ± 0.02

Table 2. Performance evaluation using fold-wise and consensus strategy on hold-out test data.

Dataset		Feature	Precision	Recall	Accuracy	F1	MCC	AUC
Male	Fold-wise	KB	0.643 ± 0.01	0.54 ± 0.02	0.620 ± 0.01	0.587 ± 0.01	0.244 ± 0.02	0.661 ± 0.01
		GA	0.629 ± 0.01	0.535 ± 0.02	0.609 ± 0.01	0.578 ± 0.01	0.222 ± 0.02	0.664 ± 0.01
		UN	0.634 ± 0.02	0.532 ± 0.01	0.612 ± 0.01	0.579 ± 0.01	0.227 ± 0.03	0.661 ± 0.01
	Consensus	1*Con	0.585	0.812	0.618	0.68	0.255	0.639
		2*Con	0.667	0.713	0.678	0.689	0.357	0.679
		3*Con	0.676	0.423	0.610	0.520	0.238	0.628
Female	Fold-wise	KB	0.617 ± 0.01	0.566 ± 0.01	0.608 ± 0.01	0.591 ± 0.01	0.216 ± 0.02	0.667 ± 0.01
		GA	0.641 ± 0.01	0.600 ± 0.01	0.632 ± 0.01	0.62 ± 0.01	0.265 ± 0.01	0.686 ± 0.004
		UN	0.622 ± 0.01	0.566 ± 0.02	0.611 ± 0.01	0.593 ± 0.01	0.223 ± 0.02	0.684 ± 0.004
	Consensus	1*Con	0.593	0.792	0.624	0.678	0.264	0.64
		2*Con	0.799	0.706	0.764	0.749	0.532	0.768
		3*Con	0.800	0.447	0.668	0.573	0.373	0.708
Combined	Fold-wise	KB	0.586 ± 0.02	0.475 ± 0.01	0.57 ± 0.01	0.525 ± 0.01	0.142 ± 0.02	0.597 ± 0.01
		GA	0.608 ± 0.02	0.486 ± 0.02	0.586 ± 0.02	0.54 ± 0.02	0.176 ± 0.03	0.625 ± 0.01
		UN	0.605 ± 0.02	0.472 ± 0.02	0.581 ± 0.02	0.53 ± 0.02	0.167 ± 0.03	0.615 ± 0.01
	Consensus	1*Con	0.654	0.719	0.669	0.685	0.340	0.671
		2*Con	0.679	0.669	0.676	0.674	0.353	0.676
		3*Con	0.612	0.374	0.568	0.464	0.148	0.580

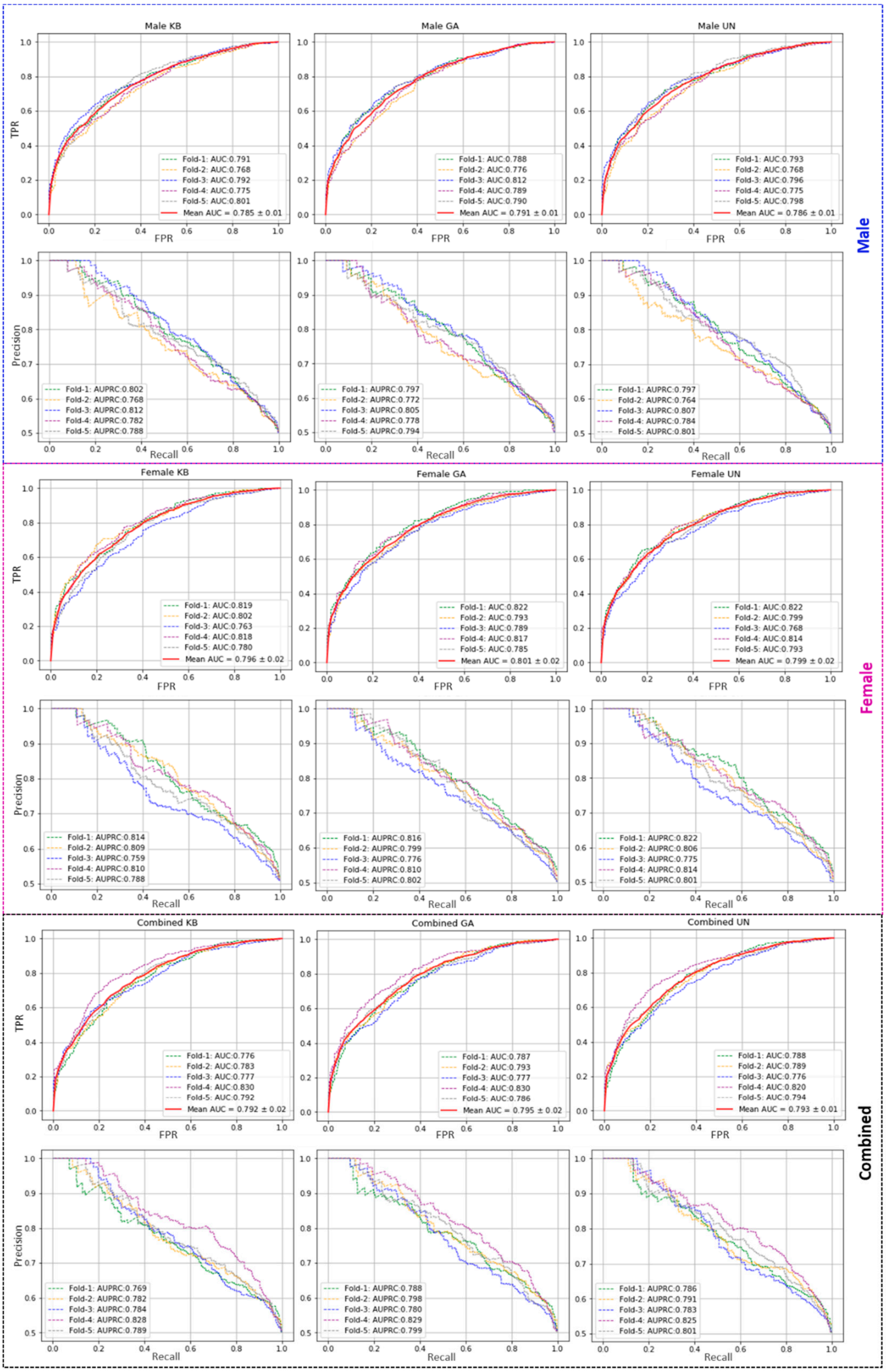


Figure 1. Performance evaluation on three datasets, Male, Female, and Combined. Plots in the 1st, 3rd, and 5th rows show the AUC, and the 2nd, 4th, and 6th rows represent AUPRC, respectively. The 1st, 2nd, and 3rd column-wise plots represent KB, GA, and UN type features-based evaluation.

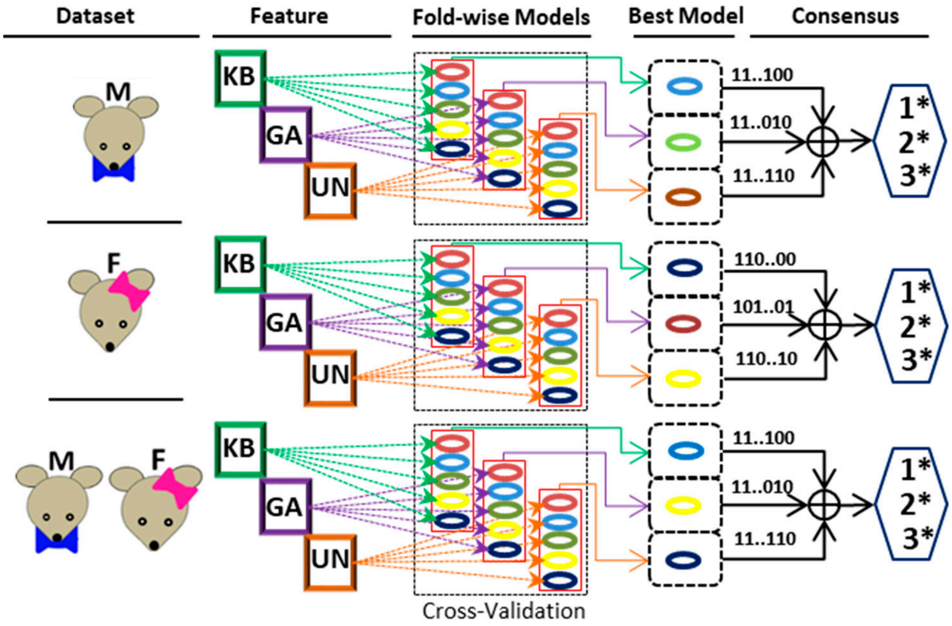


Figure 2. A schematic diagram depicting the underlying consensus strategy for S-PALM prediction.

2.2. Comparison with the State-of-the-Art Approaches

To demonstrate the performance of our proposed method, we have compared our approach with existing PTM prediction models. We have identified three *state-of-the-art* approaches for benchmarking purposes, CapsNet [23], MusiteDeep [24,30], and ModPred [31]. The CapsNet [23] is a deep learning-based architecture that provides prediction models for different PTM sites. MusiteDeep [24,30] is a deep-learning-based system that can predict general and kinase-specific phosphorylation sites from primary sequence information. ModPred [31] is a sequence-based PTM prediction tool developed on the structural and functional signatures of proteins. The CapsNet, provides a 10-fold cross-validation result on the benchmark dataset of animal species (metazoa), extracted from the NCBI taxonomy database [32], which has been curated by collecting annotations from Uniprot/Swiss-Prot (August 2007 release) [33] with less than 30% sequence similarity.

Our approach has also been trained with the similar dataset used in CapsNet [23] for S-palmitoylated cysteine prediction for comparison purposes. When compared with all three existing approaches on similar datasets, the performance scores are directly incorporated from Wang et al. [23]. In the proposed model, we have also presented the class-imbalanced learning by imposing a positive-negative ratio at 1:2 along with the balanced learning (1:1). The performance has been compared with the existing approaches concerning the AUC and AUPRC scores (Table 3). Our proposed method outperforms the *state-of-the-art* methods in both metrics. The AUC and AUPRC have improved by 8% in comparison with the earlier best-performing method. Additionally, the proposed approach has surpassed the prior approaches by 32% in the AUPRC score, as depicted in Table 3. The detailed fold-wise evaluation scores are shown in the Supplementary Table S4 (balanced) and Table S5 (imbalanced).

Table 3. Performance comparison with the *state-of-the-art* methods for S-PALM prediction.

Methods	AUC	AUPRC	Accuracy	F1	MCC
CapsNet [23]	0.780 ± 0.02	0.500 ± 0.07	NA	NA	NA
MusiteDeep [24]	0.771 ± 0.02	0.484 ± 0.05	NA	NA	NA
ModPred [31]	0.8553 ± 0.01	0.5973 ± 0.04	NA	NA	NA
Proposed Method (1:1)	0.936 ± 0.01	0.889 ± 0.02	0.824 ± 0.03	0.799 ± 0.04	0.669 ± 0.05
Proposed Method (1:2)	0.928 ± 0.02	0.785 ± 0.04	0.816 ± 0.02	0.645 ± 0.06	0.577 ± 0.06

To investigate the significance of our proposed model on a novel S-PALM dataset, we have evaluated and compared the performance with two web servers MusiteDeep [30] and CSS-Palm [25]. MusiteDeep [24,30] is a web resource with a deep-learning framework that can predict and visualize different post-translational modification (PTM) sites from protein sequence information. CSS-Palm [25] is developed based on clustering and scoring strategy (CSS) algorithm and Group-based Prediction System (GPS) algorithm. CSS-Palm is evaluated with two high-performing thresholds, as stated by the authors in [25]. The novel hold-out test data from male, female, and combined sets has been submitted in the above two web servers, and performances have been recorded for comparison purposes (see Table 4). The proposed method has achieved a better result in more balanced metrics F1, and MCC compared to each of these web servers in S-PALM prediction depicting the efficacy of the proposed method on S-PALM prediction. In all three datasets, male, female, and combined, the proposed approach has improved the F1 score by 54%, 52%, and 48%, and MCC score by 7%, 32%, and 13%, respectively.

Table 4. Performance comparison with MusiteDeep [24,30] and CSS-Palm [25] web server with holdout dataset.

Method		Type of Data	Precision	Recall	Accuracy	F1	MCC
MusiteDeep [30]		Male	0.827	0.088	0.535	0.159	0.155
		Female	0.808	0.107	0.51	0.188	0.151
		Combined	0.555	0.0719	0.507	0.127	0.029
CSS-Palm [25]	High Threshold	Male	0.857	0.132	0.555	0.229	0.206
		Female	0.783	0.147	0.524	0.247	0.168
		Combined	0.75	0.129	0.543	0.22	0.153
	Medium Threshold	Male	0.768	0.158	0.555	0.262	0.182
		Female	0.761	0.177	0.532	0.288	0.173
		Combined	0.735	0.179	0.557	0.289	0.176
Proposed Method		Male	0.628	0.539	0.609	0.58	0.222
		Female	0.639	0.583	0.627	0.61	0.254
		Combined	0.623	0.504	0.599	0.556	0.202

In this novel hold-out data set, both web servers show high precision (0.827 in MusiteDeep and 0.857 in CSS-Palm) and very low recall (0.0882 in MusiteDeep and 0.1324 in CSS-Palm). A high precision score depicts low false positivity, and low recall depicts the increase in false-negative data, which can be interpreted as a failure for predicting the positive data. This may lead to a biased classification. Low recall also results in a low F1 score, which is the harmonic mean of precision and recall. Not only the recall score, but the MCC score for both the web servers are low, which depicts the failure of the class imbalance issue [34]. In contrast, our proposed method achieves 0.638 precision, and 0.583 recall scores on this hold-out dataset, which shows a more balanced scenario of classification outcome. In addition, our proposed method shows the highest accuracy for all three categories of the data, which outperforms the other two (accuracy improvement by 9%, 15%, and 7% in male, female, and the combined dataset).

3. Discussion

Our method, RFCM-PALM, computationally predicts the S-palmitoylation sites using the primary sequence information of the synaptic group of proteins from three categories of mouse data, designed as sex-dependent (male, female) and sex-independent (combined) mode. The computational model has been developed through a rigorous feature selection strategy and optimal model selection for predicting the S-PALM modification sites in a given subsequence window. The proposed model has been evaluated with five-fold cross-validation, and model performances have been compared with the *state-of-the-art* approaches using three different feature sets; KB, GA, and UN. Finally, a consensus strategy is designed based on the feature-specific best models from their cross-validated models.

The performance of the consensus model improved significantly compared to *state-of-the-art* approaches. The significant performance improvement in predicting S-PALM modification sites portrayed the efficacy of the proposed method.

The performance of the method may further be enhanced by incorporating deep-learning models. However, the major bottleneck lies with the limitation of adequate training samples. Furthermore, due to the complex nature of the biological experiments, scalability of the experimentally validated samples may not be easy. The development of the RFCM-PALM web server is also in our plans. We also plan to extend the method for other PTM types to predict protein nitrosylation sites in the synaptic proteins.

4. Materials and Methods

4.1. Dataset Preparation

Experimental S-Palmitoylated datasets are categorized into three groups, male, female, and combined (includes both male and female), where each category contains two types of data: weight (WT) and knock-out (KO). Weight data is used for training, and knock-out data is considered for testing. The dataset was derived using the mass spectrometry-based PANIMoni method from WT and koZDHHC7 mouse brains. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD025286.

The benchmark dataset for this experiment is constructed with the data available in the said repository. In this experiment, all three benchmarking datasets, namely, male, female, and combined, weight data is considered a train set, and knock-out data is considered the test set for classification. Both male and female datasets contain peptides, modified sites, and assigned proteins. All the modified cysteines are labeled. The cysteines which are labeled with Carbamidomethyl are palmitoylated and are considered as positive data. The cysteines which are labeled as N-ethylmaleimide are not palmitoylated and they constitute the negative data. In this approach, to retrieve the high-quality negative samples, the cysteine positions, which are not within the selected fragments of positive samples, are considered. However, the cysteine position that belongs to the same protein but not in the selected fragment is considered as the negative data for the classification. The cysteine positions with both Carbamidomethyl and N-ethylmaleimide modification create ambiguity in S-PALM identification and thus are discarded from this experiment. The number of positive and negative sites for S-PALM prediction is given in Table 5. In all experiments, the positive and negative ratio is kept as 1:1 for balanced classification. The details of the three benchmark datasets are shown in Table 5.

Table 5. Dataset details of positive and negative sites for all three benchmark data; Male, Female, and Combined.

Category	Type	# Protein	# Cysteine Sites
Male	Positive (P_D)	1077	1870 (Experimental)
	Negative (N_D)	1175	9279 (Identified)
Female	Positive (P_D)	1036	1773 (Experimental)
	Negative (N_D)	1131	8934 (Identified)
Combined (Male + Female)	Positive (P_D)	1180	2083 (Experimental)
	Negative (N_D)	1293	10,403 (Identified)

4.2. Features

In this work, we have incorporated amino acid physicochemical properties to design the features for the classification task [28]. The position-specific amino acid propensity is computed from the primary sequence of proteins using the physicochemical properties of each amino acid. We have extracted a λ-length sequence window for each cysteine site with the cysteine at the center of the subsequence.

4.2.1. Position-Specific Amino Acid Propensity (PSAAP)

The position-specific feature of amino acid is introduced for feature design. First, the position-specific amino acid composition is computed for all λ-length sub-sequences in the positive dataset (say, P_D). Initially, the positive data set is divided into five different non-overlapping subsets. For any subset of positive data, the amino acid composition for $i-th$ position is defined as, $\left(A^P_{1,i},\ A^P_{2,i},\ A^P_{3,i},\ A^P_{4,i}\ldots\ldots\ A^P_{20,i}\right)^T$ where, $i=1,2,\ 3,\ \ldots\lambda$ and 20 amino acids are ordered alphabetically according to their single letter code. Then, the position-specific amino acid composition is computed as the position-wise average over all five subsets, denoted as $\overline{A}^P_{1,i}$. Similarly, the negative dataset is partitioned into five equal partitions where each subset size = $|N_D| = |P_D|$. The position-wise amino acid composition is computed for all negative subsets (as done in the case of P_D). The position-wise amino acid composition for individual negative subsets is calculated as, $\left(A^N_{1,i},\ A^N_{2,i},\ A^N_{3,i},\ A^N_{4,i}\ldots\ldots\ A^N_{20,i}\right)^T$ where, $i=1,2,\ 3,\ \ldots\lambda$. The average of amino acid composition over five negative subsets is represented as $\overline{A}^N_{1,i}$.

Finally, the propensity of the $j-th$ amino acid at position i in the cysteine sites is computed as:

$$\chi_{i,j} = \frac{\overline{A}^P_{j,i} - \overline{A}^N_{j,i}}{\overline{\sigma}^N_{j,i}},$$

where, $\overline{\sigma}$ represents the standard deviation of $j-th$ amino acid at position i overall negative subsets. With these propensity values, final propensity matrix $ProP_{20\times\lambda}$ is constructed as

$$ProP_{20\times\lambda} = \begin{bmatrix} \chi_{1,1} & \cdots & \chi_{1,\lambda} \\ \vdots & \ddots & \vdots \\ \chi_{20,1} & \cdots & \chi_{20,\lambda} \end{bmatrix}$$

4.2.2. Physicochemical Properties Based PSAAP

In the next level, a physicochemical property-based feature is generated by incorporating the PSAAP ($ProP$). Currently, there are 566 physicochemical features in the AAIndex database [28]. A numeric score is assigned to each amino acid in the AAIndex database representing any particular physicochemical property scale. Then, the scores are normalized by [0, 1] for all amino acids for individual AAIndex using max–min normalization. From any target subsequence ($length = \lambda$), the final feature for any amino acid θ at position ι is for amino acid property φ defined as

$$\tau(\theta,\iota) = ProP(Ordx(\theta),\iota) \times PHY_\varphi(\theta,\iota)$$

where, $Ordx(\theta)$ represent the ordering index of amino acid θ in $ProP$ matrix and $PHY_\varphi(\theta,\iota)$.

4.3. Sub-Sequence Length Selection

To prepare the dataset, protein sequences are segmented into equal-length windows containing the cysteine at the center position. Amino acid sequences before and after the cysteine position in the sequence window are referred to as backward (BW) and forward (FW) subsequences, respectively. The window size (λ) is varied from 31 to 41 (i.e., $|BW| = |FW| = n$ is varied from 5 to 20 and $\lambda = (2\times n+1)$). Different length-wise experimental analysis has been carried out to find the optimal subsequence length (window size). Based on the AUC score, it has been found that the performance is optimum when $n = 19$ (window size $= 2\times 19+1$) as depicted in Table 6. Thus, the length of the subsequence in this approach is set to 19 for all consecutive experiments.

Table 6. Performance with different length of sub-sequences.

Length (n)	Precision	Recall	Accuracy	F1	AUC
15	0.657	0.792	0.69	0.718	0.765
16	0.701	0.731	0.709	0.715	0.781
17	0.699	0.722	0.706	0.71	0.777
18	0.72	0.731	0.722	0.725	0.788
19	0.724	0.717	0.723	0.72	0.79
20	0.715	0.731	0.719	0.723	0.789

4.4. Feature Selection

In the present work, we have introduced two different types of feature optimization strategies for predicting the S-palmitoylation sites in mouse protein. The method includes a K-Best (KB)-based feature optimization strategy and a genetic algorithm (GA)-based feature optimization strategy. We have employed both strategies on three types of datasets (discussed above) and recorded their performances, evaluated on the cross-validated test set, and hold-out test set. A detailed discussion of each feature optimization strategy is discussed in the following section.

4.4.1. K-Best Feature Selection

We have introduced the K-Best feature selection strategy to identify significant and non-redundant features from 566 physicochemical property-based PSAAP features. Initially, individual physicochemical property-wise performance has been evaluated with different varying subsequence lengths (31 to 41). Based on these performances (AUC score), physicochemical properties are sorted/ranked for individual subsequence length. Top-performing K features are extracted from each subsequences length-wise evaluation with four different thresholds of K (as top 25, 50, 75, and 100). Finally, two sets of features are constructed by considering the intersection of K-best (IB-K) and union of K-best (UB-K) features from different length-wise evaluations.

Once retrieving these K-best feature sets, performance has been evaluated with the merged feature where individual features are concatenated into a single feature vector for final representation. The concatenated feature is generated for the window length 39 (=2 * n + 1, where n = 19) as it shows superior performance compared to other window lengths. The Union and Intersection-based performance evaluation with four different thresholds (25, 50, 75, and 100) are depicted in Table 7. Based on AUC and accuracy scores, we concluded that at window length 39 with IB25 gives the best result with the highest AUC score among all (see Table 7), thus constitute the K-best features (KB). Figure 3 shows the detailed workflow for selecting the K-Best feature from the 566 feature set. Finally, the KB feature results in 19, 20 and 21 features in male, female, and the combined datasets, respectively.

Table 7. Performance of top K features.

Feature	Precision	Recall	Accuracy	F1	AUC
IB25	0.724	0.717	0.722	0.72	0.79
IB50	0.715	0.713	0.715	0.714	0.784
IB75	0.702	0.673	0.694	0.687	0.772
IB100	0.707	0.702	0.705	0.704	0.775
UB25	0.72	0.722	0.72	0.721	0.789
UB50	0.714	0.715	0.714	0.714	0.782
UB75	0.709	0.706	0.708	0.707	0.778
UB100	0.703	0.700	0.702	0.701	0.771

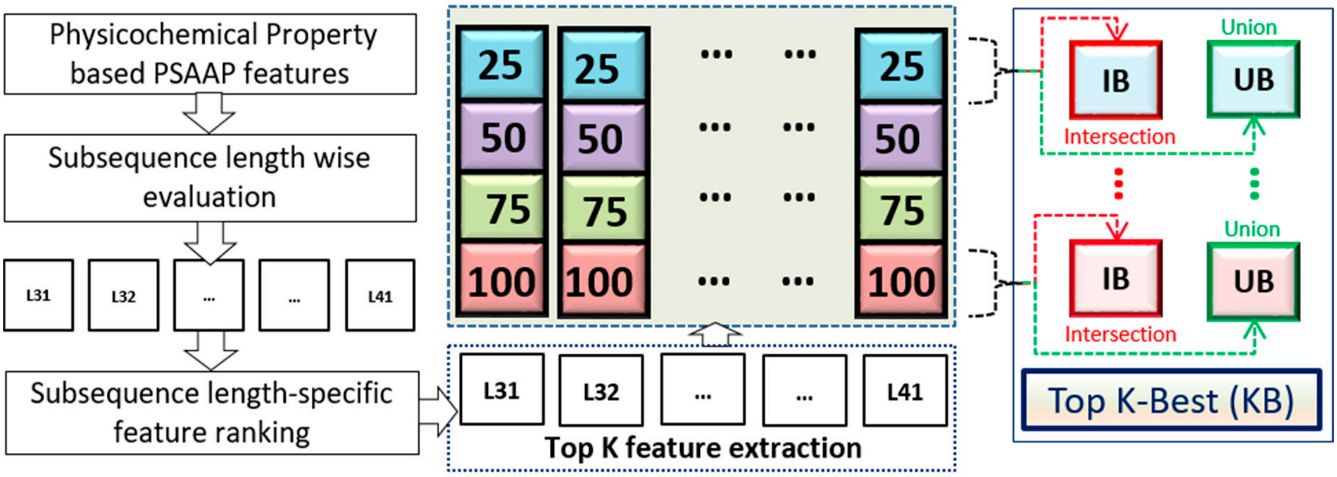

Physicochemical Property based PSAAP features
Subsequence length wise evaluation
L31
L32
...
...
L41
Subsequence length-specific feature ranking
25
50
75
100
L31
L32
...
...
L41
Top K feature extraction
IB
UB
Intersection
Union
Top K-Best (KB)

Figure 3. A detailed flow chart for K-Best feature selection.

4.4.2. Genetic Algorithm Based Feature Selection

Genetic algorithm (GA), which is inspired by the natural selection and evolution process, is a guided random optimized search technique that results in an excellent semi-optimal solution to the feature selection problem [35]. Under GA, fitter children (chromosome) populated from the earlier generation (parents) have a better chance of survival. The feature subsets are encoded as chromosomes are considered as individual and the collection of such chromosomes represent the population. Here, the chromosomes are encoded as a binary string where '1' at any position *i* of represents the selection of *i-th* feature and '0' represents the refusal. Each chromosome representing a subset of features is given a fitness score, which is obtained as the AUC in predicting the correct S-PALM modification using this feature subset and RF classifier.

Initially, the 566 physicochemical properties are hierarchically clustered based on the amino acid properties. Then, the hierarchical cluster tree is partitioned into 331 non-singleton and 185 singleton clusters using the same splitting strategy proposed in [36]. In this experiment, GA has used in two steps:

- First, GA is employed over the non-singleton clusters to obtain the best performing feature among the cluster members.
- Second, GA is applied with the newly identified features from the non-singleton clusters and with the remaining features from singleton clusters.

In our proposed method, RF is used for classification purposes while evaluating the performance of feature(s) at each generation. However, the AUC score is incorporated in fitness/objective computation. In this experiment, roulette wheel selection strategy and uniform crossover are employed. The crossover probability (p) and uniform mutation probability (q) is set to 0.7 and 0.01, respectively, to populate the next generation chromosome. The positive and negative data ratio is kept as 1:1 for evaluation purposes. The tie between equally performing chromosomes, the one with the lesser number of features, is retained. The method results in the globally best chromosomes. Finally, the GA based approach identified 6 features in male, 7 in female and 21 features in the combined dataset, respectively, for final classification. The overall workflow of GA-based feature design is detailed in Figure 4.

In a nutshell, our tool RFCM-PALM has been developed with effective feature selection and consensus strategy for in silico prediction of S-palmitoylation in mouse protein and shows significant improvement. Sample datasets, supplementary files, and the prediction tool are available at https://github.com/anupgth/RFCM-PALM (accessed on 10 September 2021).

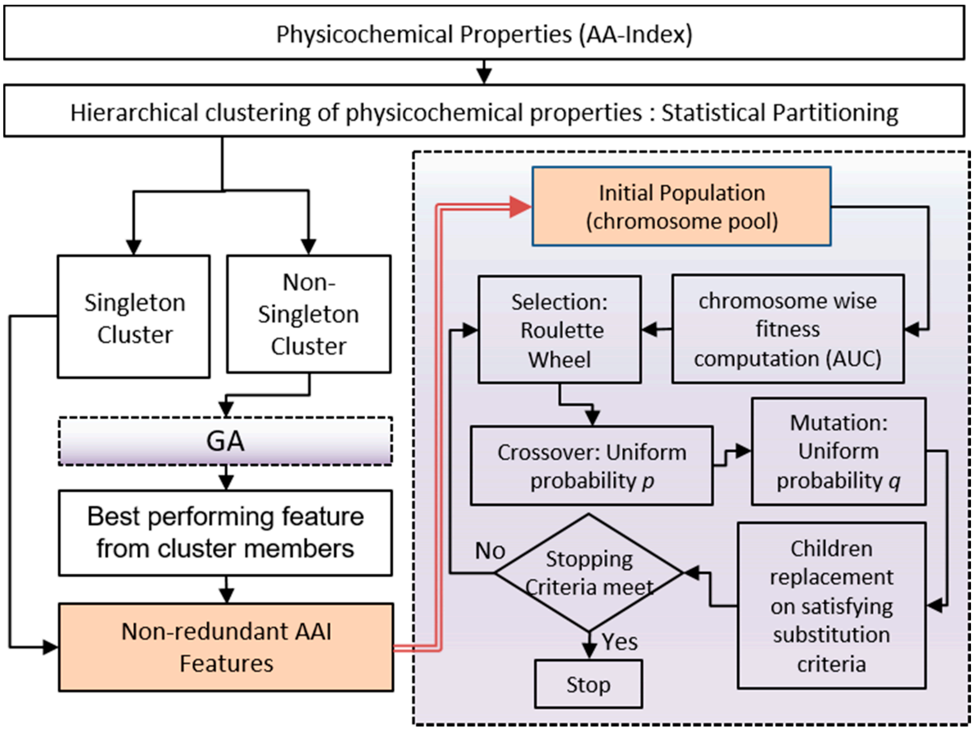

Physicochemical Properties (AA-Index)
Hierarchical clustering of physicochemical properties : Statistical Partitioning
Singleton Cluster
Non-Singleton Cluster
GA
Best performing feature from cluster members
Non-redundant AAI Features
Initial Population (chromosome pool)
Selection: Roulette Wheel
chromosome wise fitness computation (AUC)
Crossover: Uniform probability *p*
Mutation: Uniform probability *q*
No
Stopping Criteria meet
Yes
Children replacement on satisfying substitution criteria
Stop

Figure 4. Detailed workflow of GA based feature selection.

Supplementary Materials: Supplementary materials can be found at https://www.mdpi.com/article/10.3390/ijms22189901/s1.

Author Contributions: Conceptualization, S.B. and J.W.; methodology, A.K.H., A.D., S.S.B., M.Z.-K., A.B.-K. and T.W.; formal analysis, A.K.H., S.S.B., P.C., S.B., M.Z.-K. and J.W.; investigation, P.C., M.N., S.B., M.Z.-K. and J.W.; writing—original draft preparation, A.K.H., S.S.B., S.B., M.N., P.C., T.W., M.Z.-K. and J.W.; supervision, S.B., M.N., P.C., M.Z.-K. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Centre, Poland, grant number 2017/26/E/NZ4/00637. T.W. is financially supported by the National Science Centre, Poland, grant number 2019/34/E/NZ4/00387. S.B. is funded by the Department of Biotechnology grant (BT/PR16356/BID/7/596/2016), Government of India.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD025286.

Acknowledgments: This work is partially supported by the CMATER research laboratory of the Computer Science and Engineering Department, Jadavpur University, India, PURSE-II and UPE-II grants.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

GA	Genetic Algorithm
IB	Intersection Based
UB	Union Based
KB	K-Best
RF	Random Forrest
COMBI	Combined

SD	Standard Deviation
CV	Cross Validation
AAI	Amino Acid Index
AUC	Area under the curve
AUPRC	Area under the precision-recall curve
MCC	Matthews Correlation Coefficient

References

1. Barber, K.W.; Rinehart, J. The abcs of ptms. *Nat. Chem. Biol.* **2018**, *14*, 188–192. [CrossRef] [PubMed]
2. Jiang, J.; Suppiramaniam, V.; Wooten, M.W. Posttranslational modifications and receptor-associated proteins in AMPA receptor trafficking and synaptic plasticity. *Neurosignals* **2006**, *15*, 266–282. [CrossRef]
3. Lussier, M.P.; Sanz-Clemente, A.; Roche, K.W. Dynamic regulation of N-methyl-d-aspartate (NMDA) and α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors by posttranslational modifications. *J. Biol. Chem.* **2015**, *290*, 28596–28603. [CrossRef]
4. Ghosh, H.; Auguadri, L.; Battaglia, S.; Thirouin, Z.S.; Zemoura, K.; Messner, S.; Acuna, M.A.; Wildner, H.; Yévenes, G.E.; Dieter, A. Several posttranslational modifications act in concert to regulate gephyrin scaffolding and GABAergic transmission. *Nat. Commun.* **2016**, *7*, 1–16. [CrossRef]
5. Vallejo, D.; Codocedo, J.F.; Inestrosa, N.C. Posttranslational modifications regulate the postsynaptic localization of PSD-95. *Mol. Neurobiol.* **2017**, *54*, 1759–1776. [CrossRef]
6. Bradley, S.A.; Steinert, J.R. Nitric oxide-mediated posttranslational modifications: Impacts at the synapse. *Oxid. Med. Cell. Longev.* **2016**, *2016*, 5681036. [CrossRef] [PubMed]
7. Fukata, Y.; Fukata, M. Protein palmitoylation in neuronal development and synaptic plasticity. *Nat. Rev. Neurosci.* **2010**, *11*, 161–175. [CrossRef]
8. Kang, R.; Wan, J.; Arstikaitis, P.; Takahashi, H.; Huang, K.; Bailey, A.O.; Thompson, J.X.; Roth, A.F.; Drisdel, R.C.; Mastro, R. Neural palmitoyl-proteomics reveals dynamic synaptic palmitoylation. *Nature* **2008**, *456*, 904–909. [CrossRef]
9. Zhang, M.M.; Hang, H.C. Protein S-palmitoylation in cellular differentiation. *Biochem. Soc. Trans.* **2017**, *45*, 275–285. [CrossRef] [PubMed]
10. Fröhlich, M.; Dejanovic, B.; Kashkar, H.; Schwarz, G.; Nussberger, S. S-palmitoylation represents a novel mechanism regulating the mitochondrial targeting of BAX and initiation of apoptosis. *Cell Death Dis.* **2014**, *5*, e1057. [CrossRef]
11. Yeste-Velasco, M.; Linder, M.E.; Lu, Y.-J. Protein S-palmitoylation and cancer. *Biochim. Biophys. Acta (BBA)-Rev. Cancer* **2015**, *1856*, 107–120. [CrossRef] [PubMed]
12. Meckler, X.; Roseman, J.; Das, P.; Cheng, H.; Pei, S.; Keat, M.; Kassarjian, B.; Golde, T.E.; Parent, A.T.; Thinakaran, G. Reduced Alzheimer's disease β-amyloid deposition in transgenic mice expressing S-palmitoylation-deficient APH1aL and nicastrin. *J. Neurosci.* **2010**, *30*, 16160–16169. [CrossRef] [PubMed]
13. Pinner, A.L.; Tucholski, J.; Haroutunian, V.; McCullumsmith, R.E.; Meador-Woodruff, J.H. Decreased protein S-palmitoylation in dorsolateral prefrontal cortex in schizophrenia. *Schizophr. Res.* **2016**, *177*, 78–87. [CrossRef] [PubMed]
14. Zaręba-Kozioł, M.; Figiel, I.; Bartkowiak-Kaczmarek, A.; Włodarczyk, J. Insights into protein S-palmitoylation in synaptic plasticity and neurological disorders: Potential and limitations of methods for detection and analysis. *Front. Mol. Neurosci.* **2018**, *11*, 175. [CrossRef] [PubMed]
15. Chen, B.; Zheng, B.; DeRan, M.; Jarugumilli, G.K.; Fu, J.; Brooks, Y.S.; Wu, X. ZDHHC7-mediated S-palmitoylation of Scribble regulates cell polarity. *Nat. Chem. Biol.* **2016**, *12*, 686–693. [CrossRef] [PubMed]
16. De, I.; Sadhukhan, S. Emerging roles of DHHC-mediated protein S-palmitoylation in physiological and pathophysiological context. *Eur. J. Cell Biol.* **2018**, *97*, 319–338. [CrossRef] [PubMed]
17. Greaves, J.; Chamberlain, L.H. DHHC palmitoyl transferases: Substrate interactions and (patho) physiology. *Trends Biochem. Sci.* **2011**, *36*, 245–253. [CrossRef]
18. Zaręba-Kozioł, M.; Bartkowiak-Kaczmarek, A.; Roszkowska, M.; Bijata, K.; Figiel, I.; Halder, A.K.; Kamińska, P.; Müller, F.E.; Basu, S.; Zhang, W. S-Palmitoylation of Synaptic Proteins as a Novel Mechanism Underlying Sex-Dependent Differences in Neuronal Plasticity. *Int. J. Mol. Sci.* **2021**, *22*, 6253. [CrossRef]
19. Gorinski, N.; Wojciechowski, D.; Guseva, D.; Galil, D.A.; Mueller, F.E.; Wirth, A.; Thiemann, S.; Zeug, A.; Schmidt, S.; Zareba-Kozioł, M. DHHC7-mediated palmitoylation of the accessory protein barttin critically regulates the functions of ClC-K chloride channels. *J. Biol. Chem.* **2020**, *295*, 5970–5983. [CrossRef]
20. Zareba-Koziol, M.; Bartkowiak-Kaczmarek, A.; Figiel, I.; Krzystyniak, A.; Wojtowicz, T.; Bijata, M.; Wlodarczyk, J. Stress-induced Changes in the S-palmitoylation and S-nitrosylation of Synaptic Proteins. *Mol. Cell. Proteom.* **2019**, *18*, 1916–1938. [CrossRef]
21. Woodley, K.T.; Collins, M.O. Quantitative analysis of protein S-acylation site dynamics using site-specific acyl-biotin exchange (ssABE). *Methods Mol. Biol.* **2019**, *1977*, 71–82. [PubMed]
22. Basu, S.; Plewczynski, D. AMS 3.0: Prediction of post-translational modifications. *BMC Bioinform.* **2010**, *11*, 1–15. [CrossRef] [PubMed]
23. Wang, D.; Liang, Y.; Xu, D. Capsule network for protein post-translational modification site prediction. *Bioinformatics* **2019**, *35*, 2386–2394. [CrossRef]

24. Wang, D.; Zeng, S.; Xu, C.; Qiu, W.; Liang, Y.; Joshi, T.; Xu, D. MusiteDeep: A deep-learning framework for general an kinase-specific phosphorylation site prediction. *Bioinformatics* **2017**, *33*, 3909–3916. [Cross

25. Ren, J.; Wen, L.; Gao, X.; Jin, C.; Xue, Y.; Yao, X. CSS-Palm 2.0: An updated software for palmitoylation sites prediction. *Protei Eng. Des. Sel.* **2008**, *21*, 639–644. [CrossRef] [PubMed]

26. Xue, Y.; Chen, H.; Jin, C.; Sun, Z.; Yao, X. NBA-Palm: Prediction of palmitoylation site implemented in Naive Bayes algorithm. *BMC Bioinform.* **2006**, *7*, 1–10. [CrossRef]

27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

28. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **2007**, *36*, D202–D205. [CrossRef]

29. Tan, F.; Fu, X.; Zhang, Y.; Bourgeois, A.G. A genetic algorithm-based method for feature subset selection. *Soft Comput.* **2008**, *12*, 111–120. [CrossRef]

30. Wang, D.; Liu, D.; Yuchi, J.; He, F.; Jiang, Y.; Cai, S.; Li, J.; Xu, D. MusiteDeep: A deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res.* **2020**, *48*, W140–W146. [CrossRef]

31. Pejaver, V.; Hsu, W.; Xin, F.; Dunker, A.K.; Uversky, V.N.; Radivojac, P. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci.* **2014**, *23*, 1077–1093. [CrossRef] [PubMed]

32. Wheeler, D.L.; Barrett, T.; Benson, D.A.; Bryant, S.H.; Canese, K.; Chetvernin, V.; Church, D.M.; DiCuccio, M.; Edgar, R.; Federhen, S. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2007**, *36*, D13–D21. [CrossRef] [PubMed]

33. Bairoch, A.; Apweiler, R.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M. The universal protein resource (UniProt). *Nucleic Acids Res.* **2005**, *33*, D154–D159. [CrossRef] [PubMed]

34. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 1–13. [CrossRef] [PubMed]

35. Holland, J.H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; MIT Press: Cambridge, MA, USA, 1992; ISBN 0262581116.

36. Halder, A.K.; Chatterjee, P.; Nasipuri, M.; Plewczynski, D.; Basu, S. 3gClust: Human Protein Cluster Analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *16*, 1773–1784. [CrossRef]

Soumyendu Sekhar Bandyopadhyay
21/9/23

Subhadip Basu.
24.09.2023

Subhadip Basu, Ph.D.
Professor
Computer Sc. & Engg. Department
Jadavpur University
700032