

Integrated Molecular Modelling and Machine Learning Strategies to Investigate Potential Drug Targets Against Poliovirus



By

Rijja Hussain Bokhari

(NUST00000318810-MSBI-Fall19)

(MS Bioinformatics)

Supervised by:

Dr. Ishrat Jabeen

**Research Centre for Modelling and Simulation
National University of Sciences and Technology
Islamabad, Pakistan.**

November 2021

Integrated Molecular Modelling and Machine Learning Strategies to Investigate Potential Drug Targets Against Poliovirus

A thesis submitted in partial fulfilment of the requirement for
the degree of Master's in Bioinformatics



By

Rijja Hussain Bokhari

(NUST00000318810-MSBI-Fall19)

(MS Bioinformatics)

Supervised by:

Dr. Ishrat Jabeen

Research Centre for Modelling and Simulation
National University of Sciences and Technology
Islamabad, Pakistan.

November 2021

Dedication

*I dedicate this work to my parents who inspire me to better
every day.*

Certificate of Originality

I hereby certified that this thesis is based on my own research and struggle. Furthermore, none of its contents have been plagiarized or submitted for a higher degree. Other people's contributions to this project are acknowledged and referenced.

Rijja Hussain Bokhari

(NUST00000318810-MSBI-Fall21)

ACKNOWLEDGMENT

First, I would like to pay my gratitude towards Allah Almighty as it was my faith and believe that reminded me that for every sickness there is cure which help in inclination towards this research. This research journey would not have been possible without constant guidance and support of my Supervisor Dr. Ishrat Jabeen, who taught me about hard work and dedication research work necessitate.

I would like to acknowledge my parents Mutahhar Hussain and Ghazala Parveen, who believe in me, when I was failing to do the same. They not only provided me with the financial aid to acquire this degree but also remain a constant emotional support during the overwhelming research phase. I would like to express thanks to my brothers Sarmad and Osama who taught me to ask questions and search for answers not matter how hard to find.

I am grateful towards my fellow research group fellows Aniq Bokhari, Aiman Raauf, Ammara Naz and Maham Ahmad. They not only supported me during my research but also helped in get back on my feet during research setbacks. I would also like to pay my thanks to my seniors Maria Ehsan and Fatima Iqbal as they share their knowledge and research experience which help me during my research.

Last but not the least, I would like to acknowledge National Institute of Information Technology and Research Centre for Modelling and Simulation department and especially IT staff for providing the infrastructure that made this research possible.

TABLE OF CONTENTS

Acknowledgment	i
List Of Abbreviation	iv
List Of Tables	vi
List Of Figures	vii
Abstract	ix
ChAPTER 1	11
INTRODUCTION	11
1 Poliovirus:	12
1.1 Background of poliovirus:	12
1.2 Pathogenesis:	13
1.3 Replication of Poliovirus:	14
1.4 Poliovirus Proteases:	16
1.5 Research Strategy:.....	17
1.6 Objectives of this study:	17
CHAPTER 2	18
LITERATURE REVIEW	18
2 Previous Targets Against Poliovirus:	19
2.1 Capsid Proteins:.....	19
2.2 RNA polymerase:.....	19
2.3 2B protein:	20
2.4 Proteases as Antiviral Targets:	20
2.5 Poliovirus Proteases:	21
2.5.1 Role of 2A and 3C in taking control of cell machinery:	21
2.5.2 Role of 2A in termination of mRNA transport:.....	22
2.5.3 Termination of transcription:	22
2.5.4 Inducing Cell death:	23
2.5.5 Interaction with Innate Immune system:.....	23
2.6 Inhibitor of 2A protease:.....	24
2.7 Inhibitor of 3C Protease:.....	25
2.8 2A and 3C protease as potential drug targets:	26
2.9 Computational Studies on Poliovirus:	26
CHAPTER 3	29
METHODOLOGY	29
3 Methodology Overview:	30

Structure-Based Method:.....	31
3.1 Data collection:.....	31
3.2 Homology Modelling:.....	38
3.3 Molecular Dynamic Simulation of Homology Model:.....	39
3.4 Molecular Docking:	40
3.5 Pose Analysis:	41
3.6 Molecular Dynamic Simulation of the Docking complexes:	42
Machine Learning Methodology:	43
3.7 Data collection and Pre-processing:.....	43
3.8 Generation of Machine learning Model:.....	43
3.9 Ensemble Methods:.....	45
CHAPTER 4	46
RESULT	46
4 Molecular Modelling Results	47
4.1 Homology Modelling:.....	47
4.2 Molecular Dynamic Simulation of Proteins:	48
4.3 Molecular Docking:	50
4.4 PLIF Analysis:.....	54
4.5 Molecular Dynamic Simulation of Docking Complexes:	56
Machine Learning Model	62
4.6 Decision Tree:	62
4.7 Ensemble Methods:.....	63
CHAPTER 5	65
DISCUSSION	65
Chapter 6	70
Conclusion	70
References	73
Appendix-1	78

List Of Abbreviation

Ala	Alanine
eIF4G	Eukaryotic Translation Initiation Factor 4 G
FP	Total Number of False Positive
GOLD	Genetic Optimization for Ligand Docking
GPEI	Global Polio Eradication Initiative
IPV	Inactive Poliovirus Vaccine
IRES	Internal Ribosome Entry Site
logP	Octanol-Water Partition Coefficient
MAP4	Microtubule Associated Protein 4
MD	Molecular Dynamics
MOE	Molecular Operating Environment
mRNA	Messenger Ribonucleic Acid
MSA	Multiple Sequence Alignment
NPCs	Nuclear Pore Complexes
OPV	Oral Poliovirus Vaccine
ORF	Open Reading Frame
PABP/PAB	Poly(A)-Binding Protein
PARP	Poly (ADP-Ribose) Polymerase
PDB	Protein Data Bank
PLIF	Protein Ligand Interaction Fingerprints
Pro	Protease
PVR	Poliovirus Receptor
RGR	Retinal G Protein

RMSD	Root Mean Square Deviation
rRNA	Ribosomal Ribonucleic Acid
snRNA	Small Nucleic Ribonucleic Acid
TBP	Tata Binding Protein
TN	True Negatives
TP	True Positive
Vpg	Viral Protein Genome

List Of Tables

Table 3.1: Structural data of Inhibitor of 2A protease along with IC ₅₀ values in range of 1-98 μM against 2A and poliovirus collected from literature and ChEMBL.	32
Table 3.2: Chemical structural data of Inhibitors of 3C protease along with IC ₅₀ values against poliovirus and HRV14/16 3C protease collected from literature and ChEMBL database.....	35
Table 4.1 ERRAT score and Ramachandran plot residues at start and end of molecular dynamic simulation of 2A and 3C protease.....	49
Table 4.2 The 2A selected complex binding residues and binding interaction at 0ns and 300ns.....	58
Table 4.3 The binding residue of 3C proteases with selected inhibitors at 0 ns and 50 ns respectively.	61
Table 4.4 Decision tree model accuracy on training and test set.....	63
Table 4.5 Ensemble method accuracy parameters at batch size and test options.....	63

List Of Figures

Figure 1.1 Replication Cycle of Poliovirus inside a host cell.	16
Figure 2.1 Proteolytic cleavage by 2A and 3C at different locations of the single polyprotein. The VP1,VP2,VP3 and VP4 are structural protein while the 2A, 2B, 2C , 3A, 3C 3D and Vpg are nonstructural proteins. 2A protease performs the first cleavage and the rest are cleaved by 3C.	21
Figure 2.2 Multiple proteins that are cleaved to control cell machinery. 2Aprotease cleaves the nuclear pore proteins nup153, nup62, and nup93 and translational protein elf4g1. 3C protease cleaves translational co-factor pabpc1 and elf5b as well as a transcriptional factor.	24
Figure 2.3: The chemical structures of homophthalimides with substitution at R1 and R2.	25
Figure 2.4 Chemical structure of (A) dipeptidyl aldehyde, (B) aldehyde bisulfite adduct salt, and (C) Rupintrivir active against 3C, having IC ₅₀ in the range of 1.7 to 2.0 μM.	25
Figure 3.1 The Overall Methodology of Research. It consists of Data collection of biological and Chemical data. Module-II used biological data and chemical data for homology and docking analysis. Module-II used Viral and Human protease sequence data for machine learning classification model.	30
Figure 3.2 Homology Modeling Steps.	39
Figure 3.3 Molecular Dynamic Simulation Steps	40
Figure 4.1 Sequence alignment of 2A protease of Poliovirus (sp P03300) and enterovirus (4FVB) in T-coffee.	47
Figure 4.2 Ramachandran plot of the 2A protease (a) and 3C protease (b) at 0 ns. Both proteases have the majority of their residues in the favored region (red) and additionally allowed region (yellow) while very few in the generously allowed region (off white) and non in the disallowed region (white).	48
Figure 4.3 Ramachandran plot of 2Aprotease (c) and 3C protease (d) at 500ns and 50 ns respectively. It can be seen that only one residue of 2A protease is in the disallowed region.	49
Figure 4.4 Molecular Dynamic Simulation of 2A Protease for 500 ns. It shows that at 0 ns the minimum RMSD is 0.6 (Å) and at 500 ns it has the RMSD. Add range of RMSD!!	50
Figure 4.5 Molecular Dynamic simulation of 3C Protease for 50ns. It shows that the minimum RMSD is 0.15 nm and at 50 ns it has the maximum RMSD of 0.22 nm. .	50
Figure 4.6 The docking results of 2A protease. Representation of Gold Score on X-axis and biological activity of pIC ₅₀ on Y-axis. The red data points A10(ly3553349), A2(ly3553352), A12(elastase inhibitor II), A9(ly3553348) and A3(iodoacetamide). .	52
Figure 4.7 The correlation plot of biological activity of 2A protease and Molecular weight (g/mol). The green data points A10(ly3553349), A2(ly3553352), A12(elastase inhibitor II), A9(ly3553348) and A3(iodoacetamide).	52

Figure 4.8 The 3C protease docking score Gold Score (x-axis) and PIC50 (y-axis). The red data points C23(CHEMBL4212167), C19(CHEMBL4212167), C6 (Rupintrivir), and C1(CHEMBL4229177)	53
Figure 4.9 3C protease logP(o/w) correlation plot with pIC ₅₀ . The green data points C23(CHEMBL4212167),C19(CHEMBL4212167),C6(Rupintrivir),and C1 (CHEMBL4229177).	54
Figure 4.10 The PLIF analysis of 2A protease. The bars represent the overall percentage abundance of the residues in the interaction with the inhibitor data set. .	55
Figure 4.11 The PLIF analysis of 3C protease. The bars represent the overall percentage abundance of the residues in the interaction with the inhibitor data set. .	55
Figure 4.12 RMSD plot of 2A protease selected complexes for 300ns. The highest fluctuation.	56
Figure 4.13 The interaction of binding residues of 2A proteases before the MD simulation A10(ly3553349), A2(ly353352), A12(elastase inhibitor II), A9(ly355348) and A3(iodoacetamide).....	57
Figure 4.14 The interaction of binding residues of 2A proteases after the MD simulation A10(ly3553349), A2(ly353352), A12(elastase inhibitor II), A9(ly355348) and A3(iodoacetamide).....	57
Figure 4.15 The 3C protease selected complexes RMSD plot for MD simulation of 50ns. The highest RSMD is approximately 0.3nm (3Å) for all the complexes.	60
Figure 4.16 The 3C protease binding residue interaction before MD with C23(CHEMBL4212167), C19(CHEMBL4212167), C6 (Rupintrivir), and C1(CHEMBL4229177).....	60
Figure 4.17 The 3C protease binding residue interaction after MD with C23(CHEMBL4212167), C19(CHEMBL4212167), C6 (Rupintrivir), and C1(CHEMBL4229177).....	61
Figure 4.18 Decision Tree for classification of human and virus sequences.	63

ABSTRACT

Poliovirus is a highly pathogenic virus causing the crippling disease of Poliomyelitis. Poliovirus mostly infects infants with a weak immune system and is still infecting many in developing countries like Pakistan and Afghanistan. Poliovirus virus drug designing efforts might help in the eradication of poliovirus. Poliovirus is a non-enveloped +ssRNA virus that can cause paralytic polio by the chromatolysis of the motor neurons residing in the spinal cord or brain stem. The non-structure proteins of poliovirus are involved in the proteolysis of the single polypeptide into functional proteins. These are the cysteine proteases i.e., 2A and 3C proteases. A previous proof of concept study identifies poliovirus protease 3C and 2A also play a crucial role in the apoptosis of the motor neuron. The 2A and 3C protease initiate apoptosis by caspase-independent and dependent pathways respectively. The 2A protease also cleaves the nuclear pore proteins Nup62, Nup98, and EIF4G1. This blocks the transport of host mRNA required for the viability of cells and results in the nuclear localization of protease 3C. Additionally, 3C protease degrades the DNA and cleaves the poly (ADP-ribose) involve in DNA repair. The 3C protease also cleaves cytoskeletal protein MAP4 and translocates cytochrome c from mitochondria. These morphological changes by the 3C protease induce apoptosis by activation of the caspase pathway. Moreover, protease 2A and 3C cleaves the protein Eukaryotic translation initiation factor 4 G (eIF4G) and Poly(A)-binding protein (PAB or PABP) which terminate the translation of the host cells. This involvement of 2A and 3c proteases makes them a significant target for a drug against poliovirus.

This study aims to explore 2A and 3C protease as a potential drug target against poliovirus. For this purpose, approaches like homology modeling and MD simulation have been used to get a stable molecular structure of proteases. The docking experiments have been used to probe the best binding confirmation of the ligands with the proteases. The MD simulation of some ligand complexes was performed to evaluate the ligand-protein interaction profiles. One of the challenges faced while targeting the viral proteases is the conserved nature of the viral proteases with human proteases. This highly similarity of viral proteases with humans can lead to the off-target toxicity which can be controlled by increasing the specificity of drug towards the viral proteases. In order to identify the unique classification features of viral

proteases we have used machine learning technique of decision tree. Additionally, ensemble methods like the random forest, bagging, and boosting have been used to remove the bias and variance in the data. These strategies identified that hydrogen bonding is the most crucial interaction required for the inhibition of 2A and 3C protease. Moreover, the residues Cys147 and Gln146 have displayed stable interaction in more than one complex of 3C and 2A protease respectively. The machine learning techniques highlights sequence features like the length of the sequence, frequency of proline, and alanine as the most significant feature for the classification of viral protease from human protease. This project explores the poliovirus conserved proteases 2A and 3C as therapeutic targets which might help in antiviral drug development.

CHAPTER 1
INTRODUCTION

1 Poliovirus:

Poliomyelitis is a paralytic disease caused by a positive RNA virus known as Poliovirus. Polioviruses belong to the class of Picornaviruses which are small in size but highly pathogenic viruses. Poliomyelitis is preventable by two types of vaccine i.e., the Oral Poliovirus Vaccine (OPV) and Injectable Poliovirus vaccine (IPV). Although poliomyelitis is eradicated in most of the world, it is declared as endemic in only two countries Pakistan and Afghanistan. According to the Global Polio Eradication Initiative (GPEI), 84 cases and 7 environmental samples of Wild Poliovirus type 1 (WPV1) reported in November 2020 [1]. Currently, only poliomyelitis symptoms are treated through physical and heat therapy. This lack of immunity against poliovirus in developing countries and risk of resurfacing poliovirus after few decades have led towards the efforts for the development of the drug against poliovirus. Polio is an ancient disease that dates back to many centuries before it was recognized as a viral disease.

1.1 Background of poliovirus:

Although poliovirus pre-exists to the Egyptian civilizations, the first well-known report on poliovirus surfaces in 1894 in the United State, Rutland County [2]. In 1905, it was recognized that polio is a contagious disease and can spread from person to person. After a report of many episodes of poliovirus endemic, Ivar Wickman a Swedish scientist describe poliovirus as a contagious disease and termed the non-symptomatic condition of poliovirus as abortive cases [3]. Scientists at Rockefeller Institute for Medical Research in New York identified and reported germicidal substances in polio survived monkey blood in 1910. In 1916, poliovirus cause the death of approximately 6000 people in the United States alone [4]. During the early 1950s, the scientist in Vienna reported that infectious agent of polio is a virus which helped in development towards the oral vaccine.

Hilary Koprowski researched oral poliovirus vaccine, conducted a trial vaccine on 20 children, and demonstrated that not a single patient was affected by poliovirus while all of them developed antibodies against poliovirus [5]. In 1951, the researchers discovered the method of cultivation of poliovirus in kidney tissues of monkeys which help in the mass production of vaccines [6]. In 1954, a vast polio vaccine trial started that enrolled 1.3 million children. In 1952, there was a wave of

increase in poliovirus cases that lead to 57,628 cases in the US, this was control by the awareness of the vaccine. During this, Salk and his teams also started the human trial of the killed-virus polio vaccine which also displayed the production of antibodies against poliovirus. The Salk vaccine was also later registered by the United States [7].

In 1959, Albert Sabin began his work on the less expensive vaccine against poliovirus. He dedicated his research to the Oral poliovirus vaccine [8]. He conducted the trial of his vaccine on 10 million soviet children without any control group. The result proved that Sabin's vaccine produces antibodies faster than the Salk vaccine which makes it very useful in endemic. Soon after the Sabin vaccine was also licensed by the United States. this vaccine replaced the Salk injectable vaccine containing inactive poliovirus for routine vaccination.

In 1988, the world health assembly launched the global polio eradication initiative, which set the goal to eradicate polio globally by 2000 [9]. In 1990 Poliovirus elimination goal was set in America by Pan American Health Organization (PAHO). In 1994, America was certified as polio-free. The developing countries which massive population like India also started their vaccination campaign in 1997. There was a 99% reduction in poliovirus cases from 1988 to 2000 after the launch of the global poliovirus eradication program [10]. In 2002, many countries eradicated poliovirus, only remain in few countries like Afghanistan, Egypt, India, Pakistan, and Nigeria. The poliovirus type 2 was successfully eradicated in 2015 while type 3 was eradicated in 2019.

1.2 Pathogenesis:

Poliovirus is a type of enterovirus; these viruses enter the body through the gastrointestinal system. Poliovirus can enter the body through contaminated food or water. Most of the time virus is secreted out of the body without causing any infection. Poliovirus replicates in the oropharynx and small intestine. This phase of pathogenesis is called the Alimentary phase. The Peyer's patches located at the oropharynx and M cells located in the inner lining of the epithelial cells are the first point of viral replication [11]. Soon the virus enters cervical and mesenteric lymph nodes leading to the Lymphatic phase. The lymphatic system helps the poliovirus to easily enter the blood circulatory system in a state known as viremia. Viremia is an

accelerated replication phase of virus in the blood, the constant viremia results in short period influenza-like symptoms [12]. This may result in major viremia, facilitating the virus to enter its target tissue in the central nervous system. This phase of the pathogenesis is called the neural phase. These are the target tissues of poliovirus, here poliovirus causes the chromatolysis in the motor neuron present at the different organs.

This leads to different types of poliovirus infections characterized by the location of motor neurons. If poliovirus damages the brain layers or meninges it results in meningitis-like symptoms. This is a non-paralytic infection of poliovirus in which the patient recovers in few weeks. If poliovirus damages the brain stem which is the connection of the cerebral cortex with the brain resulting in Bulbar polio [13]. This weakens the cranial nerves, glossopharyngeal nerve, vagus nerve, and accessory nerve. Bulbar polio symptoms are difficulty in breathing, swallowing, and facial weakness. If the spinal cord motor neurons are damaged, causes spinal polio. Spinal polio is the most common type of poliomyelitis, which constitutes 75% of cases of paralytic poliovirus. In this condition patient completely loss the mobility of one of the arms or legs. The third and most chronic type of poliomyelitis is bulbar-spinal polio. It is caused when the motor neurons of the brain stem as well as of the spinal cord are destroyed. In this condition, the patient is bed driven and unable to breathe without a ventilator which results in heart conditions [14]. The replication of poliovirus is important for understanding poliovirus interaction with host cells at a molecular level.

1.3 Replication of Poliovirus:

The 30nm poliovirus structure consists of two main parts. A positive single-stranded RNA genome and an icosahedral capsid protein. The genome has its 3' terminal poly-adenylated and contains three short reading frames. The 5' terminal has a small protein Vpg and contains a long noncoding region and a large Open Reading Frame (ORF). Poliovirus is enclosed in a capsid 60 copies of four small proteins VP1-VP4. The capsid protein plays a major role in the binding of poliovirus with its receptor site CD155. CD155 is commonly known as Poliovirus Receptor (PVR). It is an immunoglobulin-like protein that is located on the membrane of the cell [15].

PVR interacts with the capsid protein VP4 which introduces the conformational changes in the viral capsid. The **Figure 1.1** represents the complete replication cycle of the Poliovirus in the host cell. This helps the poliovirus pore-mediated insertion. Soon after its entry, the poliovirus genome is released in the cytoplasm. The release of the genome activates the Tyrosyl-DNA phosphodiesterase 2 (TDP2) protein which immediately cleaves off the VPg protein from the 5' end in a process called unlinkedase. Now the genome is ready for the translation [16]. The translation of the poliovirus is cap-independent and facilitated by Internal Ribosome Entry Site (IRES). As a result of translation, a polyprotein of 274-kDa is produced that is initially divided into three precursors P1, P2, P3. These precursors are further cleaved into 2A and 3C protein at a different location to produce multiple viral proteins required for the poliovirus life cycle. These non-structural proteins are 2Apro, 2B, 2C, 3A, 3Cpro/3CDpro, 3AB, 3Dpol, and Viral protein genome (Vpg) [17].

After the translation, the replication complexes are localized in the vesicles which are used for the transport between the Golgi bodies and endoplasmic reticulum. The proteins employed during replication are 3D polymerase, 3CD, 2C, and 3A. The 2C and 3A are involved in recruiting the replication complex. While the 3D polymerase and 3CD bind to the poliovirus sequence and use Vpg as a primer to initiate the replication. The replication takes place in two steps. First, the dsRNA is produced which is used to create the (+) RNA. Secondly, multiple copies of (+) RNA are produced. After replication, each copy of (+) RNA is packed into capsid proteins to produce mature poliovirus [18]. Once an abundant number of virions are generated, the viruses leave the cell either by exocytosis or by lysis.

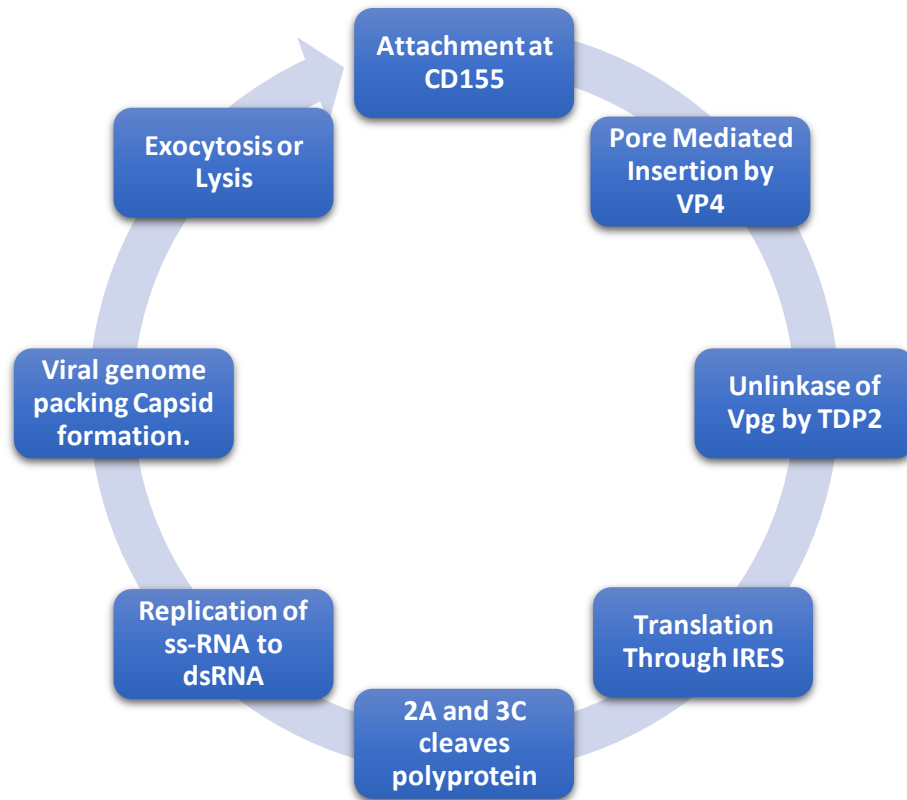


Figure 1.1 Replication Cycle of Poliovirus inside a host cell.

1.4 Poliovirus Proteases:

Many studies have identified that the poliovirus proteases 2A and 3C plays a major role in the chromatolysis of the host cell. 2A and 3C have primarily involved the proteolytic cleavage of polyprotein during translation. Moreover, 2A and 3C protease inhibit the cell translation by cleavage of the Eukaryotic translation initiation factor 4 G (eIF4G) and Poly(A)-binding protein (PAB or PABP). The 2A protease is also responsible for the termination of the transcription by cleavage of Nucleoporin proteins (Nup62, Nup98, and Nup153)[19]. Additionally, 3C protease degrades the DNA and cleaves the poly (ADP-ribose) involve in DNA repair. The 3C protease also cleaves cytoskeletal protein MAP4 and translocates cytochrome c from mitochondria [20]. With the help of an extensive study of literature, we identified that poliovirus protease 2A and 3C are playing a crucial role in the production of all viral proteins and also in the chromatolysis of host cells. Previous drug development studies lack the simultaneous targeting of 2A and 3C protease. Thus, in this study, we aim to employ molecular modeling techniques to identify the

binding interactions of potential drug targets and to module the impact of the 2A and 3C protease on poliovirus treatment.

1.5 Research Strategy:

To target 2A and 3C protease, we have employed a molecular modeling approach to 3D structural features of the respective modulators. Homology modeling of 2A protease was done due to the lack of availability of structure in Protein Data Bank (PDB). After the development, the predicted model was stabilized using molecular dynamics simulation techniques. Docking studies were used to identify the interaction between protein and ligand. The inhibitor data against 2A and 3C protease of poliovirus was collected from literature and different chemical databases. Machine learning technique was also used to differentiate between human and viral protease. This strategy will help us in the modulation of hit against poliovirus.

1.6 Objectives of this study:

Following are the objectives of this research:

- 1) To determine the 3D binding hypothesis of modulators of both 2A and 3C proteases to evade poliovirus replication by molecular modelling.
- 2) To identify the stable binding interaction of 2A and 3C proteases with respective lead molecules.
- 3) To develop a predictive machine learning model to classify viral proteases from human proteases.

Overall by feature identification and by development of classification models of human/virus proteases, we anticipate that this project will aid in???

CHAPTER 2

LITERATURE REVIEW

2 Previous Targets Against Poliovirus:

In past, there have been many efforts to identify the most effective drug target against poliovirus. These attempts were mostly in-vitro studies that used cell-based assay to report the inhibition of certain poliovirus proteins[21]–[23]. As we mentioned in section 1 that from entry in cell to replication many poliovirus proteins are involved. Inhibition many of those proteins have been tried in past but these compounds had certain limitations such as lack of efficacy and poor pharmacokinetics that prohibited their further testing[24]. Following are some poliovirus proteins that were previously studied as a potential drug target:

2.1 Capsid Proteins:

Capsid binding inhibitors have great importance in antiviral research. There are examples of many successful attempts in the development of capsid inhibitors against viruses like HIV and rhinoviruses. Poliovirus replication starts with the entry in the host cell through capsid protein binding to the Poliovirus receptor (PVR /CD155). In the case of poliovirus, capsid protein constitutes four subunits VP1-VP4. The capsid protein VP4 is playing a central role in the binding and pore-mediated insertion of the poliovirus[22]. The capsid inhibitors of other picornaviruses were not effective against poliovirus capsid, hence lacking the specificity. Furthermore, the use of capsid inhibitors for a prolonged time has triggered 3% drug resistance [25].

2.2 RNA polymerase:

One of the most common drug targets against viruses is viral polymerase. Polioviruses contain RNA polymerase also known as 3D polymerase, or RNA-dependent RNA polymerase (RdRP). It performs the poliovirus replication by forming a complex with other proteins like 2C and 3A. It uses a small protein VPg as the primer for viral replication. It performs the replication in three steps binding, initiation, and elongation. It starts the replication processing by binding itself to the 2C (cre) sequence, this helps to break the bond between UMP and VPg. This process is also known as adenylylation. Now, available VPg will be used as a primer and RNA replication will be initiated by the formation of polyA tail. As (+) RNA strand of the poliovirus is used as a template, it results in (-) RNA strand. Consequently, forming dsRNA which is used to produce (+) RNA. The efforts to target the RNA polymerase resulted in some inhibiting compounds identification but it was observed

that inhibitors were able to block the binding and initiation but failed during the elongation phase. Thus, these compounds were ineffective in termination poliovirus replication [26].

2.3 2B protein:

Another poliovirus drug target in previous studies is a viroporin 2B protein. It transforms different host cell membranes and alters multiple cellular functions. In normal condition, the cell maintains the particular balance of Ca^{2+} which regulate the permeability of membranes located at different location like Golgi body, endoplasmic reticulum, nuclear membrane, and cell membrane. 2B protein not only changes the permeability of the cell membrane but also influences autophagy. As soon as 2B is expressed in the cell the Ca^{2+} ions balance is disturbed in the cell. 2B protein decrease the Ca^{2+} ions concentration in Golgi bodies and endoplasmic reticulum and turn increase the Ca^{2+} ions concentration in mitochondria. This results in the release of cytochrome c in mitochondria which may result in activation of the autophagy pathway. Likewise, change in permeability of Golgi bodies membrane also facilitate the formation of viroplasm which is used as replication complex [27]. As protein 2B is highly conserved, it has been suggested to use a 2B gene marker for the diagnosis of poliovirus. The drug development efforts against 2B protein have not been very fruitful as general viroporin inhibitors lack specificity against poliovirus[28]. Moreover, viruses like the Hepatitis C virus and influenza A virus have to develop resistance against viroporin inhibitors, which can be a concern for the development of drug contain viroporin inhibitors.

2.4 Proteases as Antiviral Targets:

In the recent decade research on antiviral has revealed that proteases are extremely important for the viral life cycle. Proteases are not only involved in the proteolytic processing of the structural proteins but are also involved in the cleavage of the host cell proteins. Proteases have been recognized as the efficient drug target due to their conservation in many classes of viruses [29]. This has led to the idea of targeting the proteases for the development of broad-spectrum antiviral drugs. There have been successful attempts in developing the drugs targeting the proteases for deadly diseases like HIV. The drugs include the effective targeting of HIV Reverse

Transcriptase (RT) [30]. Thus, proteases can be seen as a potential drug target in the case of poliovirus as well.

2.5 Poliovirus Proteases:

The proteases of poliovirus 2A and 3C are cysteine proteases that structurally resemble trypsin-like serine proteases. These proteases are hydrolase in nature and have cysteine as nucleophile instead of serine. These proteins contain a catalytic triad and chymotrypsin-like fold. The N-terminus of these proteases contains a zinc-binding domain that is essential for the stability of protease [31]. The foremost and crucial function of poliovirus proteases 2A and 3C is proteolytic processing during translation. As a result of poliovirus genome translation, a single large polypeptide is produced. The first cleavage of this polypeptide is by 2A protease which results in P1 and P2 precursor proteins. Next P2 is successively cleaved by 3C protease to give P3 precursor. P1 is cleaved to produce structural proteins while P2 and P3 are cleaved to produce non-structural proteins as showed in (**Figure 2.1**). 2A protein cleaves the Try | Gly bond and 3C cleaves the Gly | Gln bond. The experimental studies prove that 2A cleavage is highly important as it activates the cleavage of 3C protein[19].

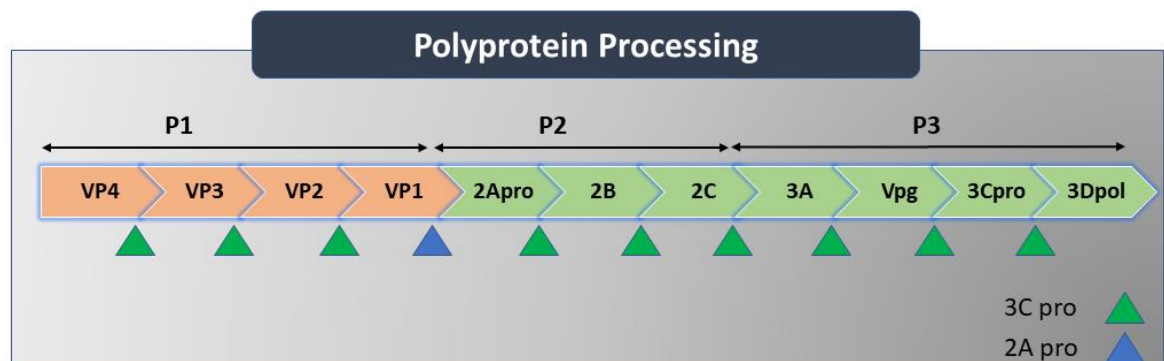


Figure 2.1 Proteolytic cleavage by 2A and 3C at different locations of the single polypeptide. The VP1, VP2, VP3 and VP4 are structural protein while the 2A, 2B, 2C, 3A, 3C 3D and Vpg are nonstructural proteins. 2A protease performs the first cleavage and the rest are cleaved by 3C.

2.5.1 Role of 2A and 3C in taking control of cell machinery:

The 2A and 3C protease are also responsible for taking over the control of all vital host cell machinery. Translational machinery is taken over by the cleavage of some subunits which block the host translation, but viral translation is taking place. 2A protease cleaves the Eukaryotic translation initiation factor 4 (eIF4G1) at positions 681/682 to give cpN (N-terminus fragment) and cpC (C-terminus

fragment). The cpN contains eIF-binding site while cpC contains eIF4A and eIF3 subunit of proteins as shown in (**Figure 2.2**). Earlier it was considered that 2A proteases also completely cleave the eIF4G11 protein as well, but studies showed that eIF4G11 is resistant to the poliovirus infection and only gets partially cleaved [32]. The 3C protease cleaves the eIF5B at position 478/479 which contains the sequence VMEQ|C₄₇₉ at C479 position. This separates the eIF5B N-terminus and its conserved central GTPase and C-terminus [33]. The 3C proteases also cleave the poly(A) binding protein (PABP) which facilitates the eIF5B during translation. The in-vitro studies showed that the 3CD also can cleaves PABP but only 3C protease cleavage is RNA directed [34]. These cleavage result in loss of cap-dependent translation and cell enter in cap-independent translation. This kind of translation is facilitated by Internal Ribosomal Entry Site (IRES).

2.5.2 Role of 2A in termination of mRNA transport:

The 2A protease interferes with the transport of mRNA which is important for the host cell protein synthesis. The transport of RNA outside the nucleus is vital and the initial step of cellular replication. 2A protease blocks the transport of messenger ribonucleic acid (mRNA), ribosomal ribonucleic acid (rRNA), and small nucleic ribonucleic acid (snRNA). This transport is blocked by the cleavage of the Nuclear Pore Complex (NPC) proteins [35]. These proteins include Nup153, Nup64, and Nup98 which are responsible for the traffic between the nucleus and cytoplasm. As a result, host cell proteins that are important for the viability of cells are not produced. Additionally, it also causes cell death by the caspase-independent pathway.

2.5.3 Termination of transcription:

The 3C protease is responsible for the shutdown of the transcription machinery. It cleaves the TATA-Binding Protein (TBP) which is crucial for the transcription which is performed by the DNA-dependent RNA polymerase as shown in Figure 2.2 (A). The 3C protease cleaves the TBP at the position 104/105 breaking glutamine and serine bond [35]. The in-vitro studies revealed that if TBP is not cleaved by the poliovirus it leads to the reduced viral RNA synthesis significantly. Hence the cleavage of TBP is highly important taking over the host cell control and for the sustainability of virus in cells.

2.5.4 Inducing Cell death:

The 3C protease is also responsible for inducing autophagy in infected cells. Apoptosis is a programmed cell death that is non-inflammatory. The 3C protease activates the caspase 3 pathway which results in apoptosis of the cell due to infection. It activates this apoptosis pathway damaging many cellular structures responsible for maintaining the integrity of the cell. The 3C protease cleaves the cytoskeleton protein Microtubule Associated Protein 4 (MAP4)[20]. Moreover, 3C protease upregulates the Bax and releases cytochrome c from the mitochondria. The 3C protease also degrades the DNA and DNA repair enzyme PARP (poly (ADP-ribose) polymerase). This results in the activation of the caspase 3 and 9 pathway which results in cell death.

2.5.5 Interaction with Innate Immune system:

Another essential function carried by 2A and 3C is interaction with the innate immune system. When poliovirus RNA is exposed in the cytoplasm it is recognized by two proteins Retinal G Protein Receptor (RGR) pathway. There are two major protein units of this pathway i.e., Retinoic acid-inducible gene-I-like (RIG-like) receptors and Mitochondrial antiviral-signaling protein (MAVS). These proteins produce the type 9I IFN β (Interferon Beta) by activation of another protein Melanoma differentiation-associated Protein 5 (MDA5). The poliovirus protease 3C cleaves RIG-1 whereas 2A cleaves the MAVS and MAD5 which inhibits the production of type I IFN β [36]. This delays the establishment of the antiviral state as shown in (Figure 2.2 B). As a result, the adaptive immune system is unable to recognize the infected cell. Moreover, the low production of type I IFN β on the surface of the host cell also triggers the apoptosis pathway[37].

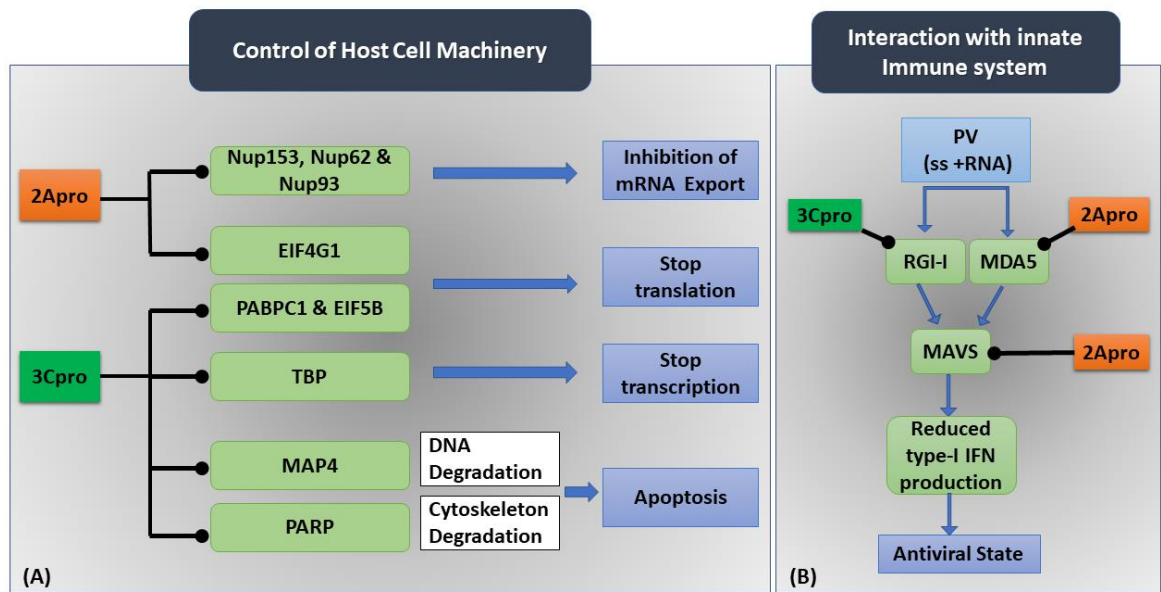


Figure 2.2 Multiple proteins that are cleaved to control cell machinery. 2Aprotease cleaves the nuclear pore proteins nup153, nup62, and nup93 and translational protein elf4g1. 3C protease cleaves translational co-factor pabpc1 and elf5b as well as a transcriptional factor.

2.6 Inhibitor of 2A protease:

There are only a few studies that have identified the 2A protease inhibitor. In vitro studies have identified the Elastase-Specific Inhibitors against poliovirus 2A. These inhibitors include Elastase Inhibitor III, Calpaininhibitor1, Iodoacetamide, and methoxysuccinyl-Ala-Ala-Pro-Val-chloromethyl ketone (MPCMK). Poliovirus 2A protease contains the active site thiol group and is inhibited by alkaline agents like Iodoacetamide. 2A protease also has high similarity with trypsin-like serine proteases so elastase-specific inhibitors like Calpaininhibitor1, Elastase Inhibitor III, and MPCMK have also shown inhibition in in-vitro studies[38].

As inhibition data against poliovirus is in literature is limited so, a similarity-based approach was used to obtain additional data. The Human Rhinoviruses (HRV) belong to the same class as polioviruses and also HRV inhibitors have shown activity against poliovirus as well. A biological essay study showed that homophthalimides dual inhibition against 2A and 3C protease. This study investigated multiple derivatives of homophthalimides and reported their inhibition activity against 2A and 3C protease[38]. As the IC_{50} values against 2A protease were in range of 3-98 μ M than 3C protease as IC_{50} values were in range of >200 μ M so this data was also considered for inhibition of 2A protease.

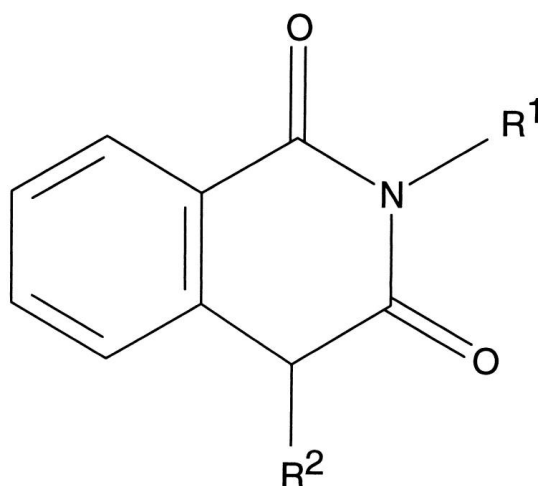


Figure 2.3: The chemical structures of homophthalimides with substitution at R¹ and R².

2.7 Inhibitor of 3C Protease:

A study reported the inhibition of 3C protease by dipeptidyl inhibitors. This study investigated the derivatives of dipeptidyl against different picornaviruses including poliovirus and reported their IC₅₀ values in range of 1.7 to 2.0 μM [39]. It also investigated a broad-spectrum Rhinoviruses 3C protease inhibitor against poliovirus i.e., Rupintrivir. The chemical structures of these inhibitors are given in **Figure 2.4**.

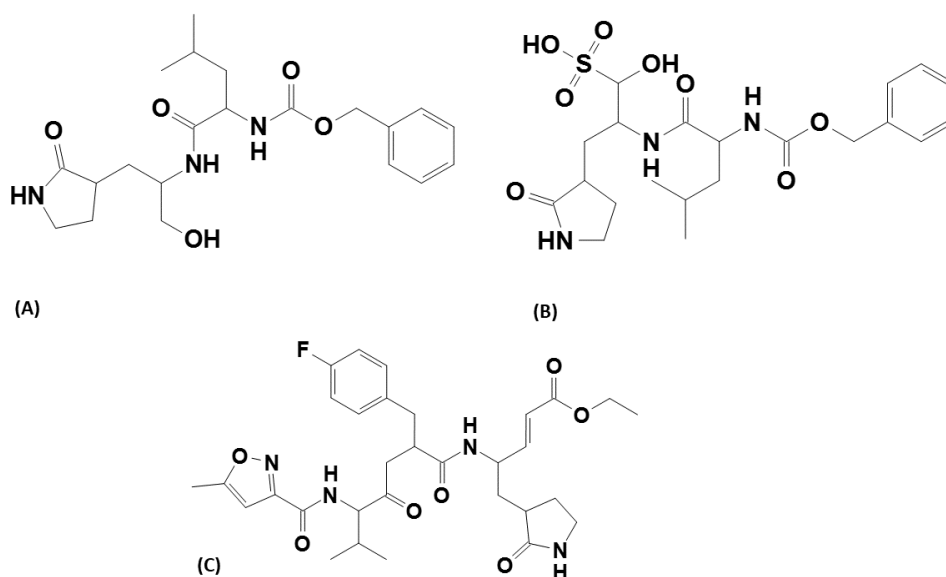


Figure 2.4 Chemical structure of (A) dipeptidyl aldehyde, (B) aldehyde bisulfite adduct salt, and (C) Rupintrivir active against 3C, having IC₅₀ in the range of 1.7 to 2.0 μM.

Moreover, similarity bases approached also used to identify further data of inhibitors. The 3C protease of Human Rhinoviruses has also shown activity against poliovirus protease so HRV inhibitor data was also considered. The inhibitor data of human rhinoviruses were collected from the chemble database, which has data of two types of HRV strains i.e., HRV14 and HRV16. Both of these strain data were collected and used against 3C protease.

2.8 2A and 3C protease as potential drug targets:

Hence, proteases are involved in more than one crucial function for the replication of poliovirus in a host which makes them suitable for drug targets for the development of antivirals against poliovirus. To prevent poliovirus replication, it is important to stop the cleavage of the single polyprotein by 2A and 3C to produce multiple viral proteins. As it is the first and most essential step for viral replication and terminating of the proteolysis by 2A and 3C will terminate the consequential steps. The in-silico studies on 2A protease are very limited as its crystallographic structure lack in Protein Data Bank (PDB). Most of the studies involve the use of mutation data to identify the important binding residue in 2A protease. There are docking studies on 3C protease to compare the binding properties of poliovirus and other picornaviruses.

In this study, we aim to identify the potential inhibitors of 2A and 3C protease using molecular modeling techniques. The known data of inhibitors of 2A and 3C protease is used in this study. We started by homology modeling of 2A protease. The inhibitor data discussed previously was used in docking studies to identify the interaction pattern of the 2A and 3C protease with their inhibitors.

2.9 Computational Studies on Poliovirus:

Although the majority of work on poliovirus protein is based on the biological assay, various structural-based studies have also been conducted on different proteins of poliovirus. Most of these studies have employed docking methodology to understand the interaction of poliovirus proteins and some studies have also provided a comparative analysis of poliovirus proteins with other viral proteins

In [40] Balasubramanian and Smriti Chawla compared the poliovirus nonstructural proteins interaction with Dengue virus non-structural proteins against the *Mentha arvensis* leaves compounds. They investigated the four ligands (

2-Cyclohexene,1,One-2-Methyl-5-(1-Methylethenyl), enzaldehyde,2-Hydroxy-6-Methyl, 2-(2-Hydroxy-2-Phenylethyl)-3,5,6-Trimethylpyrazine, and 3,7,11,15-Tetramethyl-2-Hexadecen-1-ol) extracted from the *Mentha arvensis* leaves (Tetramethyl-2-hexadecane) against the non-structural proteins of poliovirus and Dengue virus.

The binding affinities of these compounds were compared with proteins using docking techniques and it was found that ligand 3,7,11,15-Tetramethyl-2-Hexadecen-1-ol has the best binding affinity with Dengue and Poliovirus non-structural protein. In [41] S. Kashetty et al. studied the poliovirus receptor proteins. The human poliovirus-related protein-2 isoform alpha is a receptor of poliovirus that is not very well studied due to the lack of crystal lattice structure. This study used homology modeling and then used docking studies of 30 camptothecin with poliovirus CD155 receptor protein. It was concluded that poliovirus protein 2 alpha isophorm is one of the significant target that can be inhibited by camptothecin derivative inhibitors and certain interactions with residues like Arg72, Ser77 and Thr86 were found highly significant.

In [42] A.Yonus et.al. investigated the important binding residue for the activity of poliovirus protease 2A with the help of docking studies. In this study, homology modeling was used to predict the structure of 2A protease and single nucleotide polymorphism was used to identify the change in the binding affinity of the poliovirus 2A protease. This study revealed valuable information that binding residue Lys 15, His 20, Cys 55, Cys 57, Cys 64, Asp 108, Cys 109, and Gly 110 are highly significant drug binding residues.

In [39] Y.kim et.al, conducted a comparative study of 3C poliovirus and 3C-like protease of noroviruses and coronaviruses respectively. This study reported the x-ray crystallographic structure of proteins. It also visualized the co-crystallized compounds to identify the difference in binding patterns of poliovirus proteins with noroviruses and coronaviruses against the same inhibitor. There are limited *in-silico* studies on the poliovirus protease due to the lack of structural information. As mentioned previously most of the studies have not studied 2A and 3C protease simultaneously. Both proteases are important for the life cycle of poliovirus thus target both 2A and 3C is important for the termination of the poliovirus replication. Moreover, as mentioned in the previous section 2A protease activates the 3C protease cleavage during the polyprotein process and 2A is activated only after the

cleavage of 3C protease. The current research study has developed a homology model for 2A protease and used the docking technique to identify the binding pattern of above mention inhibitor with 2A and 3c protease. Moreover, this study will also develop a machine learning model to distinguish viral protease from human protease to identify unique features of viral proteases.

CHAPTER 3

METHODOLOGY

3 Methodology Overview:

The overall methodology of this research study is given in (Figure 3.1). The research methodology of this project is divided into two parts one is a structure-based method for understanding the binding pattern of 2A and 3C protease with their respective inhibitors. This will help us identify the significant binding residues of proteins and 3D features of ligands. The structure-based method started by collecting data of 2A and 3C proteases and inhibitor datasets of both these proteins. Then homology modeling was used to develop the 3D model of 2A protease. After modeling molecular dynamic simulation was used to stabilize the protease structures. Once we have the predicted structure of 2A protease and structure of 3C protease from PDB, docking was performed to identify the binding pockets of 2A and 3C protease. Pose analysis was performed after docking to select the best binding pose along with PLIF analysis. A molecular dynamic simulation was employed to stabilize the docking complexes of both proteases. The second is machine learning-based methods that employed the sequence information of human and viral proteases and train the machine learning model to differentiate respective proteases. This will help in the identification of unique classifying features of viral proteases that can be used to increase the specificity of drugs in the future.

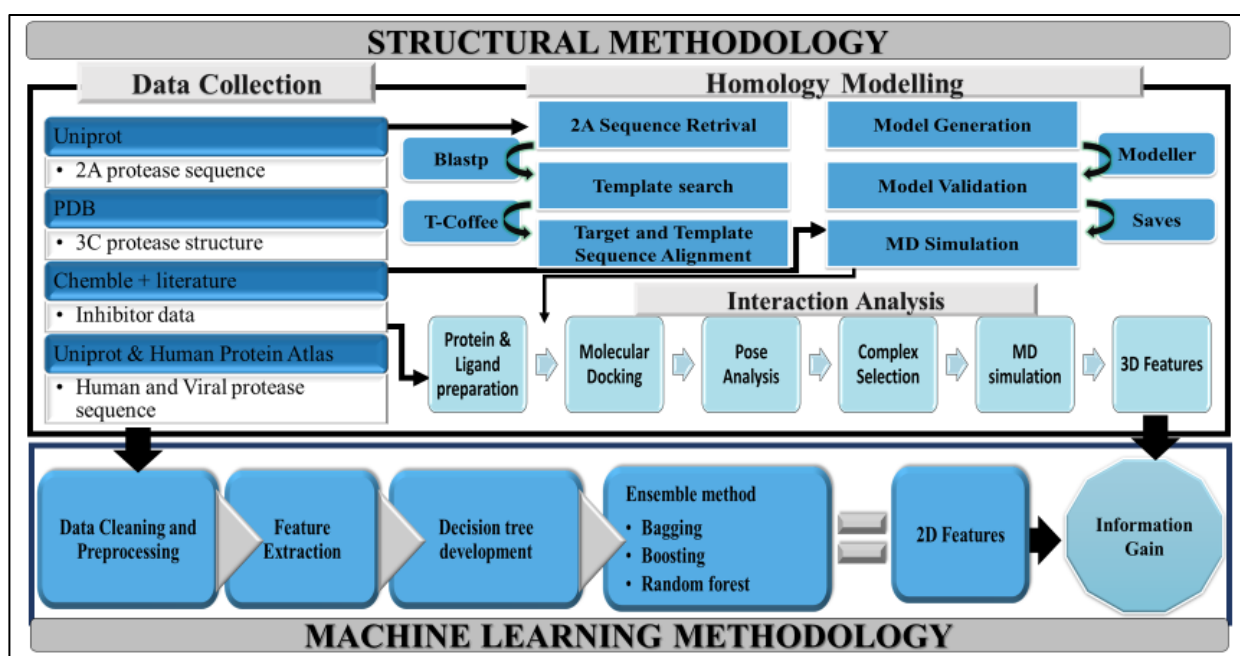


Figure 3.1 The Overall Methodology of Research. It consists of Data collection of biological and Chemical data. Module-II used biological data and chemical data for homology and

docking analysis. Module-II used Viral and Human protease sequence data for machine learning classification model.

Structure-Based Method:

The structure-based methodology is used in drug discovery research when structural information about protein is present and structural data of ligand is limited. This method aid in the recognition of unique binding features of the protein, which in turn help in the identification of potential drug compounds for a protein. The structure-based methodology is used in this research to identify the interaction pattern of a protein with respective ligands. This will facilitate us to construct a binding hypothesis for both 2A and 3C protease. For this purpose, the first homology model of 2A protease is built, and then a docking experiment is performed for both 2A and 3C protease.

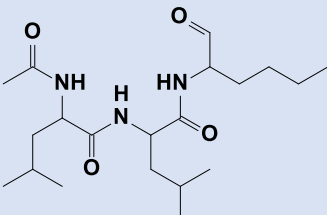
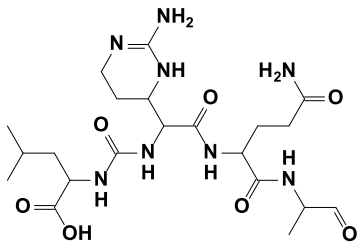
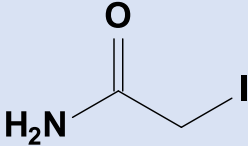
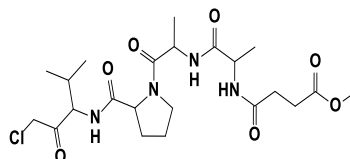
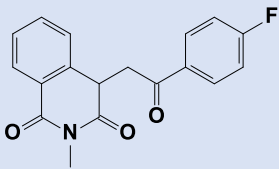
3.1 Data collection:

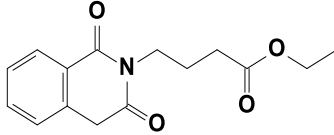
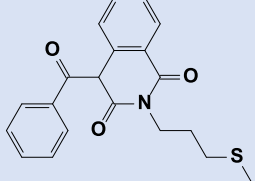
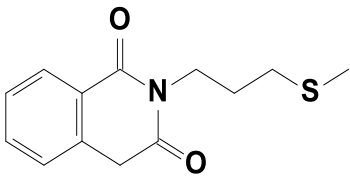
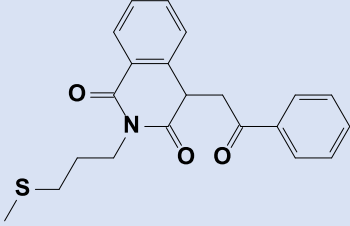
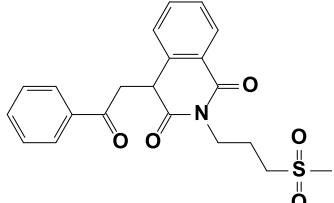
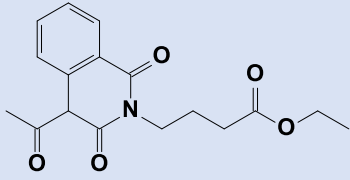
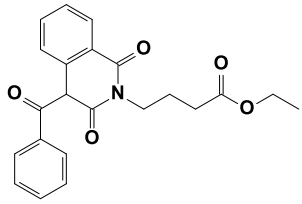
There were two types of protein data collected in structure-based methodology. The PDB (Protein Data Bank) has structural information of 3C protease. The PDB ID **4DCD** was the 3C protease structure titled “1.6A resolution structure of Poliovirus 3C Protease Containing a covalently bound dipeptidyl inhibitor”[39]. The structure was cleaned in MOE software by removing the attached ligand and all other chains except chain A. The molecular structure of 2A protease was not present in the PDB database. For the development of its homology model amino acid sequence of 2A protease was extracted from Uniprot data base against the ID **P03300**, i.e., Poliovirus type1 Mahoney strain[43].

Secondly, we collected data on poliovirus inhibitors from multiple sources. The inhibitor data against the 2A protease of poliovirus was derived from the literature that gave us the elastase-specific inhibitors. By employing a similarity-based approach it was identified that Human Rhinovirus Inhibitors homophthalimides can be used against poliovirus 2A protease. These along with other inhibitors used in this research are given in (**Table 3.1**). The inhibitor data against 3C protease was also collected from the literature. As mentioned in the chapter 2 this data was very limited. That is why a similarity-based approach was used which led to the inhibitor data of 3C protease of Human Rhinoviruses (HRV). This data was collected from the Chemble database. Chemble contains the chemical

data of different compounds and their IC_{50} against different protein targets. **Table 3.2** represents the complete inhibitor data against 3C protease that was used in this study.

Table 3.1: Structural data of Inhibitor of 2A protease along with IC_{50} values in range of 1-98 μM against 2A and poliovirus collected from literature and Chemble.

Code	Chemical Structure	Chemical Name	Viruses	IC_{50} (μM)
A1		Calpaininhibitor1	Poliovirus	28
A2		Elastase Inhibitor III	Poliovirus	4
A3		Iodoacetamide	Poliovirus	1200
A4		MPCMK	Poliovirus	20
A5		LY046601	HRV2	21.8

A6		LY343813	HRV14	18.2
A7		LY343814	HRV14	19.7
A8		LY344453	HRV14	98.3
A9		LY353348	HRV2	8.4
A10		LY353349	HRV2	3.1
A11		LY353350	HRV14	26.5
A12		LY353352	HRV2	3.9

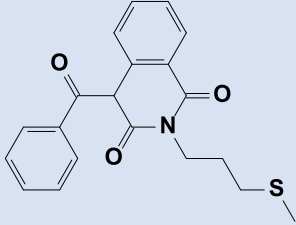
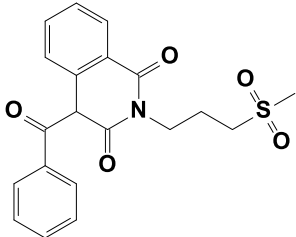
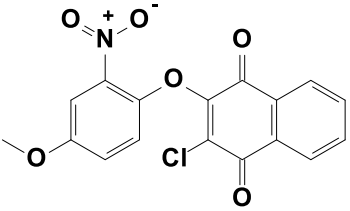
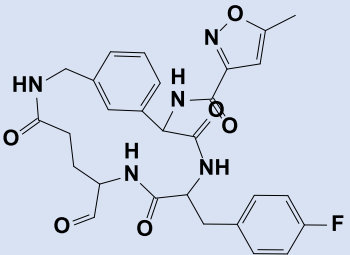
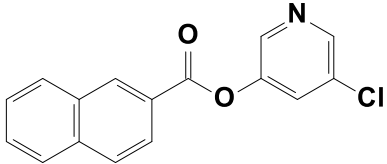
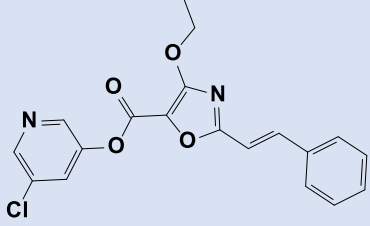
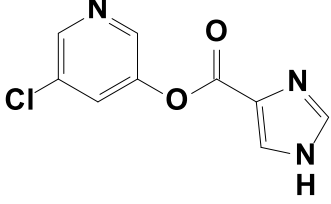
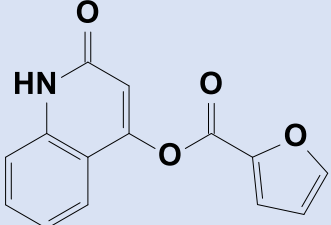
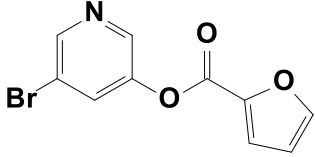
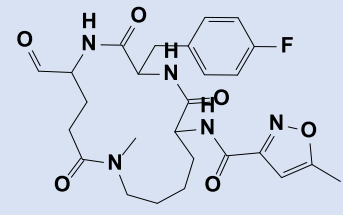
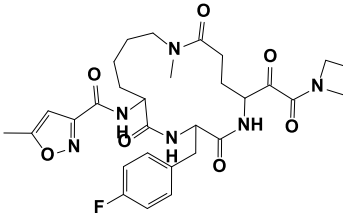
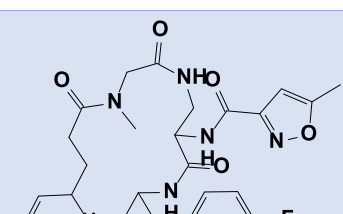
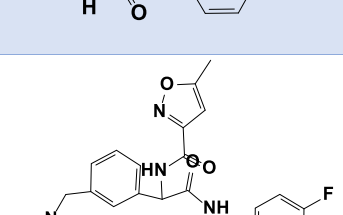
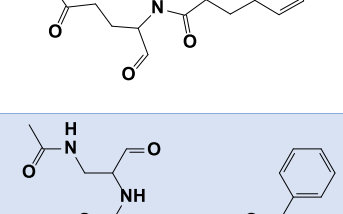
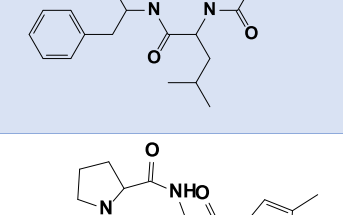
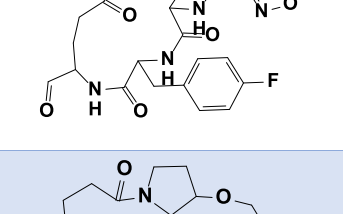
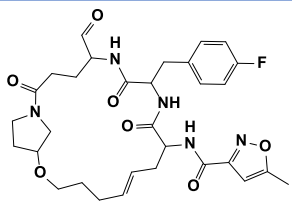
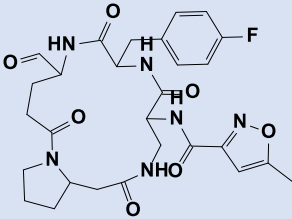
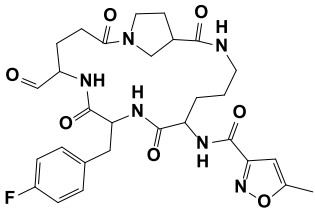
A13		LY353353	HRV2	23.1
A14		LY353354	HRV2	31.2

Table 3.2: Chemical structural data of Inhibitors of 3C protease along with IC₅₀ values against poliovirus and HRV14/16 3C protease collected from literature and Chemble database.

Code	Chemical Structure	Chemble ID	Virus	IC ₅₀ (μ M)
C1		CHEMBL4229177	HRV	8.4
C2		CHEMBL4202850	HRV14	7.37
C3		CHEMBL4207200	HRV14	3.84
C4		CHEMBL4206628	HRV14	3.2
C5		GC373	Poliovirus	2.02
C6		CHEMBL20210 (Rupintrivir)	Poliovirus	1.83
C7		GC376	Poliovirus	1.77

C8		CHEMBL4224853	HRV14	0.85
C9		CHEMBL4207281	HRV14	0.753
C10		CHEMBL466351	HRV16	0.71
C11		CHEMBL496214	HRV16	0.69
C12		CHEMBL496709	HRV16	0.29
C13		CHEMBL506171	HRV16	0.20
C14		CHEMBL222234	HRV16	0.08

C15		CHEMBL4211003	HRV14	0.063
C16		CHEMBL4203502	HRV14	0.035
C17		CHEMBL4217566	HRV14	0.018
C18		CHEMBL4212963	HRV14	0.008
C19		CHEMBL4215964	HRV14	0.007
C20		CHEMBL4207862	HRV14	0.006
C21		CHEMBL4218384	HRV14	0.003

C22		CHEMBL4216095	HRV14	0.002
C23		CHEMBL4212167	HRV14	0.002
C24		CHEMBL4202932	HRV14	0.002

3.2 Homology Modelling:

Homology modeling is an approach used to predict the 3D structure from protein sequence. The homology modeling steps are described in the (**Figure 3.2**). If the target sequence has a similarity of greater than 30% with a known structure protein, then it can be used as a template for homology modeling. As 2A protease crystalized structure is not present in the PDB (Protein Data Bank), the development of its homology model was the first step in its investigation. For this purpose, the 2A protease sequence was extracted from the Uniprot database. Uniprot is a vast database of protein sequences that contain all the information related to protein sequence. Sequence from Uniprot ID “**P03300**” was used, it is a sequence for the complete genome polyprotein of poliovirus type 1 Mahoney strain[43]. The sequence similarity search was performed using the Blastp server. Thus, the Enterovirus A71 was selected as a template for the 2A protease of poliovirus having an identity of 58%. The sequence of 2A protease of poliovirus and 2A protease of Enterovirus A71 was aligned using a T-coffee server [44]. The alignment resulted in a generation of PIR and ALI files. The PDB file of Enterovirus A71 was downloaded and removed for other chains. These files along with the sequence file of the 2A protease of poliovirus were used by Modeller 10.0 to develop a homology model of the 2A protease[45]. There were 100 models generated in total. Each model was validated

using the ERRAT score and Ramachandran Plot. ERRAT score describes the verification of a protein structure based on the atom interaction. It considers the patterns of nonbonded atom's interaction to identify the incorrectly predicted regions in the protein. Ramachandran Plot validates the protein structure based on steric clashes between the atoms of psi and phi bonds. It classifies the residues based on distance in three categories allowed, generously allowed, and disallowed. The top three models according to the ERRAT score and Ramachandran plot were energy minimized. After energy minimization, the model with the top ERRAT score and no residue in the disallowed region was selected for the next step.

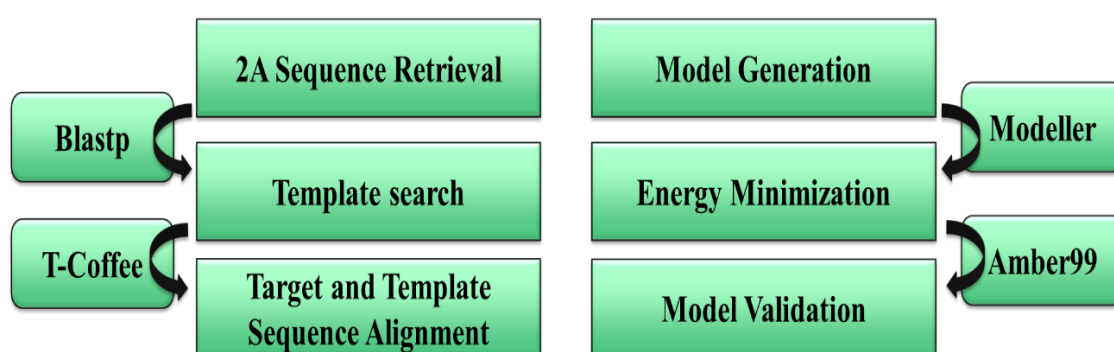


Figure 3.2 Homology Modeling Steps.

3.3 Molecular Dynamic Simulation of Homology Model:

A stable protein structure will lead to better binding interactions, therefore 2A proteases homology model and 3C protease structure was stabilized using Molecular Dynamic (MD) simulation. MD simulation uses Newton's law of motion to monitor the movement of atoms and molecules for the specified time interval usually in nanoseconds to check the stability of molecular interactions. MD simulation steps include preparation of structure (optimization and minimization), Periodic boundary conditions (selection of force fields, shape, and size of boundary box), Solvation (Addition of ions), and Energy minimization of the system and MD production (for the specified time). SCHRODINGER software was used to run MD simulations. SCHRODINGER is a GUI-based software that uses modules like Desmond and Maestro to run MD simulations[46]. The structure was prepared using a preparation wizard which optimizes and minimizes the structure at pH 7.4. Then cubic grid box with 5Å of the area was selected and OPLS (Optimized Potentials for Liquid Simulations) was selected as the force field. Then solvation was performed using Na ions. After the addition of ions again system energy was minimized. **Figure 3.3**

Molecular Dynamic Simulation Steps Now energy minimized system was ready for the simulation. For this research 2A, the protease model was simulated for the different intervals including 50ns, 100ns, 150ns, and 500ns. It was found that the structure was most stable at 500ns. Similarly, 3C protease was also simulated for 50ns. After, molecular dynamics simulation both protease structures were cleaned by removing the solvents molecules, and the structure file was saved in PDB format. This file is now ready for the docking experiment in the next step.

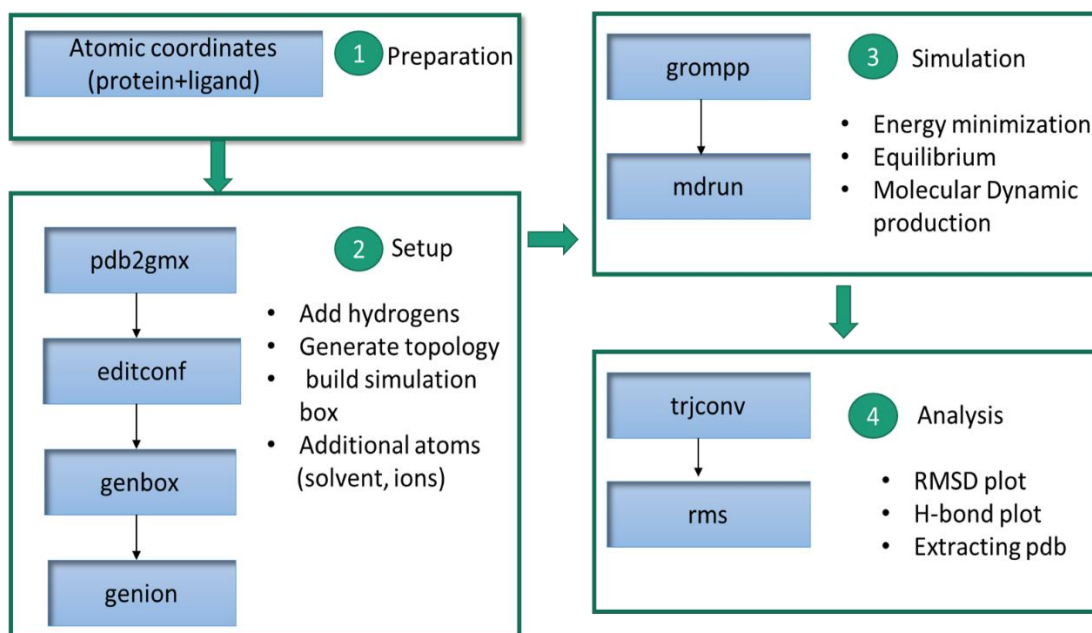


Figure 3.3 Molecular Dynamic Simulation Steps

3.4 Molecular Docking:

Docking studies help us to calculate the desired conformation of the ligand when bound to protein to produce a stable complex. The stability of a complex is calculated using different scoring functions based on different physical properties. For this research, we have used GOLD software and the Gold Score fitness function[47]. Gold Score is a sum of internal hydrogen bond, external hydrogen bond, external van der Waal energy, internal van der Waal energy, and internal torsion as given in the formula below. During the docking experiment, we performed docking of both 2A and 3C side by side.

$$\text{GoldScore} = \text{internal } H_bond + \text{external } H_bond + \text{internal } vdW + \text{external } vdW + \text{internal } tor$$

Firstly, we docked 2A protease with its fourteen ligands as given in Table 3.1. To perform docking protein was uploaded in the GOLD software and its protonation was performed which added only 1 hydrogen atom. The binding site was defined by setting the coordinates as follows $x=10.8070$, $y=10.7321$, and $z= 25.6070$ from the centroid point, 15 Å of the area around the coordinates was selected for binding cavity formation. It was considered that coordinates must include residues Gly 1, Lys 15, Cys 17, His 20, Cys 55, Cys 57, Cys 64, Asp 108, Cys109, Gly 110, Cys 115, and His 117 as their significance has been reported in literature[42]. Gold Score was used with an early termination option with default parameters was used which terminate the pose generation for a ligand if the top 3 solutions are within 1.5 Å. Maximum 100 poses were generated for each ligand.

Similarly, 3C protease docking was also performed in GOLD software. 3C protease structure was protonated by the addition of 22 hydrogen atoms. The binding site coordinates used in this study were $x=1.8070$, $y=10.7321$, and $z=20.6070$ from the centeroid point. The first 15 Å of the area was selected which was found sufficient to produce 100 poses. The coordinates were selected considering that residues Val162, Cys147, His161, Leu127, His40, Gly128, Thr142, and Gly164 are present in the selected region. These residues have been reported for interactions in the literature[39]. Gold Score was used with the early termination option enabled. After all the ligand's docking solution was completed for both proteins, it was examined for pose analysis.

3.5 Pose Analysis:

Docking resulted in different docked poses of inhibitors within poliovirus 2A and 3C proteases. In the case of 2A protease, we adopted the strategy of correlating the docking score with the pIC_{50} value. The pIC_{50} represents the activity of the inhibitor against a particular target. pIC_{50} was calculated from the IC_{50} given in **Table 3.1** using the following formula.

$$pIC_{50} = 1/LOG10[IC_{50}(M)]$$

OR

$$pIC_{50} = -[LOG10(IC_{50}(M))]$$

pIC_{50} was plotted against the Gold fitness score to . Gold fitness score represents the binding energy. Secondly, correlation of molecular weight of inhibitors with pIC_{50} for 2A protease was also studied as it represents the biological activity due to the transport of ligand towards the target. For 3C protease LogP (o/w) was studied in correlation with pIC_{50} . LogP (o/w) is the partition coefficient of the between octanol and water also known as Lipophilicity. It is a physicochemical feature describe the solubility of given substance in fat. As to reach the target a drug compound has to cross many lipophilic and hydrophobic barrier, but the high lipophilicity can dissolve the drug hence low activity.

Moreover, Protein Ligand Interaction Fingerprints (PLIF) analysis is a method that can help in the identification of the significant interacting residues of the protein for a data set of inhibitors. The PLIF analysis was performed on the docking results of 2A and 3C proteases. The PLIF analysis represents the interaction type and abundance percentage of interacting residue with inhibitor data set. It uses the bit score to represents the following type of interactions Surface contact (Surf), side chain doner (ChDon), side chain acceptor (ChAcc), Backbone doner (BkDon), backbone acceptor (BkDon) and Hydrophobic interaction (Arene). PLIF uses the graphical of the residues with the percentage of abundance in the interaction.

3.6 Molecular Dynamic Simulation of the Docking complexes:

The stability of the complexes generated through the docking was examined using the molecular dynamic simulation. GROMACS 2019.6 is molecular dynamic simulation software that was used to evaluate the stability of binding residues and hydrogen bonds of the selected docked complexes by the help of pose analysis[48]. The Charmm36 all-atom force field was used to develop the topology of ligand compounds and protein. The solvation was performed using SPC216 water model in a periodic box, neutralized by the addition of Na^+ and Cl^- . Initially energy minimization was completed using steepest descent minimization algorithm for 500 steps and tolerance of $1000J/\text{\AA}$ to eliminate the steric clashes of the complexes. Equilibrium of energy minimization was performed under constant temperature and pressure for 1000ps (default value). The Molecular Dynamic simulation runs were accomplished using the Berendsen thermostat and barostat under constant temperature 300K and pressure 1 atm. The fast smooth Particle-Mesh Ewald (PME) summation were employed for the Long-range electrostatic interactions using the

cut-of value of 1 nm for the direct interactions. The MD runs were analyzed for different time period of 50ns, 100ns, 200ns and 300ns. It was found that complexes were most stable at 300ns by the help of RMSD plots.

Machine Learning Methodology:

3.7 Data collection and Pre-processing:

Data collection is the first and most important step for the development of machine learning models. As humans have a high number of proteases consisting of different types, it was a better approach to collect only human cysteine proteases sequences as 2A and 3C are also cysteines in nature. There were two types of data collected sequence data of all viral 2A and 3C protease and Human cysteine protease. In the case of viral protease sequence similarity search using blast was performed which resulted in 42 sequences of 2A protease and 44 sequences of 3C protease from different viruses given in appendix-1 . The resulting file was also provided with a uniprot ID of each sequence which was used to extract each sequence fasta file separately. In the case of human proteases data was collected from The Human Protein Atlas version 20.1. This database contains all the information about the protein present in humans. This database provides the human proteases ID in Uniprot. This information was used to extract sequences from uniprot manually. Once all the viral and human sequences were collected. A python script was used to count each amino acid in all sequences separately. The frequency of each amino acid (Alanine, Arginine, Asparagine ,Cysteine, Glutamine, Glycine, Histidine, Isoleucine, Leucine, Lysine, Methionine, Phenylalanine, Proline, Serine, Threonine, Tryptophan, Tyrosine, Valine) was calculated along with the length of each sequence. The output file comprises of 21 attributes, label class and 240 entries. This file data was randomized by the help of MOE and divided into training and test data set of 80 and 20 respectively.

3.8 Generation of Machine learning Model:

Machine Learning is a modern technique that identifies the data patterns used to classify the data into desired classes. Here in this research, the aim is to classify human sequences from virus sequences by the help of attributes like sequence length and amino acid frequency as described in previous section. There are different types of machine learning models that can be applied to data depending upon the data type

of attributes as well as of classification label. The preprocessed data has 21 attributes, all of these attribute variables are quantitative variables. The variable of sequence length is discrete as it contains the number of amino acids in a given sequence, whereas the other 20 variables are continuous variables as they describe the frequency of a particular amino acid in the a given sequence. The class label is a qualitative, binomial variable, containing two values human and virus. This information helps us in deciding the Machine Learning Model.

For the purpose of this study, the Decision Tree Model is used, as the study data is extracted from 2D data of sequences and has binary class label. To build Decision Tree Model a GUI based software Weka 3.8.4 was used[49]. Weka is software that has the collection of different machine learning algorithms for data analysis. It provides with the tools for the data preprocessing, data cleaning, data preprocessing and data clustering. As described in previous section training data set of 80% data i.e., 192 instances were used to build the decision tree and tested on 20% data i.e., 48 instances. Decision Tree classifier was used to model the model using J48 algorithm. This algorithm is built on J R Quinlan C4.5 algorithm which is the implementation of the ID3 algorithm. this algorithm builds the tree on the basis of information gain to select the most informative attribute from the dataset. The training model and testing model was build using the default parameters. The decision tree performance is measured by the help of its accuracy, sensitivity and specificity values of training and testing model as given below.

$$\text{Accuracy (\%)} = \frac{(TP+TN)}{TP+TN+FP+FN} \times 100$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

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

3.9 Ensemble Methods:

Ensemble methods can be used to validate the accuracy of any machine learning models. Machine learning accuracy can also be high due to the variance in the training data which means that it is possible that if the same model is used for different data, it will give low accuracy. To validate decision tree, ensemble methods are used. Ensemble method mainly divide the data in different equal size of samples, trains the model on each sample and use average or voting to get the final model attributes.

In this study, we have used only one machine learning model i.e., Decision tree as it is one of the best approach visual and easy to understand classification model. In order to check the efficiency of the decision, three ensemble methods were used i.e., bagging, boosting, and random forest. Bagging is an ensemble method that uses the different model that are trained on different samples from the same dataset. These multiple models are combined by average or voting. Secondly, boosting was also used as ensemble method, which start with a base learner and move onto next classifier for the values that are predicted wrong by the first one, it goes on until reach the maximum accuracy. Lastly, random forest is one of the best of ensemble method to increase the accuracy of decision tree as it uses the multiple decision trees to train of different samples and then uses the bagging for the final class label prediction. This is study, all ensemble methods were tested on batch size 30, 50, 100 and number of iteration 50 and 100 and test options were 10 cross fold validation and 80% split test. For each parameter statistics are reported in results section.

CHAPTER 4

RESULT

4 Molecular Modelling Results

4.1 Homology Modelling:

The homology model of the 2A protease of poliovirus is the first step in the molecular modeling of the 2A protease. The homology model was build using the crystal structure of the 2A protease of Enterovirus A71. This structure of 2A protease was previously developed using the X-ray diffraction method and has a resolution of 1.90 Å. The 2A protease of enterovirus and poliovirus displayed the sequence similarity score of 99 in the T-coffee server. The sequence alignment in T-coffee is given in **(Figure 4.1)**. Using this alignment 100 homology models of 2A protease were developed in Modeler 10.0. Each model was validated using the ERRAT score and Ramachandran plot. The best model was selected with the ERRAT score of 76.5 and having 69.8% favored and 30.5% residues in the additionally allowed region was selected as the best model detail are given in table **(Table 4.1)**. The Ramachandran plot before the molecular dynamic simulation is given in **(Figure 4.2)**. MD simulation of 2A and 3C protease is performed to stabilize these protein structures for the generation of more accurate binding complexes during docking.

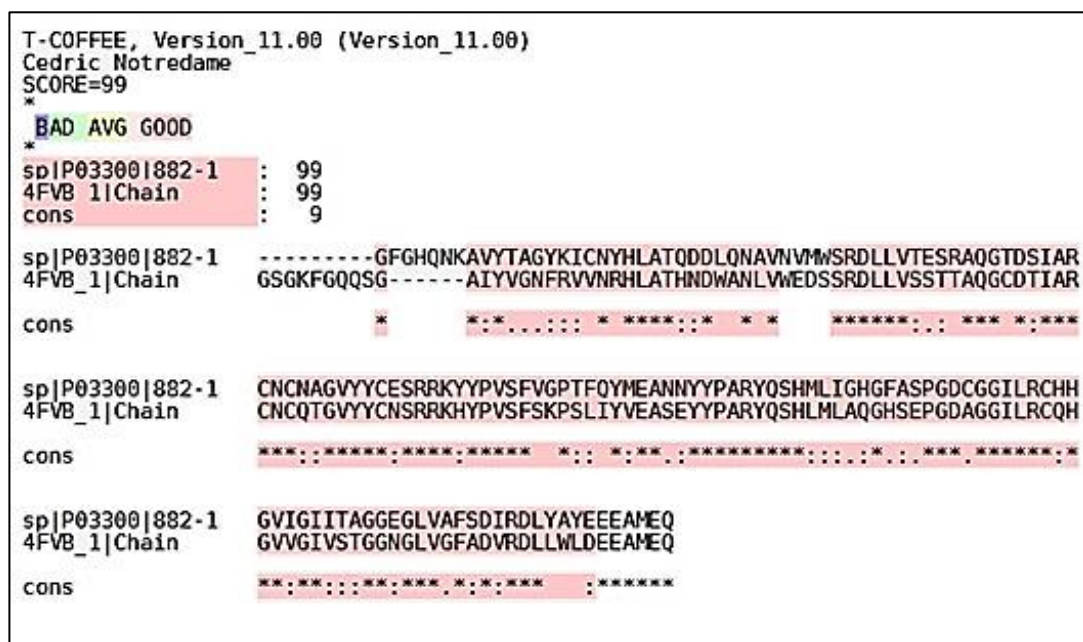


Figure 4.1 Sequence alignment of 2A protease of Poliovirus (sp |P03300) and enterovirus (4FVB) in T-coffee.

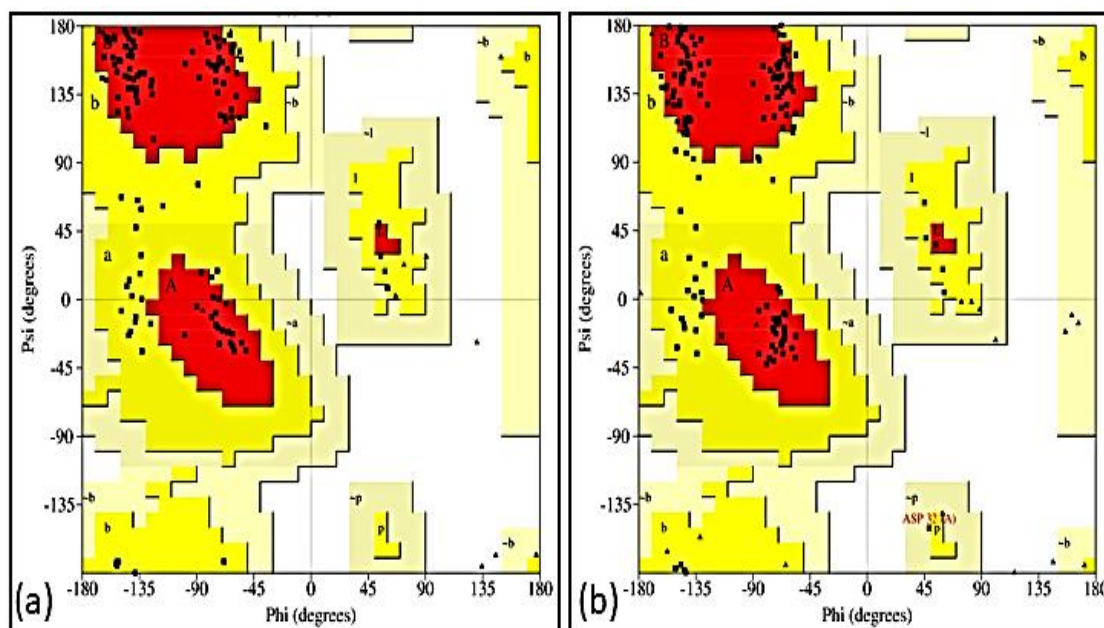


Figure 4.2 Ramachandran plot of the 2A protease (a) and 3C protease (b) at 0 ns. Both proteases have the majority of their residues in the favored region (red) and additionally allowed region (yellow) while very few in the generously allowed region (off white) and non in the disallowed region (white).

4.2 Molecular Dynamic Simulation of Proteins:

The molecular dynamic simulation of the homology model of 2A protease helps us in validating the stability of the protein in human body temperature and pressure. For this purpose, Molecular dynamic simulation was performed for multiple periods of 50ns, 100 ns, 150ns, and 500ns. It was found that 2A protease is most stable at 500ns as shown inf. It depicts that RMSD (Root Mean Square Deviation) throughout the simulation. This helps us in identifying the most stable RSMD and its time for a given protein. Moreover, the ERRAT score after MD simulation was increased to 93.54 and the Ramachandran plot (**Figure 4.3**) showed that 78.9% residues in the favored region and 20.3% residues in additionally allowed region as given in table (**Table 4.1**).

Although the structure 3C protease used in this study is extracted from PDB, it is possible that structure is not stable for molecular interaction. The ERRAT value of the 3C protease structure extracted from the PDB was 82.63 and the Ramachandran plot showed that 73.2% molecules in the favored region and 26.1% residues in the additionally allowed region as shown in **Table 4.1**. **Figure 4.3** Ramachandran plot of 2A protease (c) and 3C protease (d) at 500ns and 50 ns respectively. It can be seen

that only one residue of 2A protease is in the disallowed region. **Figure 4.2** represents the Ramachandran plot before MD simulation. To validate the stability of 3C protease MD simulation was performed for 50 ns. It was found that as this structure is predicted by the help of X-ray crystallography, it remains stable for 50ns as given in (**Figure 4.5**). The ERRAT score was increased to 87.17 and the Ramachandran plot displayed that 80.4% residues in the favored region and 18.3% residues in additionally allowed region details in (**Table 4.1**). The Ramachandran plot after the MD simulation of 3C protease is given in (**Figure 4.3 d**) which shows that no residue in the disallowed region (white).

Table 4.1 ERRAT score and Ramachandran plot residues at start and end of molecular dynamic simulation of 2A and 3C protease.

	Time (ns)	ERRAT	Ramachandran Plot			
			Favored Region(%)	Additionally allowed(%)	Generously allowed(%)	Disallowed
2A protease	0	76.42	69.5	30.5	0	0
	500	93.54	78.9	20.3	0	0.8
3C protease	0	82.63	73.2	26.1	0.7	0
	50	87.17	80.4	18.3	1.3	0

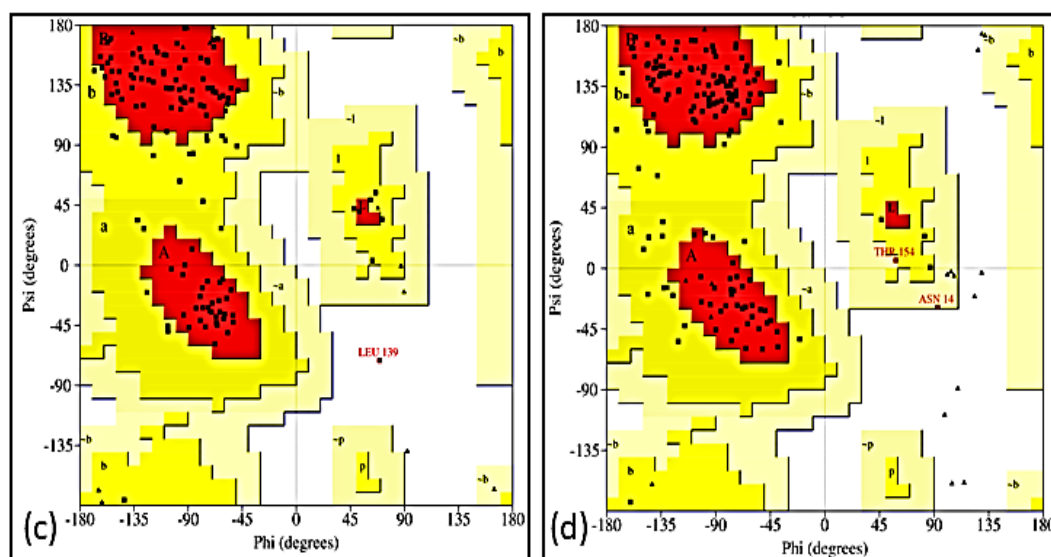


Figure 4.3 Ramachandran plot of 2A protease (c) and 3C protease (d) at 500ns and 50 ns respectively. It can be seen that only one residue of 2A protease is in the disallowed region.

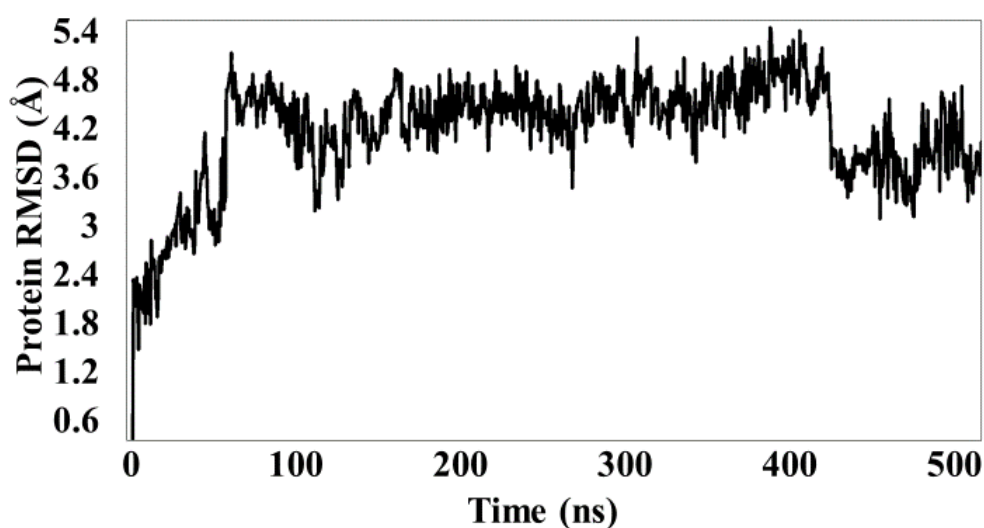


Figure 4.4 Molecular Dynamic Simulation of 2A Protease for 500 ns. It shows that at 0 ns the minimum RMSD is 0.6 (Å) and at 500 ns it has the RMSD from 3-3.6 (Å).

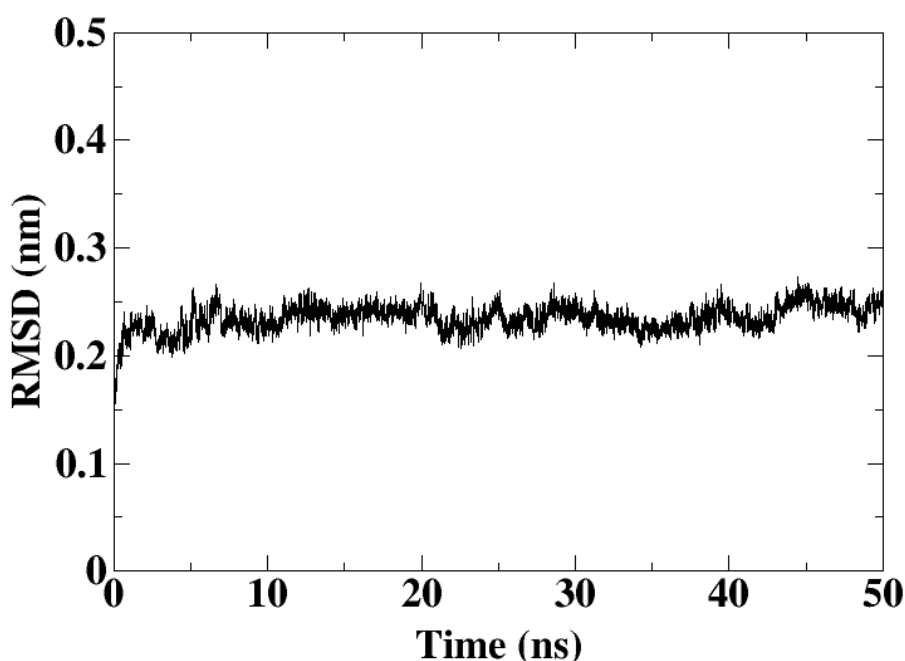


Figure 4.5 Molecular Dynamic simulation of 3C Protease for 50ns. It shows that the minimum RMSD is 0.15 nm and at 50 ns it has the maximum RMSD of 0.22 nm.

4.3 Molecular Docking:

The molecular docking of 2A and 3C protease helped us in the identification of the binding hypothesis of these proteins with their respective inhibitors. The Homology model of 2A protease after MD simulation was used in the molecular docking and inhibitors given in (**Table 3.1**). The docking protocol generated the 100 poses of each inhibitor compound within the defined coordinates. After the analysis

of the top 10 poses of each inhibitor final poses were selected on the basis of highest Gold score. To check the validity of the docking protocol, biological activity pIC_{50} was plotted against the Gold score in order to understand the relationship between the binding affinity represented by docking score and biological activity represented by pIC_{50} which should be ideally directly correlated. As given in (**Figure 4.6**) the Gold Score has a correlation of $R= 0.83$ and $R^2 = 0.69$. The correlation value represents that Gold Score has a high correlation with the pIC_{50} which means the biological activity value of each inhibitor is defined by the interaction pattern of each inhibitor with the 2A protease. Moreover, it was also observed that inhibitors have a high pIC_{50} as well as high Gold score as displayed in the (**Figure 4.6** displayed in red color) but difference in Gold score is not of greater than 5 which indicated that pIC_{50} behavior is also depended by an additional physiochemical property i.e., Molecular weight. These inhibitors biological activity value was further investigated with the help of Molecular weight. (**Figure 4.7**) shows the correlation of pIC_{50} with Molecular Weight (g/mol). It shows that the correlation between pIC_{50} and Molecular weight is $R=0.69$ and $R^2 = 0.47$, which also represents a strong positive correlation. This means that some inhibitor may have activity due to the proximity towards the target. To validate the stability of these complexes MD simulation of the complexes of these inhibitors with the 2A protease was performed. The complexes of inhibitors with high pIC_{50} like A10(ly3553349), A2(ly353352), A12(elastase inhibitor II), A9(ly355348) has low gold score as well whereas inhibitor with low pIC_{50} like A3(iodoacetamide) has low Gold score as given in **Table 3.1**. This validates our docking protocol as biological activity (pIC_{50}) have direct correlation with binding affinity i.e., Gold score.

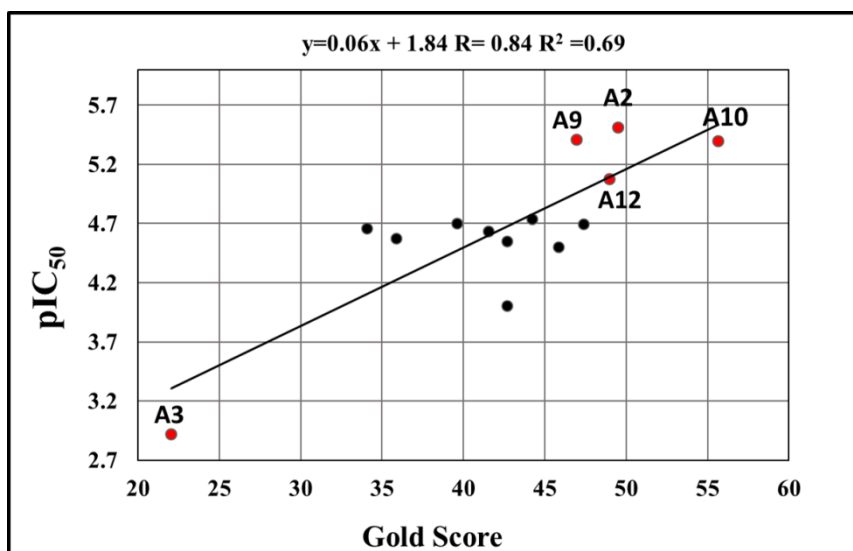


Figure 4.6 The docking results of 2A protease. Representation of Gold Score on X-axis and biological activity of pIC₅₀ on Y-axis. The red data points A10(ly3553349), A2(ly353352), A12(elastase inhibitor II), A9(ly355348) and A3(iodoacetamide).

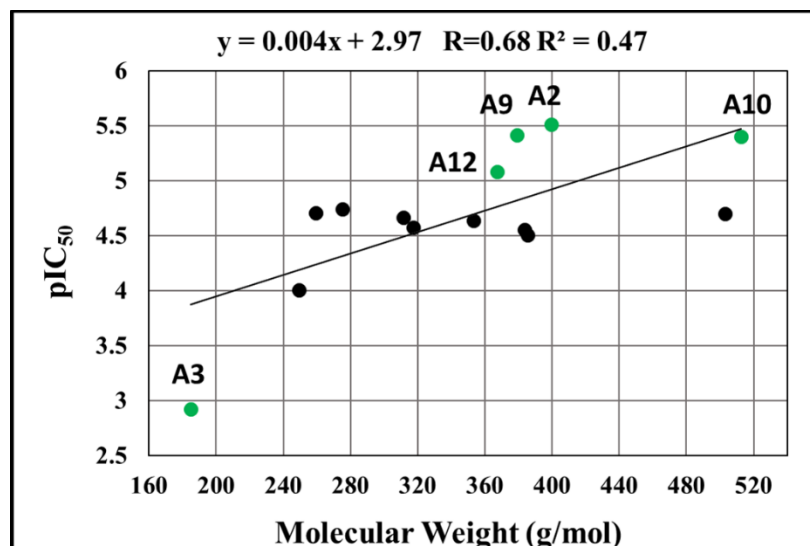


Figure 4.7 The correlation plot of biological activity of 2A protease and Molecular weight (g/mol). The green data points A10(ly3553349), A2(ly353352), A12(elastase inhibitor II), A9(ly355348) and A3(iodoacetamide).

The docking of 3C protease used the protein structure after MD simulation. It used the Gold score to generate 100 poses within the defined coordinated described in the previous section. The poses with the best score were selected as the final pose. It was found the Gold Score has a strong correlation with pIC₅₀ as given in (**Figure 4.8**). It shows that $R=0.7705$ and $R^2=0.5937$ which validate the docking protocol as docking score of ligands with high pIC₅₀ is also high. This shows that docking score is directly correlated with the pIC₅₀ and thus validating the pIC₅₀. This describes that biological activity i.e., pIC₅₀ is defined by the binding affinity of the inhibitors with

3C protease. The analysis of (Figure 4.8) shows two anomaly cases can be seen, the red datapoints C23(CHEMBL4212167) and C19(CHEMBL4212167) represent two inhibitors with a greater difference in the pIC_{50} but a small difference in the Gold Score. The red data points C6 (Rupintrivir) represent the data points that have less pIC_{50} but greater score and C1(CHEMBL4229177) less Gold Score with greater pIC_{50} which indicates these inhibitors pIC_{50} is not sufficiently defined by the docking interaction. With the aim of additional investigation, the biological activity of these inhibitors, $\log P$ (o/w) was also studied with pIC_{50} . As it can be seen in (Figure 4.9) the C23(CHEMBL4212167) and C19(CHEMBL4212167) have differences in the $\log P$ (o/w) value. The high pIC_{50} of C23(CHEMBL4212167) is due to a low value of $\log P$ (o/w). The biological activity pIC_{50} gave a weak negative correlation with $\log P$ (o/w) of $R = -0.496$ and $R^2 = 0.2232$. The $\log P$ (o/w) is insufficient to describe the behavior of pIC_{50} of C6(Rupintrivir) and C1(CHEMBL4229177) as these points are not following the trend, as there might be an additional physicochemical property contributing towards the behavior of pIC_{50} . To validate the binding pattern and explore the stability of these inhibitors' complexes with the 3C protease MD simulation was used.

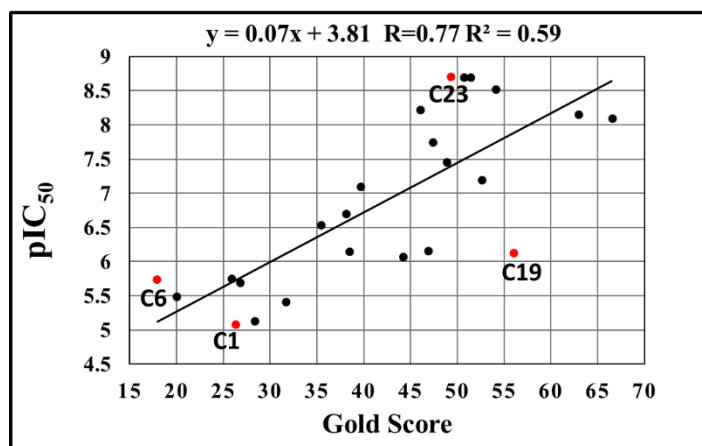


Figure 4.8 The 3C protease docking score Gold Score (x-axis) and pIC_{50} (y-axis). The red data points C23(CHEMBL4212167), C19(CHEMBL4212167), C6 (Rupintrivir), and C1(CHEMBL4229177).

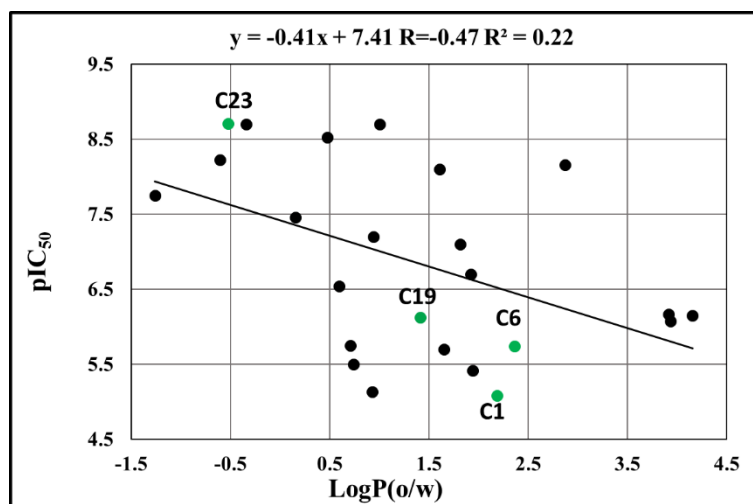


Figure 4.9 3C protease logP(o/w) correlation plot with pIC₅₀. The green data points C23(CHEMBL4212167),C19(CHEMBL4212167),C6(Rupintrivir),and C1 (CHEMBL4229177).

4.4 PLIF Analysis:

The PLIF analysis was performed on the final docking poses of 2A proteases. The PLIF analysis of 2A protease in (Figure 4.10) displayed that the highest overall abundance was 71.4% of residue Lys15 which constitutes in hydrogen bond formation. The residue Asn6 displayed an overall abundance of 64.3%, constituting hydrogen bonding and Van der Waal's interactions. The other significant residue was Gly3, Asp26, Try63, and Arg65 which consist of Hydrogen bonding, hydrophobic, and Van der Waal's interaction. This was interesting to observe that most of the interacting residues reported previously were not in high percentage in the PLIF. The binding position of the inhibitor data set was equally on two different binding sites involving the previously reported binding residues.

The PLIF analysis of 3C protease shows that the highest overall abundance was 66.7% displayed by the Ala144 consisting of the hydrophobic and Van der Waal's interactions with inhibitors as given in (Figure 4.11). Secondly, the Asn165 displayed the 65% interaction that were consisting of hydrogen bonding and Van der Waal's interaction. The other significant residues were Arg130, His 40, Cys 147 consisting of the hydrogen bonding and hydrophobic interaction. These interacting residues exist in the binding site as reported by previous docking studies on the 3C protease. It was also identified that all the inhibitors bind in the same binding position.

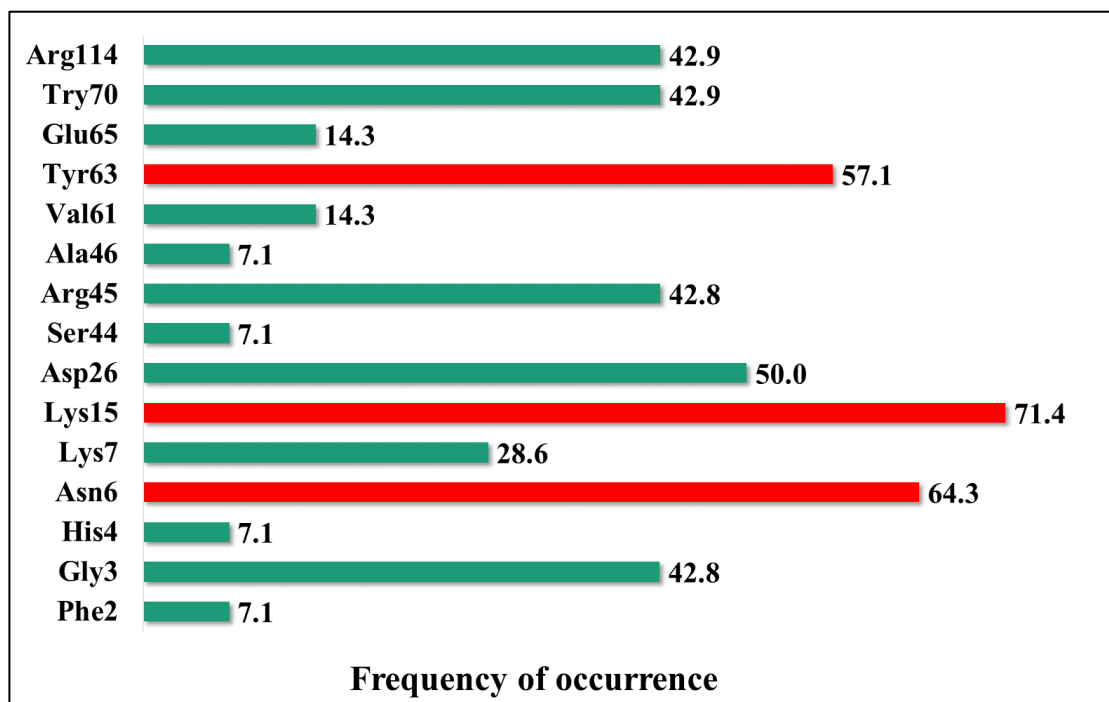


Figure 4.10 The PLIF analysis of 2A protease. The bars represent the overall percentage abundance of the residues in the interaction with the inhibitor data set.

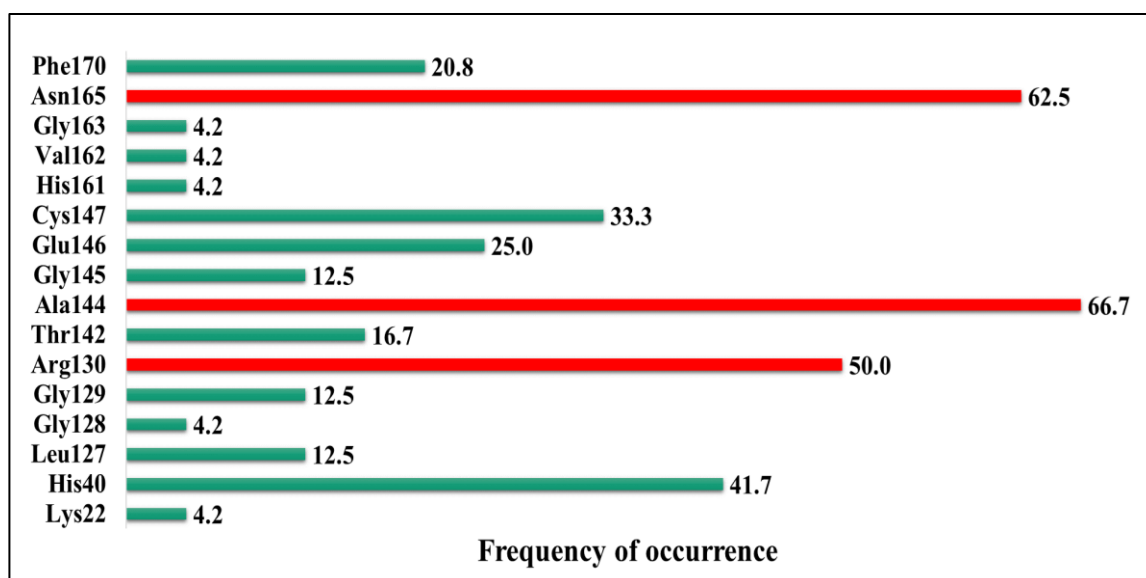


Figure 4.11 The PLIF analysis of 3C protease. The bars represent the overall percentage abundance of the residues in the interaction with the inhibitor data set.

4.5 Molecular Dynamic Simulation of Docking Complexes:

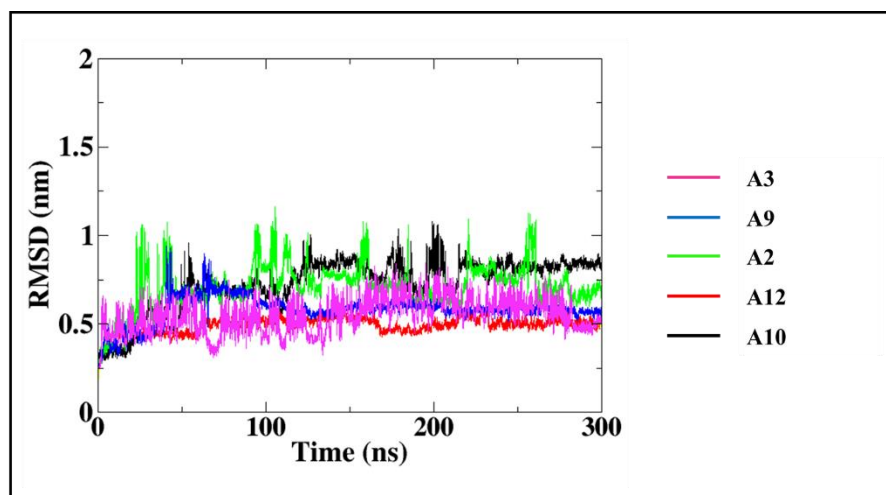


Figure 4.12 RMSD plot of 2A protease selected complexes for 300ns. The highest fluctuation.

The molecular dynamic simulation of the selected inhibitors complexes with 2A protease will help us in achieving stability at the human body temperature and pressure. The selected inhibitors i.e., A10(ly3553349), A2(ly353352), A12(elastase inhibitor II), A9(ly355348) and A3(iodoacetamide) RMSD plot as given in **Figure 4.12**. represents that the complexes with A10(ly3553349) and A12(elastase-II) were highly unstable till 100 ns with approximately 1nm RMSD (**Figure 4.13**). The oscillations were relatively less between 100 to 200 ns at remain below 1 nm (10 Å) and the complex stabilize itself at around 0.75 nm (7.5 Å) RMSD for 200 to 300 ns. It was identified that complex with A2(ly353352) and A9(ly355348) gained stability after 100ns at approximately 0.5 nm (5 Å) and remain stable for 200 ns. At last, the complex with A3(iodoacetamide) having the lowest pIC₅₀ was unstable for 250 ns and gain stability of below 0.5 nm (5 Å) for the last 50 ns (**Figure 4.12**). This can be due to the small molecular weight of A3(iodoacetamide). The binding side of 2A proteases was changed during the MD simulation of 100ns for all the selected complexes. This new identified binding site is more stable as compared to binding site identified by the docking which only represents the single snapshot during the docking experiments. **Table 4.2** represents the binding residue of 2A protease binding residue before and after MD simulation it can be identified that complex with A10(ly3553349) and A9(ly355348) has hydrogen bond formed by the Gln149 was found common as shown in **Figure 4.14**. This interacting residue was not found

in any other inhibitor after the MD simulation, the high of these two inhibitors can be due to this interaction.

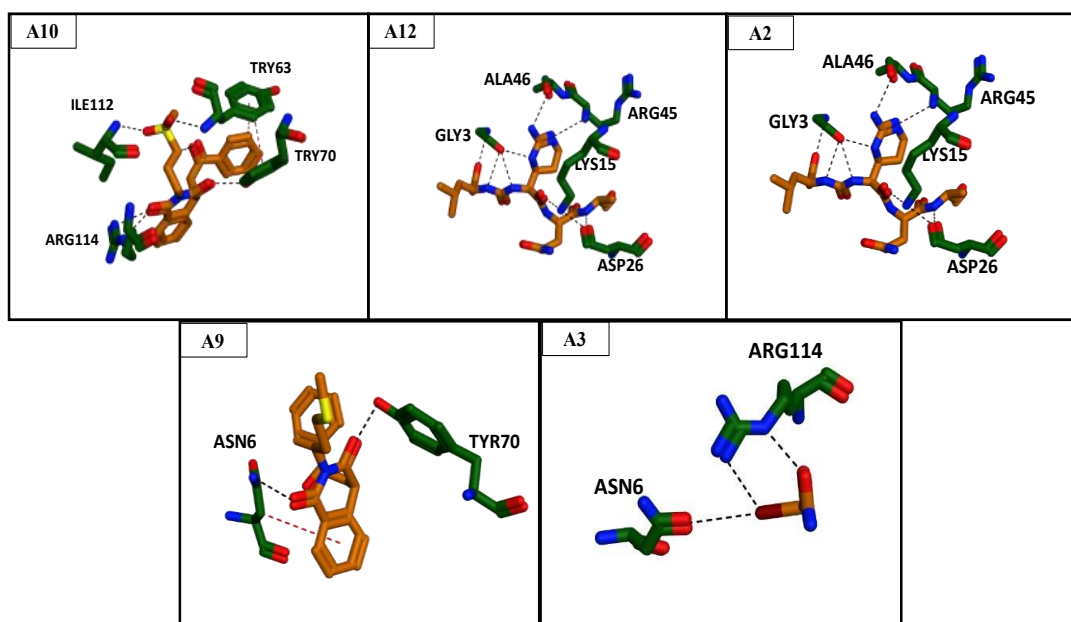


Figure 4.13 The interaction of binding residues of 2A proteases before the MD simulation A10(ly3553349), A2(ly353352), A12(elastase inhibitor II), A9(ly355348) and A3(iodoacetamide).

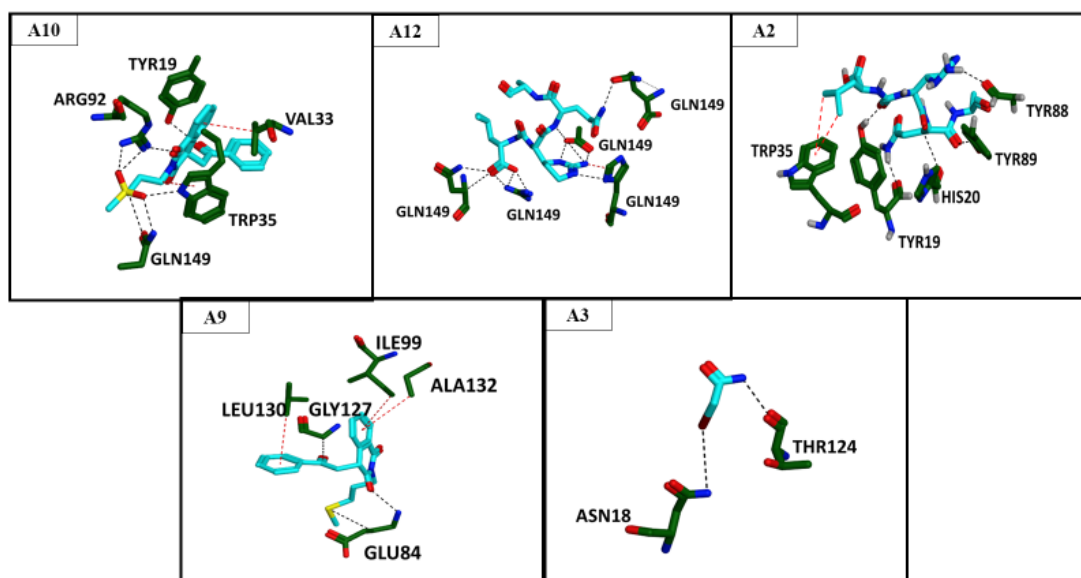


Figure 4.14 The interaction of binding residues of 2A proteases after the MD simulation A10(ly3553349), A2(ly353352), A12(elastase inhibitor II), A9(ly355348) and A3(iodoacetamide).

Table 4.2 The 2A selected complex binding residues and binding interaction at 0ns and 300ns.

Before MD				After MD		
Code	Binding residues Interaction	Amino acid atom	Ligand atom	Binding residues Interaction	Amino acid atom	Ligand atom
A10	Ile112 (H-bond) Arg114 (H-bond) Try70 (H-bond) Try63(H-bond)	N N N N	O O O O	Try19 (H-bond) Arg92 (H-bond) Trp35 (π – H bond) Gln149 (H-bond) Val33 (π – H bond)	O N N C N C	H O O Benzene ring O Benzene ring
A12	Lys15 (H-bond) Arg45 (H-bond) Asp26 (H-bond) His4 (H-bond) Gly3 (H-bond)	N N O N N	O N N O O	Gln149 (H-bond) Glu148 (H-bond) Ser36 (H-bond) Arg37 (H-bond)	N O O C N	O N N O O
A2	Ala46 (H-bond) Gly3 (H-bond) Lys15 (H-bond) Arg45 (H-bond) Asp26 (H-bond)	O H H H O	H O O N H	Try88 (H-bond) Try89 (H-bond) Trp35 (π – H bond) His20 (H-bond) Try19 (H-bond)	O H Benzene ring O O OH	H O H H H H
A9	Try70 (H-bond) Asn6 (π – H bond)	O H N	O Benzene ring O	Ile99 (π – H bond) Gly127 (H-bond) Glu84 (H-bond) Ala132(π – H bond) Leu130(π – H bond)	H H N H H H	Benzene ring O O S Benzene ring Benzene ring
A3	Arg114(H-bond) Asn6 (H-bond)	N N O	O I I	Thr124 (H-bond) Asn18 (H-bond)	O N	N O

In case of 3C protease it was found that the binding site remain same before after MD simulation. This indicate that interaction identified as a result of docking protocol were highly stable. Moreover, as the selected ligands were simulated for 50ns as given in () and it was observed that all the complexes have RMSD of 0.35 nm (3.5 Å) at 10ns and maintain their RMSD till 50ns. This early gain of stability for less time duration can be due to the reason that the 3C protease structure was extracted through x-ray crystallography. This also validates the docking protocol as (Table 4.3) shows that the binding cavity remains the same although there was a change in some binding residue. As Table 4.3 The binding residue of 3C proteases with selected inhibitors at 0 ns and 50 ns It shows that Complex with C23, C19 and C6 have residue Cys147 as common interacting residue which is also one of the significant residues a. It was also identified that complexes C23 and C19 have the common interacting residues Arg130, these residues were not found in the other two complexes, this residue was not found in the other complexes hence these inhibitors can have high activity due to this residue. Moreover, complex with ligands C19 and C1 have had residue Gly164 common. It was found that Cys147 was a stable interaction and remain stable throughout the simulation in complex with ligands C23 (CHEMBL4212167), C19 (CHEMBL4212167), and C6 (Rupintrivir). It was found another interaction of Arg130 was also stable and remains constant in complex 2b and 3b, whereas complex 4b had only one stable interaction that was Gly164 which was also observed in the complex C23 (CHEMBL4212167) and C6 (Rupintrivir) after 50 ns of MD simulation. Hence MD simulation of 3C selected complexes represents that more or less the interaction resulted at the end of docking simulation were stable and remain the same during 50 ns of MD simulation.

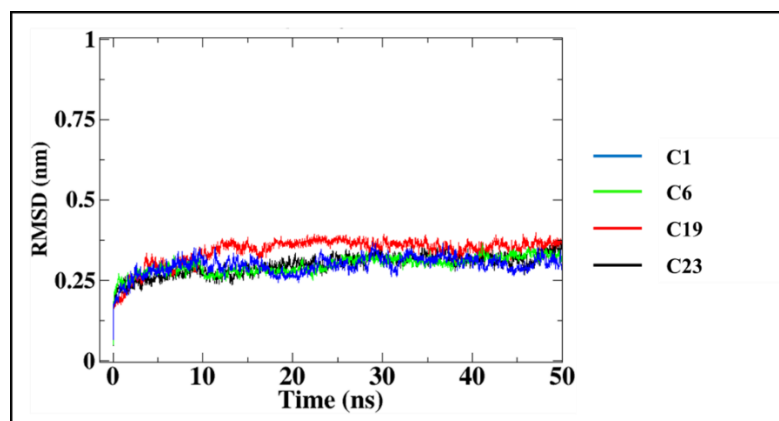


Figure 4.15 The 3C protease selected complexes RMSD plot for MD simulation of 50ns. The highest RMSD is approximately 0.3nm (3Å) for all the complexes.

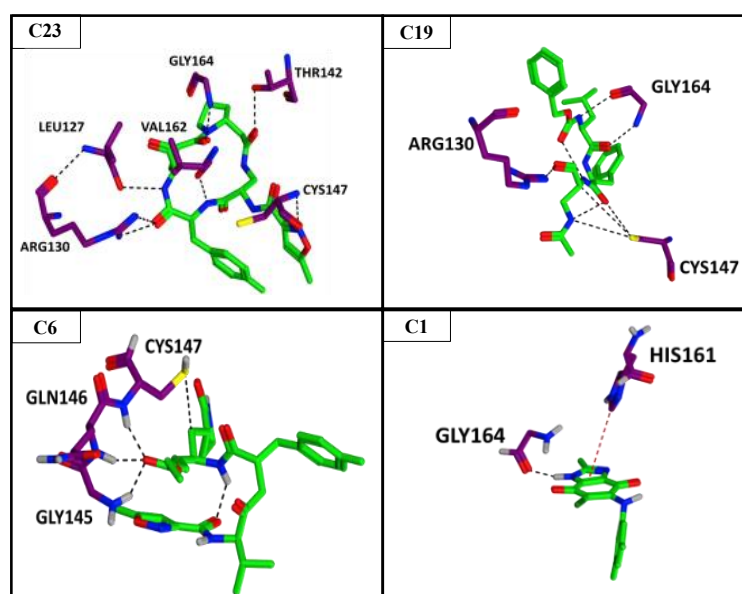


Figure 4.16 The 3C protease binding residue interaction before MD with C23(CHEMBL4212167), C19(CHEMBL4212167), C6 (Rupintrivir), and C1(CHEMBL4229177)

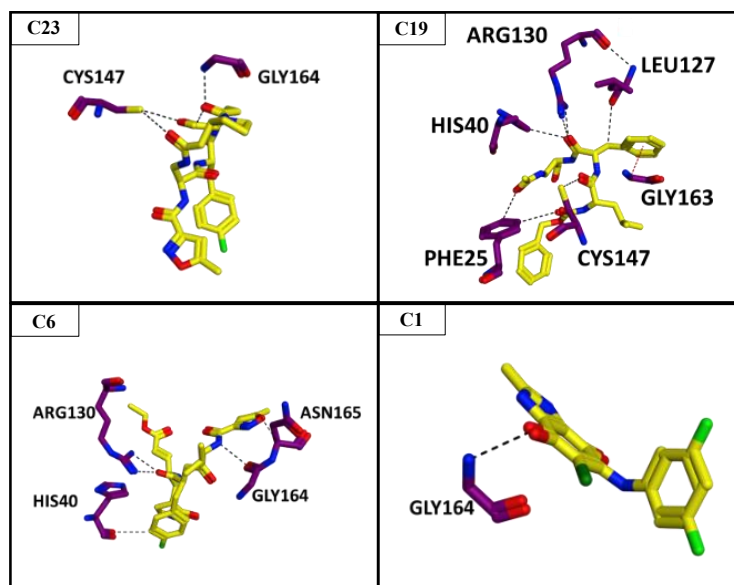


Figure 4.17 The 3C protease binding residue interaction after MD with C23(CHEMBL4212167), C19(CHEMBL4212167), C6 (Rupintrivir), and C1(CHEMBL4229177).

Table 4.3 The binding residue of 3C proteases with selected inhibitors at 0 ns and 50 ns

Code	Before MD			After MD		
	Binding residues Interaction	Amino acid atom	Ligand atom	Binding residues Interaction	Amino acid atom	Ligand atom
C23	Gly164 (H-bond)	N	O	Gly164 (H-bond)	NH	O
	Thr142 (H-bond)	O	O	Cys147 (H-bond)	H	O
	Val162 (H-bond)	O	N			
	Arg130 (H-bond)	N	O			
	Leu127 (H-bond)	O	N			
	Cys147 (H-bond)	N	O			
C19	Gly164 (H-bond)	O	H	Arg130 (H-bond)	SH	O
	Arg130 (H-bond)	NH	O	Leu127 (H-bond)	O	H
	Cys147 (H-bond)	NH2	O	His40(H-bond)	H	O
		SH	O	Phe25(H-bond)	H	O
				Cys147 (H-bond)	N	O
				Gly163 (π – H bond)	H	Benzene ring
C6	Cys147 (H-bond)	NH	O	Asn165 (H-bond)	N	H
	Gln146 (H-bond)	NH	O	Arg130 (H-bond)	H	O
	Gly145 (H-bond)	NH	O	Gly164 (H-bond)	O	H
				His40 (H-bond)	O	H
C1	His161 (H-bond)	Pentene	Benzene ring	Gly164 (H-bond)	N	O
	Gly164 (H-bond)	NH	O			

1. Overall, after the Structural methodology analysis, it can be concluded that the new identified binding site of 2A protease is highly significant for its inhibition as it remains stable from 200 to 300 ns.
2. The analysis of binding residues and interaction suggest that the residue Gln 149 forming the hydrogen bond is highly significant as it was found common in the highly active inhibitors A10(Iy3553349) and A12(elastase inhibitor II), after the MD simulation.
3. Moreover, binding site before and after the MD simulation is hydrogen doner residues so the inhibitor of the 2A protease should be hydrogen bond acceptor.
4. The 3C protease binding site identified by the docking protocol remain stable 50ns MD simulation.
5. The analysis of binding residues and interaction suggest that the residue Cys147 forming the hydrogen bond is highly significant as it was found common in the highly active inhibitors i.e., C23(CHEMBL4212167) and C19(CHEMBL4212167) before and after the MD simulation.

Machine Learning Model:

4.6 Decision Tree:

The decision tree was on the training dataset as mentioned in the previous chapter. This decision tree model was tested on the test set and model performance was checked with the help of parameters like accuracy, precision, specificity, and sensitivity by taking a human as positive and virus as a negative class. The data of sequences with the calculated frequency of each amino acid and length of sequence is given in appendix-1. **Table 4.4** represent the accuracy of the decision tree is high which is 100% on training and test data. It also has the precision, specificity, and sensitivity 1. The decision tree model is perfectly classifying the data in the respective class label. As given in (**Figure 4.18**) decision tree shows that number of leaves is 4 and the size of tree 7. Moreover, the length of a sequence is selected as root node and frequency of Proline and Alanine(the count of proline and alanine in total length of sequence)as internal nodes. This shows that from 21 attributes in the data only these three are highly important for the classification of human and viral sequences.

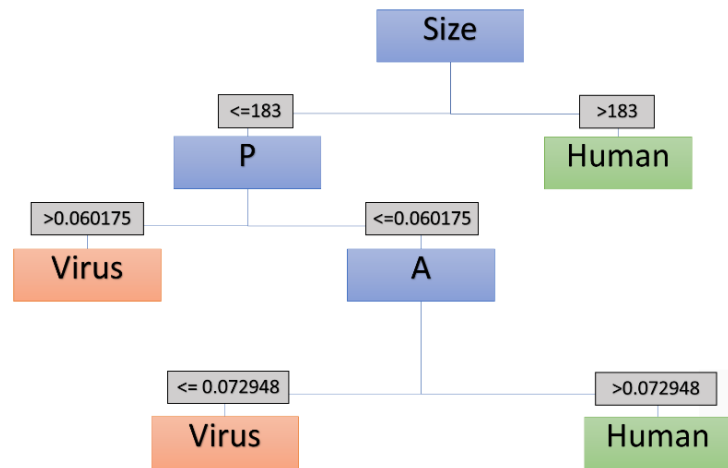


Figure 4.18 Decision Tree for classification of human and virus sequences.

Table 4.4 Decision tree model accuracy on training and test set

	Specificity	Sensitivity	Accuracy	Precession
Training set	1.00	1.00	100%	1.00
Test set	1.00	1.00	100%	1.00

4.7 Ensemble Methods:

Ensemble methods were used to further validate the results of the decision tree. **Table 4.5** shows different parameters to check the performance of ensemble methods. As it can be seen that all the models have more than 90% accuracy which means that data does not have variance. This also validates that model will give high accuracy on different training data set.

Table 4.5 Ensemble method accuracy parameters at batch size and test options.

Ensemble methods	Batch size	Test options	Sensitivity	Specificity	Accuracy	Precision
Bagging	50	10-fold	0.986	1.00	99.1667%	1.00
	50	80% split	0.974	1.00	97.9167%	1.00
	100	10-fold	0.986	1.00	99.1667%	1.00
	100	80% split	0.974	1.00	97.9167%	1.00
Boosting	50	10-fold	0.993	1.000	99.5833%	1.000
	50	80% split	0.974	1.000	97.9167%	1.00
	100	10-fold	0.993	1.000	99.5833%	1.000

	100	80% split	0.974	1.000	97.9167%	1.00
Random	50	10-fold	0.987	0.989	98.75%	0.993
Forest	50	80% split	1.00	1.00	100%	1.00
	100	10-fold	0.987	0.989	98.75%	0.993
	100	80% split	1.00	1.00	100%	1.00

CHAPTER 5
DISCUSSION

The poliovirus proteases play a very vital role in the proteolysis and taking over the host cell machinery. As both of these functions are significant for the replication and virulence of poliovirus with humans, this makes proteases a very significant and effective target for therapeutic drugs. The 2A and 3C protease are the two major proteins of poliovirus and are crucial for poliovirus replication. The aim of this study was the investigation of the inhibitors and protease binding interaction and identification of the important residues for the inhibition of proteases. The study of the interaction pattern is not possible without the 3D molecular structure of the protein. In this case, the PDB had the 3D crystallographic structure of 3C protease, but the 3D structure of 2A protease was unavailable. This led to the homology modeling of the 2A protease based on the 2A protease of EV7. After the development of the homology model development, the structure was validated by the high ERRAT score of 76.4 and no residues in the disallowed region. An approach to further increase these scores and stable the binding of complexes MD simulation was employed. The homology model was simulated for 500 ns with the RMSD between 3 and 4.2 Å. This step not only reduces the number of loops in the structure but also increases the ERRAT score to 97. Although the Ramachandran plot displayed one amino acid residue in the disallowed region this change in residue can be due to the change molecular structure due energy minimization during MD simulation. A similar approach was used for 3C protease although the structure was PDB derived. MD simulation before docking protein will lead to better binding poses as well as a more stable protein-ligand complex during the docking. The 3C protease was stable at 50 ns with RMSD of 2.5 Å as it is relatively more stable than a homology model of 2A protease.

After the structure is stabilized by the MD simulation, the docking of proteins was next the step. As poliovirus proteases are not an easy target, the major challenge was the literature data availability of 2A and 3C protease inhibitor. The exhaustive literature review led to the inhibitor activity of Elastase Inhibitor III, Calpaininhibitor1, Iodoacetamide, and methoxy succinyl – Ala – Ala – Pro – Val – chloromethyl ketone (MPCMK) reported previously against 2A protease of Poliovirus. Moreover, the only inhibitory activity of dipeptidyl inhibitors like dipeptidyl aldehyde, aldehyde bisulfite adduct salt, and rupintrivir was also reported

against 3C protease of poliovirus. This led to the similarity-based approach for the investigation of inhibitors against 2A and 3C protease. It found that databases like ChEMBL and PubChem have compounds with the inhibition of 2A and 3C protease of Human Rhinovirus. As HRV and poliovirus have a high sequence similarity score of 100 in T-coffee sequence alignment. Furthermore, there have been studies of inhibition poliovirus proteases by the protease inhibitor of similar viruses. This resulted in the total data of 15 inhibitors against 2A proteases and 24 inhibitors against 3C protease. Docking of both proteases was performed by selecting the binding cavity near the residues reported in the previous docking studies. For each inhibitor 100 poses were generated within the binding cavity. The correlation between the Gold Score of the best binding poses and affinity of the compound expressed as pIC_{50} for 2A protease was found to be $R^2 = 0.6914$ which demonstrates the strong direct correlation [50]. This correlation validates the docking protocol and explains the activity of most of the inhibitors due to the binding energy of compounds. Few exceptions were also observed, these inhibitors may have the activity due to the transport of compound towards the target rather than the interaction of inhibitor with the target. To get a deeper look, a correlation between the molecular weight and pIC_{50} was also studied which revealed that a positive correlation of 0.47 exists between these variables. This correlation value although not very high but looking at **Figure 4.7**, it shows that the selected inhibitor had a high difference in the molecular weight which cause the change in the pIC_{50} value [51]. Similarly, in 3C protease results, there was observed a strong direct correlation of 0.59. It was found that few data points have almost equal scores but high differences in the pIC_{50} so lipophilicity $\log P$ (o/w) correlation was also studied [51], [52]. It showed that the difference in the pIC_{50} is due to a change in $\log P$, which has an indirect correlation of $R^2 = 0.22$. This correlation plot **Figure 4.9** highlights the data points, it shows a decrease in $\log P$ (o/w) causes an increase in the pIC_{50} . Hence, the activity of these selected complexes is due to the change in their lipophilicity value as it helps the ligand to pass lipid-based cell membranes and reach the target protease.

MD simulation of selected complexes of 2A protease was performed and it was found that almost all the complexes achieved a stable RMSD between 5 and 7.5 Å for the simulation of 300 ns duration. The MD simulation is the process that helps

in the identification of the actively meaningful interaction for a specific period under the human body temperature and pressure [53]. The continuous force exerted on the molecules of complex only sustains the most stable interaction while breaking the less stable bonds. This resulted in the change of the binding site from the docking results. This change in binding site is more stable than docking results, as docking only provides a single snapshot of many interactions and docking score is also restricted by the binding cavity definition based on before docking studies. It should also be considered that ligand data used in this study against poliovirus also included ligands with reported pIC_{50} against HRV. There is a possibility that these inhibitors may bind more efficiently to the new binding position, hence optimizing their inhibitory function. Similarly, MD simulation of the 3C proteases selected complexes was also performed to check the validity of the docking results. The MD simulation of 3C proteases selected complexes achieved the stable RMSD of almost 3.5 Å in 50ns. The stabilization of 3C complexes in a short time is due to the reason that the 3C protease structure was crystallographic and also stabilized by the MD simulation before docking. The stability of the complex depends on the stability of the protein as 3c protease is more stable than 2A protease that's why it led to the more stable complexes. It was also observed that there was no change in the binding cavity rather binding residues within the same cavity change. This change in binding interaction is due to the energy minimization during MD production which makes new bonds and breaks old bonds to find the most stable conformation. Although 2A protease lack a common interaction between all the selected inhibitor, it was found that try19 and Gln149 was present in more than one highly active inhibitor interaction and hence are significant for inhibition. Whereas 3C protease Cys127 was found common in all highly active inhibitors complex before, and after which represents that it is highly significant. Moreover, the hydrogen bond interactions were common in all the inhibitors of 2a and 3C which represents that ligand with complimentary properties can be used for inhibition.

Another aim of this study was to classify the human protease from the viral protease. Due to the highly conserved nature of protease, one of the challenges of drug design is to identify the unique features that help in the differentiation of viral protease during the drug design process. Machine learning methodology is used in this study to classify viral proteases from human proteases. This classification will

help us in the recognition of unique classifying features of viral protease and human proteases. The machine learning model can classify the proteins based on structural data or sequential data. As structural data of viral proteases is not completely available in PDB. Moreover, homology model development from sequences is time taking processing for a large amount of data and needs to be verified by techniques like Molecular dynamic simulation. This led to the use of sequential data to develop the classification model with the help of machine learning algorithms. As the data was very simple with only 21 variables and decision tree is the right approach to describe the best classifying features here [54]. Results of the decision tree as discussed in the previous chapter shows that model accuracy and precision were 100 and 1 respectively. There is a possibility that the machine learning model is giving high accuracy because data have variance and biasness, so ensemble methods are the right approach to check the validity of the decision tree. The high accuracy of all the ensemble methods i.e., bagging, boosting, and random forest displays the evidence that data is not biased [55]. Thus, this decision tree shows that for the classification of human and viral sequence size of sequence and frequency of P and A are the most significant features.

CHAPTER 6
CONCLUSION

The poliovirus is a circulating virus that still infects many in some parts of the world. Poliovirus drug designing is an effort to combat this disease along with the vaccination. In this continuous effort, our research project is a small contribution towards the drug development against poliovirus proteases. One of the challenges drug designing is the target identification and validation. Poliovirus proteases 2A and 3C are conserved not only within poliovirus variants but also within the class of enteroviruses. Poliovirus lifecycle study helps us in understanding the vital role played by the proteases 2A and 3C in the multiplication and sustainability of poliovirus in the host cell. To block the poliovirus replication targeting proteases might be the first step in the right direction. Poliovirus proteases 2A and 3C have been studied in the past for drug designing purposes but the comparative analysis of these proteins was missing in the literature. This project aims to understand the binding interaction between 2A and 3C proteases with their currently available inhibitor data. This inhibitor data used against 2A and 3C proteases of poliovirus not only contains the poliovirus inhibitors but also the compounds used to inhibit the respective proteases in HRV. This approach will help us in the identification of such compounds that can be used to inhibit more than one viral protease. 2A proteases unavailability of 3D molecular structure is one of the challenges in studying the inhibition of poliovirus. This study makes use of techniques like homology modeling for molecular model development. 3C protease of poliovirus is already structure using X-ray crystallography and can be extracted from PDB. The method of MD simulation helped in the identification of the stable molecular structure for both proteases. Docking protocol aided in exploring the best binding pose of the inhibitors with the binding cavity of 2A and 3C protease. Docking is the first step in the interaction analysis, it was analyzed alongside pIC_{50} , molecular weight, and lipophilicity to get a comprehensive understanding of the binding affinity of the inhibitors with proteases. The 2A and 3C proteases docking analysis revealed that Gold scores have a direct correlation with the binding activity (pIC_{50}) whereas some compound's activity is better correlated with molecular weight and lipophilicity. These compounds have biological activity due to the transport towards the proteases. This helped us in evaluating the physiochemical feature responsible for the binding affinity. The MD simulation was performed for some selected complexes of 2A and

3C proteases with their respective inhibitors. The selected complexes were the inhibitors that were outliers from the overall data pattern. This analysis validated the docking protocol for 3C proteases interaction and also identified a more stable binding site for 2A protease. This interaction analysis comparison helped in differentiating the binding pattern between 2A and 3C protease. This represented that importance of Try19 and Gln146 in 2A protease and Cys147 significance in 3C protease for inhibition. The feature of ligands binding with the binding site residue is that most of these are hydrogen bond acceptor so the drug should also have hydrogen bond acceptor present at the binding site. The second aim of this study was to develop a classification model to differentiate between human and viral proteases using amino acid sequences. For this machine learning model decision tree was used. Moreover, ensemble methods were also employed to remove data biases and variation. The results represented the size of amino acid and frequency of proline and alanine as the most significant variables for the differentiation of human and viral sequences. This will serve as an approach towards the designing of drugs with specificity towards viral proteases as it is one of the major challenges in drug designing against poliovirus. In a nutshell, this study identified the significant protein residues of 2A and 3C proteases involved in the inhibition and also unique amino acid residues within the viruses. In the future, the development of better homology model pipelines, for the extraction of protein 3D features and training of the machine learning models on structural features will help in finding unique binding site for viral proteases. Moreover, using biological experiments to test more diverse given HRV inhibitors against poliovirus protease will help in the recognition of new compounds for the inhibitor of poliovirus replication.

References

- [1] B. Aylward and R. Tangermann, “The global polio eradication initiative: lessons learned and prospects for success,” *Vaccine*, vol. 29, pp. D80–D85, 2011.
- [2] E. R. Alexander, “Landmark perspective: Inactivated poliomyelitis vaccination. Issues reconsidered.,” *JAMA*, vol. 251, no. 20, pp. 2710–2712, May 1984.
- [3] A. B. Sabin and L. R. Boulger, “History of Sabin attenuated poliovirus oral live vaccine strains,” *J. Biol. Stand.*, vol. 1, no. 2, pp. 115–118, 1973.
- [4] H. J. Eggers, “Milestones in early poliomyelitis research (1840 to 1949),” *J. Virol.*, vol. 73, no. 6, pp. 4533–4535, Jun. 1999, doi: 10.1128/JVI.73.6.4533-4535.1999.
- [5] A. B. Sabin, “Oral poliovirus vaccine: history of its development and prospects for eradication of poliomyelitis,” *Jama*, vol. 194, no. 8, pp. 872–876, 1965.
- [6] J. F. Enders, F. C. Robbins, and T. H. Weller, “The cultivation of the poliomyelitis viruses in tissue culture,” *Rev. Infect. Dis.*, vol. 2, no. 3, pp. 493–504, 1980.
- [7] S. M. Lambert and H. Markel, “Making history: Thomas Francis, Jr, MD, and the 1954 Salk Poliomyelitis Vaccine Field Trial.,” *Arch. Pediatr. Adolesc. Med.*, vol. 154, no. 5, pp. 512–517, May 2000, doi: 10.1001/archpedi.154.5.512.
- [8] A. B. Sabin *et al.*, “Landmark article Aug 6, 1960: Live, orally given poliovirus vaccine. Effects of rapid mass immunization on population under conditions of massive enteric infection with other viruses. By Albert B. Sabin, Manuel Ramos-Alvarez, José Alvarez-Amezquita, Will,” *JAMA*, vol. 251, no. 22, pp. 2988–2993, Jun. 1984, doi: 10.1001/jama.251.22.2988.
- [9] S. Blume and I. Geesink, “A Brief History of Polio Vaccines,” *Science (80-)*, vol. 288, no. 5471, pp. 1593 LP – 1594, Jun. 2000, doi: 10.1126/science.288.5471.1593.
- [10] P. E. Dept, G. I. Program, and UNICEF, “The global polio eradication initiative Stop Transmission of Polio (STOP) program—1999–2013,” *MMWR. Morb. Mortal. Wkly. Rep.*, vol. 62, no. 24, p. 501, 2013.
- [11] B. Blondel, F. Colbère-Garapin, T. Couderc, A. Wirotius, and F. Guivel-Benhassine, “Poliovirus, Pathogenesis of Poliomyelitis, and Apoptosis,” *Curr. Top. Microbiol. Immunol.*, vol. 289, pp. 25–56, Feb. 2005, doi: 10.1007/3-540-27320-4_2.

-
- [12] D. M. Horstmann, R. W. McCollum, and A. D. Mascola, "Viremia in human poliomyelitis," *J. Exp. Med.*, vol. 99, no. 4, p. 355, 1954.
- [13] M. M. Mehndiratta, P. Mehndiratta, and R. Pande, "Poliomyelitis: historical facts, epidemiology, and current challenges in eradication," *The Neurohospitalist*, vol. 4, no. 4, pp. 223–229, Oct. 2014, doi: 10.1177/1941874414533352.
- [14] R. L. Bruno, "Paralytic vs. nonparalytic" polio: distinction without a difference?," *Am. J. Phys. Med. Rehabil.*, vol. 79, no. 1, pp. 4–12, 2000.
- [15] O. O. Adeyemi *et al.*, "Involvement of a Nonstructural Protein in Poliovirus Capsid Assembly," *J. Virol.*, vol. 93, no. 5, pp. e01447-18, Feb. 2019, doi: 10.1128/JVI.01447-18.
- [16] R. Virgen-Slane *et al.*, "An RNA virus hijacks an incognito function of a DNA repair enzyme," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 36, pp. 14634–14639, Sep. 2012, doi: 10.1073/pnas.1208096109.
- [17] M. Niepmann, "Internal translation initiation of picornaviruses and hepatitis C virus.," *Biochim. Biophys. Acta*, vol. 1789, no. 9–10, pp. 529–541, 2009, doi: 10.1016/j.bbagr.2009.05.002.
- [18] J. K. Daley, L. A. Gechman, J. Skipworth, and G. F. Rall, "Poliovirus replication and spread in primary neuron cultures," *Virology*, vol. 340, no. 1, pp. 10–20, 2005, doi: <https://doi.org/10.1016/j.virol.2005.05.032>.
- [19] A. Castelló, E. Alvarez, and L. Carrasco, "The multifaceted poliovirus 2A protease: regulation of gene expression by picornavirus proteases," *J. Biomed. Biotechnol.*, vol. 2011, p. 369648, 2011, doi: 10.1155/2011/369648.
- [20] A. Barco, E. Feduchi, and L. Carrasco, "Poliovirus Protease 3Cpro Kills Cells by Apoptosis," *Virology*, vol. 266, no. 2, pp. 352–360, 2000, doi: <https://doi.org/10.1006/viro.1999.0043>.
- [21] H. Guan, J. Tian, C. Zhang, B. Qin, and S. Cui, "Crystal structure of a soluble fragment of poliovirus 2CATPase," *PLoS Pathog.*, vol. 14, no. 9, p. e1007304, 2018.
- [22] S. K. Tsang, P. Danthi, M. Chow, and J. M. Hogle, "Stabilization of poliovirus by capsid-binding antiviral drugs is due to entropic effects," *J. Mol. Biol.*, vol. 296, no. 2, pp. 335–340, 2000.
- [23] B. A. Heinz and L. M. Vance, "The antiviral compound enviroxime targets the 3A coding region of rhinovirus and poliovirus," *J. Virol.*, vol. 69, no. 7, pp. 4189–4197, 1995.
- [24] S. Crotty, M.-C. Saleh, L. Gitlin, O. Beske, and R. Andino, "The poliovirus replication machinery can escape inhibition by an antiviral drug that targets a host cell protein," *J. Virol.*, vol. 78, no. 7, pp. 3378–3386, 2004.
-

-
- [25] M. A. McKinlay *et al.*, “Progress in the Development of Poliovirus Antiviral Agents and Their Essential Role in Reducing Risks That Threaten Eradication,” *J. Infect. Dis.*, vol. 210, no. suppl_1, pp. S447–S453, Nov. 2014, doi: 10.1093/infdis/jiu043.
- [26] G. Campagnola, P. Gong, and O. B. Peersen, “High-throughput screening identification of poliovirus RNA-dependent RNA polymerase inhibitors,” *Antiviral Res.*, vol. 91, no. 3, pp. 241–251, Sep. 2011, doi: 10.1016/j.antiviral.2011.06.006.
- [27] I. V Sandoval and L. Carrasco, “Poliovirus infection and expression of the poliovirus protein 2B provoke the disassembly of the Golgi complex, the organelle target for the antipoliovirus drug Ro-090179.,” *J. Virol.*, vol. 71, no. 6, pp. 4679–4693, Jun. 1997, doi: 10.1128/JVI.71.6.4679-4693.1997.
- [28] Z. Li, Z. Zou, Z. Jiang, X. Huang, and Q. Liu, “Biological Function and Application of Picornaviral 2B Protein: A New Target for Antiviral Drug Development,” *Viruses*, vol. 11, no. 6. 2019, doi: 10.3390/v11060510.
- [29] A. A. Agbowuro, W. M. Huston, A. B. Gamble, and J. D. A. Tyndall, “Proteases and protease inhibitors in infectious diseases,” *Med. Res. Rev.*, vol. 38, no. 4, pp. 1295–1331, 2018.
- [30] E. Scholar, “HIV Protease Inhibitors,” S. J. Enna and D. B. B. T. T. C. P. R. Bylund, Eds. New York: Elsevier, 2007, pp. 1–4.
- [31] C. Esneau, S. Croft, S.-L. Loo, and R. Ghildyal, “Chapter 2 - Rhinovirus diversity and virulence factors,” N. Bartlett, P. Wark, and D. B. T.-R. I. Knight, Eds. Academic Press, 2019, pp. 25–59.
- [32] B. J. Lamphear, R. Kirchweger, T. Skern, and R. E. Rhoads, “Mapping of Functional Domains in Eukaryotic Protein Synthesis Initiation Factor 4G (eIF4G) with Picornaviral Proteases: IMPLICATIONS FOR CAP-DEPENDENT AND CAP-INDEPENDENT TRANSLATIONAL INITIATION (*),” *J. Biol. Chem.*, vol. 270, no. 37, pp. 21975–21983, 1995, doi: <https://doi.org/10.1074/jbc.270.37.21975>.
- [33] S. de Breyne, J. M. Bonderoff, K. M. Chumakov, R. E. Lloyd, and C. U. T. Hellen, “Cleavage of eukaryotic initiation factor eIF5B by enterovirus 3C proteases,” *Virology*, vol. 378, no. 1, pp. 118–122, 2008, doi: <https://doi.org/10.1016/j.virol.2008.05.019>.
- [34] C. I. Rivera and R. E. Lloyd, “Modulation of enteroviral proteinase cleavage of poly(A)-binding protein (PABP) by conformation and PABP-associated factors.,” *Virology*, vol. 375, no. 1, pp. 59–72, May 2008, doi: 10.1016/j.virol.2008.02.002.
- [35] W. Tian, Z. Cui, Z. Zhang, H. Wei, and X. Zhang, “Poliovirus 2Apro induces the nucleic translocation of poliovirus 3CD and 3C' proteins,” *Acta Biochim.*
-

- Biophys. Sin. (Shanghai)*, vol. 43, no. 1, pp. 38–44, Jan. 2011, doi: 10.1093/abbs/gmq112.
- [36] Q. Feng *et al.*, “Enterovirus 2Apro targets MDA5 and MAVS in infected cells,” *J. Virol.*, vol. 88, no. 6, pp. 3369–3378, Mar. 2014, doi: 10.1128/JVI.02712-13.
- [37] A. Dotzauer and L. Kraemer, “Innate and adaptive immune responses against picornaviruses and their counteractions: An overview,” *World J. Virol.*, vol. 1, no. 3, pp. 91–107, Jun. 2012, doi: 10.5501/wjv.v1.i3.91.
- [38] A. Molla, C. U. Hellen, and E. Wimmer, “Inhibition of proteolytic activity of poliovirus and rhinovirus 2A proteinases by elastase-specific inhibitors,” *J. Virol.*, vol. 67, no. 8, pp. 4688–4695, Aug. 1993, doi: 10.1128/JVI.67.8.4688-4695.1993.
- [39] Y. Kim *et al.*, “Broad-Spectrum Antivirals against 3C or 3C-Like Proteases of Picornaviruses, Noroviruses, and Coronaviruses,” *J. Virol.*, vol. 86, no. 21, pp. 11754 LP – 11762, Nov. 2012, doi: 10.1128/JVI.01348-12.
- [40] Balasubramanian, S. Chawla, and B. Sathyamurthy, *IN SILICO STUDIES ON DENGUE AND POLIO VIRAL NON STRUCTURAL PROTEINS WITH SELECTED MENTHA ARVENSIS LEAVES CONSTITUENTS*. 2019.
- [41] S. K. Kashetty, M. Bandela, U. K. Suryavanshi, and V. Lokirevu, “In silico modelling and docking studies of camptothecin derivatives,” *Int. J. ChemTech Res.*, vol. 9, no. 9, pp. 274–284, 2016.
- [42] A. Younus *et al.*, “Structure-Function Mutational Analysis and Prediction of the Potential Impact of High Risk Non-Synonymous Single-Nucleotide Polymorphism on Poliovirus 2A Protease Stability Using Comprehensive Informatics Approaches,” *Genes (Basel)*, vol. 9, no. 5, Apr. 2018, doi: 10.3390/genes9050228.
- [43] C. L. Schoch *et al.*, “NCBI Taxonomy: a comprehensive update on curation, resources and tools,” *Database (Oxford)*, vol. 2020, p. baaa062, Jan. 2020, doi: 10.1093/database/baaa062.
- [44] C. Notredame, D. G. Higgins, and J. Heringa, “T-Coffee: A novel method for fast and accurate multiple sequence alignment,” *J. Mol. Biol.*, vol. 302, no. 1, pp. 205–217, Sep. 2000, doi: 10.1006/jmbi.2000.4042.
- [45] B. Webb and A. Sali, “Comparative Protein Structure Modeling Using MODELLER,” *Curr. Protoc. Bioinforma.*, vol. 54, pp. 5.6.1-5.6.37, Jun. 2016, doi: 10.1002/cpbi.3.
- [46] K. Zhu, T. Day, D. Warshaviak, C. Murrett, R. Friesner, and D. Pearlman, “Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction,” *Proteins Struct.*

- Funct. Bioinforma.*, vol. 82, no. 8, pp. 1646–1655, Aug. 2014, doi: <https://doi.org/10.1002/prot.24551>.
- [47] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, “Development and validation of a genetic algorithm for flexible docking 1 Edited by F. E. Cohen,” *J. Mol. Biol.*, vol. 267, no. 3, pp. 727–748, 1997, doi: <https://doi.org/10.1006/jmbi.1996.0897>.
- [48] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, “GROMACS: A message-passing parallel molecular dynamics implementation,” *Comput. Phys. Commun.*, vol. 91, no. 1, pp. 43–56, 1995, doi: [https://doi.org/10.1016/0010-4655\(95\)00042-E](https://doi.org/10.1016/0010-4655(95)00042-E).
- [49] E. Frank, M. A. Hall, and I. H. Witten, “The WEKA workbench,” *Data Min.*, pp. 553–571, 2017, doi: 10.1016/b978-0-12-804291-5.00024-6.
- [50] M. Atanasova, N. Yordanov, I. Dimitrov, S. Berkov, and I. Doytchinova, “Molecular docking study on galantamine derivatives as cholinesterase inhibitors,” *Mol. Inform.*, vol. 34, no. 6-7, pp. 394–403, 2015.
- [51] M. P. Edwards and D. A. Price, “Role of physicochemical properties and ligand lipophilicity efficiency in addressing drug safety risks,” *Annu. Rep. Med. Chem.*, vol. 45, pp. 380–391, 2010.
- [52] F. Sadeghi, A. Afkhami, T. Madrakian, and R. Ghavami, “A new approach for simultaneous calculation of pIC50 and logP through QSAR/QSPR modeling on anthracycline derivatives: a comparable study,” *J. Iran. Chem. Soc.*, vol. 18, no. 10, pp. 2785–2800, 2021, doi: 10.1007/s13738-021-02233-9.
- [53] R. Shukla and T. Tripathi, “Molecular Dynamics Simulation of Protein and Protein–Ligand Complexes,” in *Computer-Aided Drug Design*, Springer, 2020, pp. 133–161.
- [54] Z. He, Z. Wu, G. Xu, Y. Liu, and Q. Zou, “Decision Tree for Sequences,” *IEEE Trans. Knowl. Data Eng.*, 2021.
- [55] D. Che, Q. Liu, K. Rasheed, and X. Tao, “Decision tree and ensemble learning algorithms with their applications in bioinformatics,” *Softw. tools algorithms Biol. Syst.*, pp. 191–199, 2011.

Appendix-1

Uniprot_ID	Length of Sequence	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Q8N7X0	1667	0.06	0.04	0.03	0.04	0.01	0.09	0.04	0.04	0.02	0.05	0.09	0.10	0.02	0.04	0.06	0.09	0.06	0.02	0.03	0.06
Q8WYN0	398	0.05	0.03	0.04	0.07	0.03	0.08	0.05	0.06	0.02	0.05	0.11	0.07	0.03	0.05	0.06	0.07	0.05	0.03	0.03	0.05
Q9Y4P1	393	0.07	0.05	0.03	0.07	0.03	0.06	0.04	0.07	0.03	0.05	0.09	0.03	0.03	0.06	0.06	0.07	0.06	0.02	0.03	0.06
Q96DT6	458	0.05	0.04	0.03	0.07	0.02	0.08	0.03	0.06	0.03	0.06	0.08	0.08	0.02	0.06	0.03	0.08	0.07	0.02	0.04	0.05
Q86TL0	474	0.07	0.09	0.01	0.05	0.03	0.05	0.03	0.08	0.03	0.02	0.10	0.04	0.02	0.04	0.09	0.09	0.04	0.02	0.03	0.06
Q92560	729	0.07	0.06	0.04	0.05	0.01	0.08	0.04	0.07	0.03	0.04	0.09	0.06	0.01	0.03	0.08	0.10	0.04	0.01	0.02	0.07
Q13867	455	0.05	0.04	0.05	0.06	0.02	0.08	0.04	0.05	0.03	0.04	0.09	0.07	0.04	0.06	0.04	0.05	0.05	0.02	0.03	0.08
P07384	714	0.05	0.07	0.04	0.07	0.02	0.08	0.04	0.07	0.01	0.04	0.10	0.05	0.02	0.06	0.04	0.07	0.05	0.02	0.03	0.06
Q9HC96	672	0.09	0.08	0.01	0.04	0.04	0.06	0.06	0.09	0.03	0.03	0.11	0.02	0.01	0.04	0.07	0.07	0.05	0.03	0.02	0.07
Q9UMQ6	739	0.05	0.05	0.04	0.06	0.02	0.07	0.05	0.07	0.02	0.05	0.10	0.06	0.03	0.06	0.04	0.07	0.05	0.03	0.02	0.05
Q6MZZ7	669	0.05	0.06	0.04	0.06	0.02	0.06	0.07	0.06	0.03	0.04	0.10	0.05	0.03	0.06	0.04	0.08	0.05	0.03	0.02	0.06
A8MX76	684	0.05	0.06	0.04	0.04	0.02	0.07	0.07	0.06	0.02	0.04	0.13	0.06	0.02	0.06	0.05	0.07	0.04	0.03	0.03	0.05
Q6ZSI9	719	0.08	0.08	0.02	0.05	0.03	0.07	0.05	0.09	0.03	0.02	0.13	0.02	0.02	0.05	0.05	0.06	0.05	0.03	0.02	0.06
O75808	1086	0.10	0.07	0.02	0.04	0.04	0.07	0.04	0.08	0.03	0.02	0.09	0.03	0.01	0.03	0.09	0.08	0.05	0.02	0.01	0.07
P17655	700	0.07	0.06	0.04	0.07	0.02	0.10	0.04	0.07	0.01	0.07	0.10	0.06	0.02	0.06	0.03	0.06	0.05	0.02	0.03	0.04
P20807	821	0.05	0.05	0.05	0.07	0.02	0.07	0.05	0.06	0.03	0.06	0.07	0.07	0.03	0.06	0.05	0.07	0.05	0.02	0.03	0.05
O15484	640	0.06	0.07	0.04	0.06	0.03	0.07	0.04	0.07	0.04	0.05	0.08	0.06	0.01	0.04	0.05	0.06	0.05	0.03	0.03	0.07
Q9Y6Q1	641	0.03	0.06	0.05	0.06	0.02	0.07	0.05	0.06	0.02	0.05	0.10	0.07	0.02	0.05	0.04	0.05	0.07	0.02	0.04	0.07
Q9Y6W3	813	0.06	0.04	0.04	0.05	0.02	0.06	0.05	0.06	0.02	0.06	0.08	0.07	0.02	0.05	0.06	0.07	0.06	0.02	0.05	0.06
A6NHC0	703	0.07	0.05	0.04	0.06	0.03	0.08	0.05	0.08	0.02	0.04	0.11	0.05	0.01	0.05	0.04	0.07	0.05	0.02	0.03	0.05

O14815	690	0.07	0.06	0.04	0.07	0.02	0.09	0.04	0.06	0.02	0.06	0.10	0.06	0.01	0.07	0.04	0.06	0.05	0.02	0.02	0.04
P29466	404	0.06	0.05	0.04	0.06	0.03	0.08	0.04	0.06	0.02	0.07	0.08	0.07	0.04	0.04	0.04	0.08	0.06	0.01	0.02	0.06
Q92851	521	0.06	0.05	0.05	0.04	0.02	0.08	0.05	0.04	0.03	0.05	0.12	0.07	0.02	0.04	0.04	0.09	0.05	0.01	0.02	0.06
P31944	242	0.07	0.07	0.02	0.05	0.02	0.12	0.06	0.06	0.02	0.05	0.10	0.07	0.03	0.04	0.03	0.05	0.06	0.00	0.03	0.06
P42575	452	0.07	0.05	0.03	0.06	0.04	0.07	0.06	0.07	0.04	0.02	0.13	0.06	0.03	0.04	0.05	0.06	0.05	0.00	0.02	0.06
P42574	277	0.04	0.05	0.05	0.07	0.03	0.07	0.01	0.06	0.03	0.07	0.07	0.08	0.04	0.05	0.03	0.09	0.06	0.01	0.04	0.05
P49662	377	0.06	0.06	0.05	0.06	0.03	0.10	0.04	0.04	0.03	0.05	0.09	0.07	0.03	0.05	0.05	0.07	0.05	0.01	0.02	0.05
P51878	434	0.06	0.06	0.05	0.06	0.03	0.06	0.04	0.04	0.03	0.06	0.10	0.11	0.03	0.04	0.03	0.06	0.06	0.00	0.02	0.06
P55212	293	0.06	0.06	0.04	0.07	0.03	0.07	0.02	0.06	0.04	0.04	0.09	0.07	0.02	0.06	0.03	0.06	0.05	0.01	0.03	0.06
P55210	303	0.06	0.05	0.05	0.09	0.04	0.06	0.04	0.06	0.02	0.06	0.07	0.08	0.02	0.06	0.04	0.07	0.05	0.01	0.03	0.06
Q14790	479	0.03	0.05	0.04	0.07	0.03	0.09	0.05	0.05	0.01	0.07	0.11	0.07	0.03	0.04	0.04	0.08	0.04	0.00	0.04	0.03
P55211	416	0.06	0.07	0.03	0.06	0.03	0.06	0.05	0.07	0.02	0.05	0.12	0.04	0.02	0.05	0.06	0.08	0.04	0.01	0.01	0.06
O15519	480	0.04	0.06	0.03	0.06	0.02	0.07	0.05	0.05	0.03	0.05	0.14	0.08	0.03	0.03	0.03	0.08	0.03	0.01	0.04	0.06
P07858	339	0.05	0.04	0.05	0.05	0.05	0.06	0.03	0.10	0.03	0.04	0.06	0.04	0.02	0.03	0.06	0.08	0.04	0.04	0.05	0.06
P53634	463	0.07	0.04	0.05	0.05	0.03	0.05	0.03	0.09	0.03	0.05	0.08	0.05	0.02	0.04	0.04	0.07	0.06	0.02	0.06	0.06
Q9UBX1	484	0.09	0.06	0.05	0.05	0.02	0.05	0.04	0.08	0.01	0.03	0.11	0.05	0.03	0.04	0.06	0.08	0.04	0.02	0.03	0.06
P09668	335	0.09	0.02	0.06	0.03	0.04	0.05	0.04	0.09	0.03	0.04	0.07	0.07	0.03	0.05	0.05	0.07	0.05	0.03	0.05	0.05
P43235	329	0.06	0.04	0.08	0.04	0.02	0.06	0.03	0.09	0.02	0.04	0.09	0.09	0.02	0.02	0.04	0.07	0.03	0.03	0.05	0.07
P07711	333	0.08	0.04	0.06	0.05	0.02	0.09	0.04	0.10	0.02	0.03	0.05	0.06	0.05	0.05	0.04	0.07	0.04	0.03	0.05	0.06
P43234	321	0.07	0.05	0.05	0.05	0.02	0.04	0.04	0.08	0.02	0.04	0.08	0.04	0.02	0.05	0.05	0.11	0.03	0.03	0.05	0.07
P25774	331	0.06	0.04	0.06	0.05	0.03	0.06	0.03	0.08	0.04	0.03	0.08	0.08	0.03	0.02	0.03	0.08	0.04	0.02	0.06	0.08
O60911	334	0.08	0.04	0.07	0.04	0.03	0.06	0.04	0.10	0.02	0.02	0.06	0.08	0.04	0.05	0.04	0.07	0.03	0.03	0.04	0.07
P56202	376	0.08	0.05	0.03	0.04	0.03	0.06	0.06	0.08	0.03	0.05	0.09	0.05	0.02	0.05	0.07	0.06	0.05	0.03	0.03	0.06
Q9UBR2	303	0.08	0.07	0.06	0.05	0.04	0.05	0.03	0.11	0.03	0.06	0.07	0.03	0.02	0.01	0.05	0.06	0.05	0.04	0.06	0.04
Q9NQC7	956	0.04	0.04	0.04	0.05	0.03	0.08	0.04	0.07	0.02	0.05	0.10	0.07	0.02	0.05	0.06	0.08	0.04	0.01	0.03	0.06
Q9NUU6	356	0.05	0.08	0.03	0.04	0.02	0.07	0.04	0.04	0.03	0.04	0.11	0.07	0.03	0.06	0.03	0.08	0.04	0.03	0.05	0.06

Q8NB37	220	0.14	0.05	0.03	0.05	0.04	0.05	0.03	0.06	0.03	0.02	0.10	0.03	0.01	0.05	0.05	0.11	0.04	0.01	0.01	0.09
Q92820	318	0.07	0.03	0.05	0.04	0.02	0.06	0.03	0.07	0.02	0.05	0.11	0.07	0.02	0.07	0.04	0.07	0.05	0.01	0.05	0.06
Q99538	433	0.06	0.04	0.05	0.06	0.02	0.06	0.03	0.06	0.06	0.05	0.08	0.06	0.04	0.02	0.06	0.06	0.06	0.01	0.06	0.08
Q9UDY8	824	0.06	0.04	0.04	0.06	0.03	0.07	0.04	0.06	0.03	0.04	0.12	0.06	0.02	0.04	0.06	0.07	0.05	0.01	0.03	0.06
Q96FW1	271	0.06	0.04	0.03	0.07	0.01	0.10	0.08	0.05	0.03	0.06	0.09	0.06	0.02	0.05	0.03	0.07	0.04	0.00	0.06	0.06
Q96DC9	234	0.07	0.06	0.04	0.06	0.02	0.09	0.03	0.03	0.04	0.06	0.10	0.06	0.02	0.08	0.02	0.08	0.05	0.00	0.04	0.05
Q96G74	571	0.11	0.06	0.03	0.05	0.02	0.07	0.05	0.11	0.02	0.03	0.05	0.05	0.03	0.02	0.11	0.07	0.03	0.01	0.03	0.06
Q8TE49	926	0.12	0.08	0.03	0.05	0.02	0.07	0.04	0.08	0.03	0.02	0.09	0.05	0.02	0.02	0.07	0.08	0.05	0.01	0.02	0.05
Q6GQQ9	843	0.06	0.06	0.03	0.05	0.01	0.09	0.04	0.10	0.03	0.02	0.09	0.06	0.02	0.03	0.08	0.09	0.04	0.02	0.02	0.05
Q96BN8	352	0.09	0.06	0.03	0.05	0.02	0.11	0.04	0.05	0.02	0.04	0.10	0.07	0.04	0.03	0.06	0.04	0.05	0.02	0.04	0.05
Q504Q3	1202	0.06	0.05	0.04	0.06	0.02	0.07	0.04	0.06	0.03	0.05	0.11	0.05	0.02	0.05	0.05	0.07	0.05	0.01	0.03	0.06
Q99497	189	0.13	0.04	0.03	0.05	0.02	0.08	0.02	0.10	0.02	0.05	0.10	0.08	0.03	0.02	0.05	0.05	0.04	0.00	0.02	0.10
Q9NXJ5	209	0.05	0.04	0.02	0.06	0.04	0.07	0.05	0.09	0.04	0.05	0.09	0.05	0.02	0.02	0.05	0.06	0.05	0.01	0.04	0.10
A6NFU8	196	0.07	0.07	0.02	0.05	0.03	0.07	0.04	0.09	0.03	0.04	0.11	0.06	0.03	0.02	0.06	0.05	0.04	0.01	0.03	0.11
Q92643	395	0.06	0.05	0.05	0.06	0.01	0.06	0.04	0.05	0.03	0.06	0.09	0.05	0.04	0.06	0.05	0.08	0.06	0.01	0.04	0.07
Q12765	414	0.08	0.05	0.02	0.07	0.03	0.11	0.04	0.06	0.02	0.05	0.07	0.05	0.02	0.05	0.06	0.07	0.05	0.02	0.03	0.07
Q96FV2	425	0.11	0.06	0.01	0.04	0.02	0.07	0.08	0.08	0.03	0.04	0.09	0.02	0.02	0.04	0.06	0.07	0.05	0.02	0.01	0.06
Q0VDG4	424	0.06	0.05	0.05	0.06	0.02	0.09	0.03	0.05	0.03	0.05	0.09	0.07	0.03	0.04	0.04	0.06	0.05	0.02	0.03	0.07
Q9P0U3	644	0.04	0.06	0.05	0.05	0.02	0.06	0.06	0.04	0.04	0.04	0.09	0.07	0.02	0.04	0.04	0.12	0.06	0.01	0.02	0.05
Q9HC62	589	0.02	0.08	0.05	0.04	0.02	0.07	0.04	0.07	0.03	0.04	0.10	0.08	0.02	0.04	0.05	0.09	0.06	0.02	0.03	0.05
Q9H4L4	574	0.05	0.09	0.02	0.05	0.02	0.07	0.05	0.08	0.03	0.03	0.09	0.05	0.02	0.04	0.10	0.07	0.05	0.02	0.02	0.05
Q96HI0	755	0.03	0.06	0.05	0.05	0.03	0.06	0.06	0.06	0.03	0.04	0.09	0.09	0.02	0.05	0.05	0.09	0.04	0.01	0.02	0.05
Q9GZR1	1112	0.04	0.05	0.07	0.06	0.02	0.09	0.04	0.05	0.02	0.06	0.08	0.09	0.01	0.04	0.05	0.10	0.05	0.01	0.02	0.05
Q9BQF6	1050	0.02	0.05	0.05	0.05	0.02	0.09	0.05	0.03	0.02	0.05	0.10	0.08	0.01	0.03	0.05	0.13	0.06	0.01	0.02	0.05
Q96LD8	212	0.08	0.03	0.06	0.07	0.02	0.05	0.06	0.03	0.03	0.04	0.11	0.06	0.02	0.07	0.04	0.09	0.04	0.01	0.03	0.06
Q9UJW2	476	0.07	0.06	0.05	0.04	0.05	0.07	0.05	0.07	0.02	0.04	0.06	0.06	0.02	0.04	0.04	0.08	0.06	0.04	0.05	0.04

Q9GZM7	467	0.06	0.08	0.04	0.05	0.05	0.05	0.04	0.11	0.04	0.04	0.08	0.02	0.03	0.03	0.07	0.05	0.04	0.03	0.03	0.05
P21580	790	0.06	0.07	0.05	0.04	0.05	0.07	0.05	0.06	0.04	0.03	0.08	0.06	0.03	0.04	0.07	0.07	0.06	0.01	0.02	0.04
P09936	223	0.09	0.04	0.05	0.05	0.03	0.10	0.05	0.07	0.03	0.03	0.10	0.07	0.03	0.06	0.04	0.05	0.03	0.00	0.01	0.08
P15374	230	0.07	0.04	0.04	0.06	0.01	0.12	0.03	0.05	0.04	0.05	0.10	0.06	0.03	0.05	0.05	0.06	0.05	0.01	0.03	0.07
Q9Y5K5	329	0.07	0.04	0.05	0.05	0.02	0.11	0.07	0.05	0.02	0.06	0.09	0.08	0.03	0.05	0.03	0.06	0.03	0.01	0.02	0.06
Q6NVU6	142	0.08	0.05	0.01	0.05	0.04	0.04	0.05	0.15	0.04	0.01	0.11	0.02	0.01	0.04	0.06	0.08	0.03	0.03	0.03	0.08
Q9NUQ7	469	0.05	0.04	0.04	0.06	0.01	0.05	0.04	0.07	0.04	0.09	0.10	0.06	0.03	0.04	0.06	0.06	0.05	0.02	0.04	0.06
O94782	785	0.04	0.03	0.06	0.04	0.02	0.11	0.04	0.06	0.01	0.05	0.10	0.10	0.01	0.03	0.05	0.10	0.05	0.01	0.03	0.05
Q14694	798	0.06	0.04	0.04	0.05	0.01	0.07	0.05	0.08	0.02	0.05	0.08	0.06	0.01	0.03	0.08	0.09	0.07	0.01	0.04	0.07
P51784	963	0.06	0.06	0.04	0.06	0.02	0.08	0.04	0.06	0.03	0.04	0.09	0.05	0.02	0.04	0.06	0.07	0.04	0.02	0.04	0.07
O75317	370	0.05	0.05	0.06	0.06	0.03	0.08	0.04	0.04	0.03	0.05	0.10	0.07	0.02	0.05	0.04	0.06	0.06	0.01	0.05	0.05
Q92995	863	0.07	0.06	0.05	0.06	0.02	0.08	0.04	0.07	0.03	0.05	0.08	0.05	0.03	0.05	0.06	0.07	0.03	0.01	0.03	0.06
P54578	494	0.06	0.03	0.03	0.06	0.02	0.09	0.06	0.05	0.01	0.04	0.09	0.09	0.04	0.04	0.04	0.08	0.05	0.01	0.03	0.06
Q9Y4E8	981	0.04	0.04	0.07	0.08	0.02	0.07	0.04	0.06	0.02	0.05	0.09	0.07	0.02	0.04	0.05	0.07	0.06	0.01	0.04	0.04
Q9Y5T5	823	0.04	0.04	0.07	0.05	0.03	0.09	0.06	0.05	0.03	0.04	0.09	0.10	0.02	0.02	0.04	0.07	0.05	0.01	0.02	0.07
Q7RTZ2	530	0.08	0.05	0.03	0.05	0.04	0.06	0.07	0.05	0.05	0.03	0.11	0.06	0.02	0.02	0.05	0.08	0.05	0.01	0.03	0.05
C9JH3	530	0.08	0.06	0.03	0.05	0.03	0.06	0.07	0.05	0.05	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.06
C9JVI0	530	0.08	0.05	0.03	0.05	0.03	0.06	0.07	0.05	0.05	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
C9JPN9	530	0.08	0.05	0.03	0.04	0.03	0.06	0.07	0.05	0.05	0.03	0.11	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
C9JLJ4	530	0.08	0.05	0.03	0.04	0.03	0.06	0.07	0.05	0.05	0.03	0.11	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
C9J2P7	553	0.08	0.06	0.03	0.05	0.03	0.06	0.07	0.05	0.06	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.07	0.01	0.03	0.05
D6RBQ6	530	0.08	0.05	0.03	0.05	0.03	0.06	0.07	0.05	0.05	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
D6R9N7	530	0.08	0.05	0.03	0.05	0.03	0.06	0.07	0.05	0.05	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
D6RCP7	530	0.08	0.05	0.03	0.04	0.03	0.06	0.07	0.05	0.05	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
Q6R6M4	530	0.08	0.05	0.03	0.05	0.04	0.06	0.07	0.05	0.05	0.02	0.11	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
D6RJB6	530	0.08	0.05	0.03	0.05	0.03	0.06	0.07	0.05	0.05	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05

D6R901	530	0.08	0.05	0.04	0.04	0.03	0.06	0.07	0.05	0.05	0.03	0.11	0.06	0.02	0.02	0.05	0.08	0.05	0.01	0.03	0.05
D6RA61	530	0.08	0.05	0.03	0.05	0.03	0.06	0.07	0.05	0.05	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
D6RBM5	183	0.10	0.06	0.03	0.04	0.04	0.05	0.06	0.05	0.05	0.02	0.11	0.04	0.04	0.03	0.07	0.07	0.05	0.01	0.02	0.04
Q0WX57	530	0.08	0.05	0.03	0.05	0.03	0.06	0.07	0.05	0.05	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
A6NCW0	530	0.08	0.05	0.03	0.05	0.04	0.06	0.07	0.05	0.05	0.02	0.11	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
A6NCW7	530	0.07	0.05	0.04	0.05	0.04	0.06	0.07	0.05	0.05	0.02	0.11	0.06	0.03	0.02	0.05	0.08	0.05	0.01	0.03	0.06
A8MUK1	530	0.08	0.05	0.03	0.05	0.03	0.06	0.07	0.05	0.05	0.03	0.10	0.06	0.02	0.02	0.05	0.08	0.06	0.01	0.03	0.05
P0C7H9	530	0.07	0.04	0.03	0.05	0.03	0.05	0.07	0.05	0.05	0.02	0.11	0.07	0.02	0.03	0.05	0.09	0.06	0.01	0.03	0.06
P0C7I0	530	0.07	0.05	0.04	0.05	0.04	0.06	0.07	0.05	0.05	0.02	0.11	0.06	0.03	0.02	0.06	0.08	0.05	0.01	0.03	0.05
Q9UMW8	372	0.05	0.06	0.04	0.05	0.05	0.06	0.07	0.03	0.03	0.05	0.12	0.06	0.03	0.04	0.04	0.08	0.04	0.01	0.03	0.06
O94966	1318	0.08	0.06	0.03	0.05	0.03	0.08	0.05	0.07	0.02	0.02	0.10	0.05	0.01	0.04	0.08	0.08	0.05	0.01	0.02	0.08
O75604	605	0.05	0.09	0.04	0.05	0.02	0.05	0.04	0.07	0.02	0.03	0.11	0.04	0.02	0.04	0.06	0.11	0.07	0.01	0.05	0.04
Q9Y2K6	914	0.07	0.06	0.03	0.06	0.04	0.08	0.05	0.07	0.03	0.04	0.09	0.05	0.01	0.04	0.06	0.09	0.04	0.01	0.03	0.06
Q9UK80	565	0.06	0.11	0.03	0.04	0.03	0.05	0.03	0.08	0.03	0.02	0.12	0.04	0.01	0.04	0.10	0.09	0.04	0.01	0.02	0.05
Q9UPT9	525	0.05	0.05	0.04	0.05	0.05	0.06	0.04	0.06	0.06	0.05	0.09	0.07	0.02	0.05	0.04	0.08	0.06	0.02	0.03	0.04
Q9UPU5	2620	0.06	0.05	0.04	0.05	0.02	0.07	0.05	0.06	0.02	0.05	0.12	0.06	0.02	0.04	0.04	0.10	0.05	0.01	0.03	0.06
Q9UHP3	1055	0.06	0.06	0.04	0.05	0.01	0.11	0.06	0.04	0.03	0.05	0.11	0.06	0.02	0.04	0.05	0.07	0.05	0.01	0.03	0.05
Q9BXU7	913	0.05	0.04	0.06	0.05	0.02	0.08	0.05	0.05	0.03	0.05	0.11	0.11	0.02	0.05	0.04	0.08	0.05	0.01	0.03	0.04
A6NNY8	438	0.05	0.05	0.04	0.05	0.04	0.06	0.05	0.06	0.05	0.05	0.09	0.06	0.02	0.04	0.05	0.09	0.07	0.01	0.03	0.05
Q96RU2	1077	0.07	0.05	0.04	0.05	0.02	0.10	0.06	0.04	0.02	0.04	0.10	0.06	0.03	0.03	0.05	0.09	0.04	0.01	0.04	0.06
Q9HBJ7	922	0.04	0.03	0.07	0.05	0.02	0.08	0.07	0.05	0.02	0.05	0.11	0.07	0.02	0.04	0.05	0.09	0.04	0.01	0.02	0.05
Q9Y6I4	520	0.05	0.05	0.06	0.04	0.05	0.06	0.04	0.05	0.04	0.03	0.10	0.08	0.01	0.04	0.03	0.09	0.06	0.01	0.04	0.07
Q70CQ3	517	0.06	0.06	0.03	0.04	0.03	0.05	0.05	0.05	0.06	0.03	0.11	0.05	0.03	0.04	0.07	0.10	0.06	0.02	0.02	0.06
Q70CQ4	1352	0.07	0.07	0.02	0.04	0.02	0.05	0.05	0.07	0.03	0.02	0.08	0.06	0.01	0.03	0.08	0.15	0.06	0.01	0.02	0.05
Q8NFA0	1604	0.05	0.05	0.05	0.06	0.02	0.07	0.04	0.06	0.03	0.05	0.10	0.06	0.02	0.04	0.06	0.08	0.05	0.02	0.04	0.06
Q8TEY7	942	0.04	0.04	0.05	0.06	0.04	0.08	0.05	0.05	0.03	0.06	0.09	0.07	0.02	0.04	0.05	0.08	0.06	0.01	0.03	0.06

Q70CQ2	3546	0.06	0.04	0.05	0.06	0.03	0.08	0.05	0.04	0.04	0.05	0.12	0.05	0.03	0.04	0.04	0.09	0.05	0.01	0.03	0.05
Q9P2H5	1018	0.08	0.07	0.02	0.04	0.03	0.10	0.04	0.07	0.02	0.03	0.12	0.04	0.01	0.04	0.07	0.08	0.04	0.01	0.02	0.06
Q9P275	1123	0.07	0.07	0.03	0.04	0.02	0.05	0.05	0.07	0.03	0.02	0.08	0.08	0.02	0.02	0.08	0.11	0.06	0.01	0.02	0.05
Q86T82	979	0.05	0.05	0.05	0.06	0.02	0.09	0.05	0.05	0.02	0.05	0.10	0.08	0.02	0.03	0.05	0.12	0.06	0.01	0.02	0.04
Q8NB14	1042	0.06	0.04	0.04	0.05	0.02	0.06	0.05	0.04	0.03	0.04	0.13	0.05	0.02	0.05	0.05	0.11	0.06	0.01	0.03	0.06
Q53GS9	565	0.05	0.07	0.05	0.05	0.01	0.08	0.04	0.05	0.03	0.05	0.09	0.08	0.01	0.05	0.05	0.07	0.05	0.01	0.04	0.06
Q13107	963	0.06	0.05	0.04	0.06	0.03	0.07	0.04	0.06	0.02	0.03	0.09	0.06	0.02	0.04	0.05	0.09	0.05	0.02	0.04	0.06
Q9NVE5	1235	0.05	0.05	0.04	0.06	0.02	0.08	0.05	0.06	0.02	0.05	0.13	0.07	0.01	0.04	0.05	0.07	0.05	0.02	0.03	0.05
Q3LFD5	358	0.06	0.06	0.04	0.05	0.05	0.05	0.06	0.04	0.04	0.04	0.12	0.06	0.03	0.04	0.04	0.07	0.03	0.02	0.03	0.07
Q9H9J4	1324	0.08	0.07	0.04	0.06	0.02	0.07	0.04	0.07	0.05	0.02	0.06	0.07	0.02	0.02	0.09	0.11	0.04	0.01	0.02	0.05
Q70EL4	1123	0.07	0.08	0.03	0.04	0.02	0.05	0.06	0.08	0.02	0.02	0.11	0.04	0.01	0.04	0.09	0.11	0.04	0.02	0.02	0.05
Q9H0E7	712	0.05	0.05	0.04	0.04	0.04	0.07	0.06	0.05	0.04	0.04	0.11	0.06	0.03	0.04	0.04	0.09	0.06	0.02	0.03	0.06
Q70EL2	814	0.05	0.05	0.05	0.05	0.03	0.07	0.05	0.04	0.03	0.04	0.10	0.09	0.02	0.03	0.04	0.11	0.06	0.01	0.03	0.05
P62068	366	0.05	0.04	0.07	0.05	0.04	0.08	0.05	0.04	0.03	0.05	0.10	0.08	0.02	0.05	0.03	0.06	0.05	0.01	0.04	0.05
Q96K76	1375	0.05	0.05	0.04	0.07	0.02	0.10	0.04	0.05	0.02	0.04	0.10	0.07	0.02	0.04	0.03	0.09	0.05	0.01	0.04	0.06
Q86UV5	1035	0.05	0.05	0.05	0.05	0.04	0.10	0.06	0.05	0.02	0.05	0.10	0.08	0.02	0.03	0.04	0.07	0.04	0.02	0.03	0.05
Q70CQ1	688	0.06	0.09	0.04	0.03	0.04	0.07	0.06	0.05	0.03	0.03	0.12	0.06	0.02	0.03	0.05	0.07	0.06	0.02	0.03	0.05
P45974	858	0.07	0.05	0.03	0.07	0.02	0.08	0.04	0.07	0.02	0.05	0.08	0.06	0.03	0.04	0.07	0.07	0.04	0.01	0.03	0.07
Q70EL3	339	0.06	0.03	0.04	0.05	0.06	0.05	0.05	0.04	0.03	0.06	0.10	0.06	0.01	0.06	0.04	0.06	0.08	0.01	0.05	0.04
Q70EK9	711	0.05	0.06	0.02	0.05	0.04	0.06	0.05	0.06	0.04	0.05	0.08	0.07	0.02	0.04	0.08	0.10	0.06	0.01	0.02	0.04
Q70EK8	1073	0.04	0.05	0.06	0.05	0.03	0.07	0.05	0.05	0.04	0.04	0.08	0.07	0.02	0.04	0.05	0.12	0.05	0.01	0.02	0.04
Q70EL1	1684	0.05	0.07	0.03	0.05	0.02	0.07	0.06	0.06	0.04	0.03	0.08	0.05	0.02	0.03	0.07	0.15	0.05	0.01	0.02	0.04
P35125	1406	0.05	0.06	0.04	0.06	0.03	0.06	0.05	0.07	0.03	0.05	0.09	0.06	0.02	0.03	0.07	0.09	0.05	0.02	0.03	0.04
Q93009	1102	0.04	0.05	0.04	0.08	0.02	0.08	0.06	0.05	0.03	0.05	0.09	0.07	0.03	0.05	0.05	0.05	0.04	0.01	0.04	0.06
P40818	1118	0.06	0.06	0.04	0.05	0.02	0.09	0.05	0.04	0.02	0.05	0.08	0.10	0.02	0.03	0.06	0.08	0.06	0.01	0.03	0.05
Q93008	2554	0.06	0.05	0.05	0.06	0.02	0.07	0.05	0.05	0.03	0.05	0.11	0.05	0.02	0.04	0.05	0.07	0.04	0.01	0.03	0.06

O00507	2555	0.06	0.05	0.05	0.06	0.02	0.07	0.05	0.05	0.03	0.06	0.11	0.05	0.02	0.04	0.05	0.07	0.04	0.01	0.03	0.05
Q5W0Q7	1092	0.06	0.02	0.06	0.05	0.03	0.07	0.05	0.05	0.03	0.05	0.10	0.07	0.01	0.03	0.06	0.11	0.07	0.01	0.02	0.05
Q9UGI0	708	0.07	0.07	0.03	0.06	0.04	0.08	0.05	0.05	0.02	0.04	0.09	0.05	0.02	0.03	0.04	0.09	0.05	0.03	0.03	0.05
P03300	148	0.09	0.05	0.05	0.05	0.04	0.05	0.05	0.10	0.04	0.05	0.05	0.02	0.03	0.03	0.03	0.05	0.04	0.01	0.09	0.06
P03301	149	0.09	0.05	0.06	0.04	0.04	0.06	0.05	0.11	0.04	0.05	0.05	0.02	0.03	0.03	0.03	0.04	0.05	0.01	0.09	0.06
P03302	149	0.09	0.05	0.05	0.05	0.04	0.06	0.05	0.11	0.03	0.05	0.05	0.03	0.03	0.03	0.03	0.05	0.04	0.01	0.09	0.07
P06209	149	0.09	0.05	0.04	0.05	0.04	0.06	0.05	0.11	0.03	0.06	0.05	0.02	0.03	0.03	0.03	0.07	0.04	0.01	0.09	0.06
P06210	149	0.09	0.05	0.04	0.04	0.04	0.07	0.05	0.11	0.04	0.07	0.05	0.03	0.03	0.03	0.03	0.05	0.04	0.01	0.09	0.05
P23069	149	0.09	0.05	0.04	0.04	0.04	0.06	0.05	0.11	0.04	0.08	0.05	0.03	0.03	0.03	0.03	0.05	0.03	0.01	0.09	0.06
P22055	149	0.09	0.07	0.03	0.05	0.05	0.06	0.03	0.11	0.03	0.07	0.05	0.02	0.03	0.04	0.03	0.04	0.05	0.01	0.07	0.07
P36290	150	0.09	0.05	0.05	0.05	0.04	0.05	0.03	0.11	0.03	0.05	0.05	0.03	0.05	0.03	0.03	0.03	0.05	0.01	0.08	0.08
B9VUU3	150	0.07	0.06	0.05	0.05	0.04	0.05	0.05	0.11	0.04	0.03	0.08	0.02	0.01	0.03	0.03	0.09	0.04	0.02	0.05	0.09
Q66478	150	0.07	0.06	0.03	0.05	0.04	0.05	0.05	0.11	0.04	0.03	0.08	0.02	0.01	0.03	0.03	0.09	0.05	0.02	0.05	0.09
Q9QF31	150	0.07	0.05	0.04	0.05	0.04	0.05	0.05	0.11	0.04	0.03	0.08	0.03	0.01	0.03	0.03	0.09	0.05	0.02	0.05	0.09
Q65900	150	0.07	0.06	0.03	0.06	0.04	0.05	0.05	0.11	0.04	0.03	0.08	0.02	0.01	0.03	0.03	0.09	0.04	0.02	0.05	0.09
Q9YLJ1	150	0.06	0.06	0.04	0.06	0.05	0.07	0.05	0.12	0.04	0.03	0.07	0.01	0.01	0.03	0.03	0.05	0.05	0.02	0.05	0.11
Q66479	150	0.08	0.05	0.05	0.05	0.04	0.05	0.05	0.11	0.04	0.03	0.08	0.03	0.01	0.03	0.02	0.09	0.04	0.02	0.05	0.08
P16604	150	0.09	0.05	0.03	0.07	0.05	0.06	0.05	0.12	0.04	0.03	0.07	0.02	0.01	0.03	0.03	0.03	0.05	0.02	0.05	0.11
P08292	150	0.05	0.04	0.03	0.06	0.05	0.07	0.04	0.12	0.05	0.02	0.07	0.03	0.01	0.03	0.03	0.05	0.06	0.02	0.06	0.11
O91734	150	0.06	0.05	0.03	0.06	0.05	0.07	0.05	0.12	0.04	0.03	0.07	0.01	0.01	0.03	0.03	0.05	0.05	0.02	0.05	0.11
P21404	150	0.07	0.05	0.03	0.06	0.05	0.07	0.05	0.12	0.04	0.03	0.07	0.02	0.01	0.03	0.03	0.05	0.05	0.02	0.05	0.11
Q66849	154	0.07	0.06	0.03	0.06	0.05	0.06	0.04	0.12	0.05	0.03	0.06	0.03	0.01	0.03	0.03	0.05	0.06	0.02	0.06	0.10
P03313	150	0.07	0.05	0.03	0.05	0.05	0.07	0.05	0.12	0.03	0.05	0.07	0.01	0.01	0.03	0.03	0.06	0.05	0.02	0.05	0.09
P08291	150	0.07	0.07	0.03	0.06	0.05	0.07	0.05	0.12	0.03	0.03	0.07	0.01	0.01	0.03	0.03	0.05	0.05	0.02	0.05	0.11
Q9QL88	150	0.08	0.05	0.04	0.06	0.05	0.07	0.05	0.12	0.04	0.02	0.07	0.01	0.01	0.03	0.03	0.04	0.05	0.02	0.05	0.11
Q66282	150	0.07	0.06	0.04	0.05	0.05	0.07	0.04	0.12	0.03	0.05	0.06	0.01	0.01	0.03	0.03	0.05	0.05	0.02	0.05	0.09

Q9YLG5	150	0.07	0.05	0.03	0.06	0.05	0.07	0.05	0.12	0.04	0.03	0.06	0.02	0.02	0.03	0.03	0.05	0.05	0.02	0.06	0.12
Q66575	150	0.07	0.06	0.03	0.06	0.05	0.07	0.05	0.12	0.04	0.02	0.07	0.01	0.01	0.03	0.03	0.05	0.05	0.02	0.05	0.11
Q66577	150	0.06	0.05	0.03	0.07	0.05	0.07	0.05	0.12	0.04	0.03	0.08	0.03	0.01	0.04	0.03	0.05	0.05	0.02	0.05	0.09
Q9WN78	160	0.07	0.06	0.04	0.06	0.04	0.06	0.04	0.11	0.04	0.02	0.08	0.02	0.01	0.03	0.03	0.04	0.08	0.02	0.05	0.09
Q66474	150	0.07	0.05	0.03	0.06	0.05	0.06	0.04	0.13	0.05	0.03	0.07	0.01	0.01	0.03	0.03	0.04	0.06	0.02	0.05	0.11
Q03053	150	0.06	0.07	0.03	0.06	0.05	0.07	0.04	0.12	0.03	0.03	0.08	0.01	0.01	0.03	0.03	0.05	0.05	0.02	0.05	0.10
P12915	150	0.07	0.05	0.03	0.07	0.04	0.05	0.05	0.11	0.04	0.07	0.07	0.03	0.02	0.02	0.04	0.03	0.05	0.02	0.07	0.07
Q86887	150	0.05	0.07	0.03	0.06	0.04	0.07	0.05	0.11	0.05	0.03	0.07	0.01	0.01	0.03	0.03	0.05	0.05	0.02	0.06	0.11
P29813	150	0.06	0.05	0.03	0.06	0.05	0.07	0.04	0.12	0.04	0.03	0.07	0.02	0.01	0.03	0.03	0.05	0.05	0.02	0.07	0.11
P32537	147	0.07	0.04	0.04	0.05	0.05	0.05	0.04	0.12	0.03	0.05	0.07	0.03	0.01	0.02	0.04	0.04	0.05	0.02	0.07	0.08
P13900	150	0.09	0.05	0.03	0.07	0.05	0.06	0.05	0.12	0.03	0.03	0.07	0.02	0.01	0.03	0.03	0.03	0.05	0.02	0.05	0.11
Q68T42	147	0.08	0.05	0.03	0.05	0.04	0.05	0.05	0.13	0.03	0.08	0.07	0.01	0.01	0.03	0.05	0.03	0.05	0.02	0.05	0.07
O41174	150	0.07	0.07	0.04	0.05	0.04	0.05	0.05	0.11	0.05	0.06	0.05	0.03	0.01	0.02	0.03	0.06	0.06	0.02	0.04	0.09
Q82081	146	0.07	0.05	0.04	0.05	0.04	0.03	0.02	0.14	0.03	0.08	0.06	0.02	0.01	0.03	0.08	0.04	0.05	0.01	0.08	0.06
P03303	146	0.05	0.04	0.05	0.03	0.05	0.05	0.02	0.14	0.04	0.10	0.05	0.03	0.02	0.01	0.08	0.04	0.05	0.01	0.09	0.04
Q82122	142	0.03	0.03	0.05	0.06	0.05	0.06	0.03	0.10	0.07	0.10	0.07	0.04	0.01	0.02	0.05	0.06	0.05	0.01	0.09	0.04
P12916	142	0.03	0.04	0.05	0.06	0.05	0.07	0.04	0.09	0.06	0.08	0.08	0.04	0.01	0.02	0.05	0.06	0.06	0.01	0.09	0.04
P23008	142	0.03	0.04	0.06	0.06	0.05	0.07	0.03	0.09	0.06	0.11	0.08	0.03	0.01	0.02	0.05	0.05	0.04	0.01	0.09	0.04
P07210	142	0.03	0.04	0.05	0.07	0.06	0.07	0.03	0.09	0.06	0.08	0.08	0.04	0.01	0.04	0.04	0.06	0.04	0.01	0.07	0.05
P04936	142	0.03	0.03	0.05	0.06	0.05	0.06	0.03	0.08	0.07	0.08	0.07	0.04	0.01	0.03	0.04	0.06	0.05	0.01	0.08	0.06
P03300	183	0.08	0.04	0.05	0.04	0.01	0.05	0.04	0.11	0.03	0.08	0.06	0.05	0.03	0.03	0.04	0.04	0.10	0.00	0.03	0.08
P03301	183	0.08	0.04	0.05	0.04	0.01	0.05	0.04	0.11	0.03	0.08	0.06	0.05	0.03	0.03	0.04	0.04	0.10	0.00	0.03	0.08
P03302	183	0.08	0.04	0.05	0.04	0.01	0.05	0.04	0.11	0.03	0.08	0.06	0.05	0.03	0.03	0.04	0.04	0.10	0.00	0.03	0.08
P06209	183	0.09	0.04	0.05	0.04	0.01	0.05	0.04	0.11	0.03	0.07	0.07	0.05	0.03	0.03	0.04	0.03	0.10	0.00	0.03	0.08
P23069	183	0.08	0.04	0.05	0.04	0.01	0.05	0.04	0.11	0.03	0.07	0.07	0.05	0.03	0.03	0.04	0.03	0.11	0.00	0.03	0.08
P06210	182	0.08	0.04	0.05	0.04	0.01	0.05	0.04	0.11	0.03	0.07	0.07	0.05	0.03	0.03	0.04	0.03	0.11	0.00	0.03	0.08

P36290	183	0.08	0.04	0.05	0.04	0.01	0.05	0.04	0.11	0.03	0.07	0.07	0.05	0.03	0.03	0.04	0.04	0.10	0.00	0.03	0.08
P22055	183	0.09	0.05	0.06	0.04	0.01	0.05	0.04	0.11	0.03	0.08	0.06	0.04	0.03	0.03	0.04	0.02	0.11	0.00	0.03	0.07
P08490	183	0.07	0.05	0.07	0.04	0.01	0.07	0.02	0.11	0.03	0.04	0.10	0.05	0.04	0.05	0.04	0.03	0.05	0.01	0.03	0.08
Q68T42	183	0.07	0.05	0.05	0.07	0.02	0.03	0.03	0.10	0.03	0.07	0.07	0.04	0.03	0.05	0.04	0.02	0.10	0.00	0.04	0.09
O91734	183	0.08	0.05	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.05	0.04	0.05	0.04	0.02	0.07	0.01	0.03	0.07
Q66575	183	0.08	0.05	0.07	0.04	0.01	0.07	0.03	0.11	0.03	0.04	0.10	0.05	0.04	0.05	0.04	0.02	0.07	0.01	0.03	0.07
P03313	183	0.07	0.04	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.07	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.08
Q66577	183	0.07	0.05	0.07	0.04	0.01	0.07	0.03	0.10	0.02	0.04	0.11	0.05	0.04	0.05	0.04	0.03	0.06	0.01	0.03	0.07
Q9QL88	183	0.07	0.04	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.07	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.08
Q9WN78	183	0.08	0.05	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.09	0.05	0.04	0.05	0.04	0.02	0.07	0.01	0.03	0.08
Q9YLJ1	183	0.07	0.04	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.07	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.08
Q66282	183	0.07	0.04	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.07	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.08
Q66849	183	0.07	0.04	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.07	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.08
P08291	183	0.06	0.04	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.04	0.10	0.07	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.08
Q9YLG5	183	0.07	0.04	0.07	0.04	0.01	0.07	0.03	0.10	0.02	0.03	0.10	0.07	0.04	0.05	0.04	0.03	0.06	0.01	0.03	0.08
P21404	183	0.07	0.04	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.07	0.05	0.05	0.04	0.03	0.07	0.01	0.03	0.07
P13900	183	0.07	0.06	0.07	0.04	0.01	0.07	0.02	0.11	0.02	0.03	0.10	0.05	0.04	0.05	0.04	0.02	0.07	0.01	0.03	0.08
P16604	183	0.08	0.06	0.07	0.04	0.01	0.07	0.02	0.11	0.02	0.03	0.10	0.05	0.04	0.05	0.04	0.02	0.07	0.01	0.03	0.08
Q66474	183	0.07	0.04	0.07	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.07	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.08
Q03053	183	0.07	0.04	0.06	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.07	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.08
P08292	183	0.06	0.05	0.07	0.04	0.01	0.07	0.02	0.11	0.02	0.04	0.10	0.06	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.07
O41174	183	0.06	0.03	0.06	0.03	0.02	0.05	0.02	0.12	0.03	0.03	0.08	0.06	0.03	0.04	0.06	0.03	0.08	0.00	0.03	0.11
P29813	183	0.07	0.05	0.07	0.04	0.01	0.06	0.03	0.10	0.02	0.03	0.10	0.07	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.07
P32537	183	0.07	0.05	0.07	0.06	0.01	0.05	0.03	0.09	0.02	0.08	0.07	0.05	0.03	0.05	0.04	0.03	0.08	0.00	0.05	0.09
P12915	183	0.07	0.05	0.04	0.06	0.01	0.04	0.03	0.10	0.02	0.04	0.09	0.07	0.03	0.05	0.04	0.03	0.08	0.00	0.04	0.11
B9VUU3	183	0.04	0.05	0.05	0.05	0.01	0.04	0.05	0.10	0.03	0.05	0.10	0.04	0.03	0.05	0.04	0.05	0.08	0.01	0.02	0.10

Q86887	183	0.07	0.05	0.06	0.04	0.01	0.07	0.03	0.11	0.02	0.03	0.10	0.05	0.04	0.05	0.04	0.03	0.07	0.01	0.03	0.08
Q65900	183	0.05	0.05	0.05	0.05	0.01	0.04	0.05	0.10	0.03	0.06	0.10	0.04	0.03	0.05	0.04	0.05	0.07	0.01	0.02	0.09
Q66479	183	0.05	0.05	0.04	0.05	0.01	0.04	0.05	0.10	0.03	0.06	0.10	0.04	0.03	0.05	0.04	0.05	0.08	0.01	0.02	0.09
Q9QF31	183	0.04	0.05	0.04	0.05	0.01	0.04	0.05	0.11	0.03	0.06	0.10	0.04	0.03	0.05	0.04	0.04	0.08	0.01	0.02	0.09
Q66478	183	0.05	0.05	0.04	0.05	0.02	0.04	0.05	0.10	0.03	0.07	0.10	0.04	0.02	0.04	0.05	0.05	0.08	0.01	0.02	0.09
Q82081	182	0.03	0.05	0.07	0.06	0.02	0.05	0.04	0.10	0.02	0.09	0.10	0.07	0.01	0.04	0.04	0.03	0.10	0.00	0.02	0.05
P03303	182	0.04	0.05	0.07	0.05	0.02	0.05	0.04	0.10	0.02	0.07	0.09	0.07	0.02	0.04	0.03	0.04	0.09	0.00	0.02	0.08
P23008	183	0.03	0.05	0.07	0.07	0.02	0.05	0.03	0.11	0.02	0.09	0.09	0.05	0.01	0.03	0.04	0.04	0.07	0.00	0.05	0.07
Q82122	183	0.02	0.04	0.06	0.05	0.02	0.07	0.03	0.11	0.02	0.09	0.09	0.07	0.02	0.03	0.04	0.05	0.08	0.00	0.05	0.07
P12916	183	0.03	0.05	0.07	0.07	0.02	0.05	0.03	0.11	0.02	0.09	0.09	0.05	0.01	0.03	0.04	0.04	0.08	0.00	0.06	0.07
P04936	183	0.03	0.04	0.08	0.07	0.02	0.05	0.03	0.11	0.02	0.08	0.09	0.06	0.02	0.03	0.04	0.05	0.06	0.00	0.06	0.08
P07210	183	0.04	0.03	0.04	0.07	0.02	0.05	0.02	0.11	0.03	0.06	0.10	0.08	0.02	0.03	0.04	0.07	0.05	0.00	0.05	0.09