Search for acrylonitrile-based inhibitors of SAR-Cov-19 main and papain-like proteases through covalent docking and high-throughput virtual screening.

A thesis submitted in partial fulfillment of the requirements for the degree

of

Master of Science in Bioinformatics

of

Rhodes University, South Africa

Department of Biochemistry and Microbiology

Faculty of Science

by

Yamkela Ntantiso

ABSTRACT

The sudden outbreak of SARS-CoV-2 formerly known as the 2019 novel coronavirus (2019-nCoV) quickly turned into a pandemic of coronavirus disease 2019 (COVID-19), the scale of which has never been seen before. High infection rates and mortality from COVID-19 placed pressure on global health services, and this has been to the detriment of the global economy. However, treatment options for COVID-19 are still very limited; hence, it is now as important as ever that researchers explore searching for new compounds with pharmacokinetic properties that inhibit the two COVID proteases - the main protease (Mpro) and the papain-like protease (PLpro).
The main protease is a cysteine protease; as such, it is susceptible to permanent inhibition by reactive species (warheads) that may covalently bind to cysteine residues. One such class of compounds is acrylonitriles, in which the reactive acrylonitrile is reactive towards cysteine through a Michael addition reaction. The resulting covalent interaction is permanent and inactivates the cysteine residue and hence the protease within the context of the COVID-19 life-cycle.

In this context, this study seeks to utilize computational-based approaches to identify acrylonitrile-based inhibitors of coronavirus drug targets. To do this, the ZINC database has been screened for compounds containing acrylonitrile functionality, due to its known nature as a warhead that binds to cysteine residues. Pharmacokinetic properties are computed to evaluate the viability of identified inhibitors, and covalent and non-covalent molecular docking approaches to the Mpro enzyme crystal structure have also been used to assess the identified systems. To gather more information and evaluate the most promising systems, a subset of the most promising compounds have been subjected to molecular dynamics simulation (for both covalently bound and non-covalently bound systems).

DECLARATION

I, Yamkela Ntantiso, student number g22n8724, declare that this is my own work that has not been previously submitted to this or any other university.

Signed: Y. Ntantiso
Date: February 2024

DEDICATION

This work is dedicated to Lindiwe Ntantiso, my mother, and my sister, Yandiswa Ntantiso.

Your love and support have been a pillar of strength.

ACKNOWLEDGEMENTS

I would like to thank Rhodes University for funding my MSc. studies, and much thanks to the Research Unit in Bioinformatics (RUBi) for the opportunity. Gratitude to the Center of High-Performance Computing (CHPC) for providing the computational resources to perform this study.

To Victor Barozi, the team at RUBi, and fellow candidates at the Chemistry department, especially Washington Dendera, your unwavering support and eagerness to assist are much appreciated.

To my family and friends, I would not be where I am today without your love and support.

I would like to express my eternal gratitude to my supervisor, Prof. Kevin Lobb. I truly appreciate all your support and guidance leading to the completion of this project. The opportunity to work with you has been the best experience of my post-graduate career. With the lessons and guidance you have provided, I believe I can overcome any obstacle, not only in research but also in life. Thank you so much.

TABLE OF CONTENTS

LIST OF FIGURES

Figure 10: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 48356 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the red indicating unfavorable bump interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

Figure 11: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 117238 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the red indicating unfavorable bump interactions, the light lime showing carbon-hydrogen bond interactions, blue showing halogen interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

Figure 12: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 337222 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the red indicating unfavorable bump interactions, the light lime showing carbon-hydrogen bond interactions, blue showing halogen (fluorine) interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

Figure 13: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 387305 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the red indicating unfavorable bump interactions, the light lime showing carbon-hydrogen bond interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the

non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

Figure 14: DiscoveryStudio generated 3D (left) & 2D (right) diagrams of Ligand 396939 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the red indicating unfavorable bump interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

Figure 15: DiscoveryStudio generated 3D (left) & 2D (right) diagrams of Ligand 397136 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the red indicating unfavorable bump interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

Figure 16: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 397730 covalently docked to the receptor

Figure 17: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 402091 covalently docked to the receptor

Figure 18: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 403456 covalently docked to the receptor

Figure 19: DiscoveryStudio generated 3D (left) & 2D (right) diagram of ligand 48356 non-covalently docked to chain a of the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the blue showing Halogen interactions, darker orange indicating Pi-Anion, the lime color indicating Pi-Donor Hydrogen Bond and the light orange indicating Pi-Sulfur

interactions. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

Figure 20: A 3D (left) & 2D (right) diagram of ligand 48356 non-covalently docked to chain b of the receptor

Figure 21: A 3D (left) & 2D (right) diagram of ligand 117238 non-covalently docked to chain a of the receptor.

Figure 22: A 3D (left) & 2D (right) diagram of ligand 117238 non-covalently docked to chain B of the receptor.

Figure 23: A 3D (left) & 2D (right) diagram of ligand 337222 non-covalently docked to chain a of the receptor.

Figure 24: A 3D (left) & 2D (right) diagram of ligand 337222 non-covalently docked to chain b of the receptor.

Figure 25: A 3D (left) & 2D (right) diagram of ligand 387305 non-covalently docked to chain a of the receptor.

Figure 26: A 3D (left) & 2D (right) diagram of ligand 387305 non-covalently docked to chain b of the receptor.

Figure 27: A 3D (left) & 2D (right) diagram of ligand 396939 non-covalently docked to chain a of the receptor.

Figure 28: A 3D (left) & 2D (right) diagram of ligand 396939 non-covalently docked to chain b of the receptor.

Figure 29: A 3D (left) & 2D (right) diagram of ligand 397136 non-covalently docked to chain a of the receptor.

Figure 30: A 3D (left) & 2D (right) diagram of ligand 397136 non-covalently docked to chain b of the receptor.

Figure 31: A 3D (left) & 2D (right) diagram of ligand 397730 non-covalently docked to chain a of the receptor.

Figure 32: A 3D (left) & 2D (right) diagram of ligand 397730 non-covalently docked to chain b of the receptor.

Figure 33: A 3D (left) & 2D (right) diagram of ligand 402091 non-covalently docked to chain a of the receptor.

Figure 34: A 3D (left) & 2D (right) diagram of ligand 402091 non-covalently docked to chain b of the receptor.

Figure 35: A 3D (left) & 2D (right) diagram of ligand 403456 non-covalently docked to chain a of the receptor.

Figure 36: A 3D (left) & 2D (right) diagram of ligand 403456 non-covalently docked to chain b of the receptor.

Figure 37: Shows 80 out of the 403804 high-throughput screening compounds with acrylonitrile functional group from the ZINC database before they were filtered.

Figure 38: Knime workflow of covalent docking analysis and visualization.

LIST OF TABLES

LIST OF TOOLS, LIBRARIES AND WEB SERVERS

AutoDock Vina

BIOVIA Discovery Studio Visualizer

GAUSSIAN 09

Grace (Xmgrace)

GROMACS

JupyterHub

PyMOL

RCSB Protein Data Bank: http://www.rcsb.org/pdb/home/home.do

RDKit

Schrödinger Suites

Visual Molecular Dynamics

ZINC Database: https://zinc.docking.org/

LIST OF ABBREVIATIONS

Severe acute respiratory syndrome (SARS-CoV)

Middle East respiratory syndrome coronavirus (MERS-CoV)

Receptor binding domain (RBD)

Transmembrane protease serine 2 (TMPRSS2)

Dipeptidyl peptidase 4 (DPP4)

Angiotensin-converting enzyme 2 (ACE2)

World Health Organisation (WHO)

Open reading frames (ORF)

Chymotrypsin-like protease (3CL pro)

Main protease (M pro)

Transcription regulatory sequence (TRS)

Hemagglutinin-esterase (HE)

SARS-related coronaviruses (SARSr-Cov)

SARS-CoV-2 S B for human ACE2 (hACE2)

Papain-like protease (PL pro)

Protein Data Bank (PDB)

Quantitative estimation of drug-likeness (QED)

Root Mean Square Fluctuation (RMSF)

Root Mean Square Deviation (RMSD)

Radius of gyration (Rg)

Coronavirus disease 19 (Covid-19)

Centre of High-Performance Computing cluster (CHPC)

GROningen MAchine for Chemical Simulations (GROMACS)

Principal component analysis (PCA)

Protein Data Bank (PDB)

Zinc Is Not Commercial (ZINC)

CHAPTER ONE

LITERATURE AND STUDY BACKGROUND

1.1 Introduction

Coronaviruses (CoVs) are enveloped single-stranded positive-sense RNA (ssRNA) viruses that belong to the *Coronaviridae* family in the *Nidovirales* order (International Committee of Taxonomy of Viruses). Coronaviruses are divided into the alpha, beta, gamma, and sigma subgroups, with the alpha and beta being infectious to humans (Mousavizadeh and Ghasemi, 2020; Shereen *et al*., 2020) (Figure 1). Evolutionary trend analysis of coronaviruses has revealed that alpha coronaviruses and beta coronaviruses originated from bats and rodents, and they infect only mammals, while gamma coronaviruses and delta coronaviruses were found to have originated from avian species and infect only birds (Ge *et al*., 2017) however, some of them do infect mammals. The fact that coronaviruses can cross the species barrier has resulted in some of the pathogenic coronaviruses, and the two highly pathogenic viruses are the severe acute respiratory syndrome coronavirus (SARS-CoV) and The Middle East Respiratory Coronavirus (MERS-CoV). The taxonomy of coronaviruses is after their spherical morphology that is 65 - 125 nm in diameter, with proteins that protrude at the surface giving a crown-like effect. These ssRNA viruses which comprise severe acute respiratory syndrome coronavirus (SARS-CoV), H5N1 influenza A, H1N1, and Middle East respiratory syndrome coronavirus (MERS-CoV) have posed great challenges to human health (Reperant and Osterhaus, 2017). Infection with these viruses can lead to pneumonia-like symptoms, acute lung injury, acute respiratory and renal failure, and ultimately death in some cases (Shereen *et al.*, 2020).

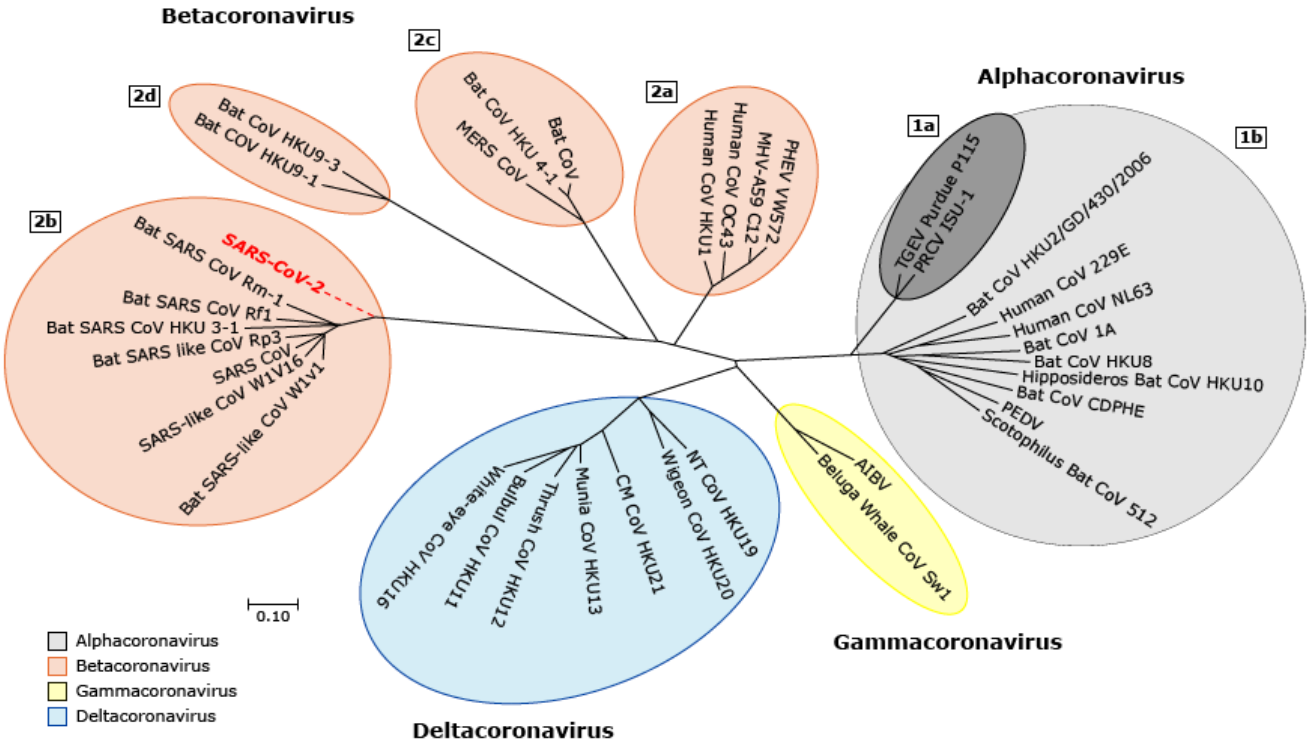


Figure 1: Coronaviruses belong to the subfamily *Coronavirinae* from the family *Coronaviridae*. The viruses in this subfamily group into four genera (prototype or representative strains shown): Alphacoronavirus (purple), Betacoronavirus (pink), Gammacoronavirus (green) and Deltacoronavirus (blue). Classic subgroup clusters are labeled 1a and 1b for the alpha coronaviruses and 2a–2d for the beta coronaviruses. The tree is reconstructed with sequences of the complete RNA-dependent RNA polymerase-coding region of the representative coronaviruses (maximum likelihood method under the GTR + I + Γ model of nucleotide substitution). Only nodes with bootstrap support above 70% are shown. Produced as found in Cui et al. (2019). https://www.uptodate.com/contents/coronaviruses/print

1.2 Origin and evolution of SARS-CoV

In 2002-2003, SARS-CoV sprung up in Guangdong Province, China, with 8000 clinical cases and 800 deaths. Since 2012, MERS-CoV has caused persistent epidemics in the Arabian Peninsula. SARS-CoV uses angiotensin-converting enzyme 2 (ACE2) as a receptor and primarily infects ciliated bronchial epithelial cells and type II pneumocytes (Li, 2003; Qian *et al.*, 2013), whereas MERS-CoV uses dipeptidyl peptidase 4 (dpp4; also known as CD26) as a receptor and infects non-ciliated bronchial epithelial cells and type II pneumocytes (Lu *et al*., 2013; Raj *et al*., 2013; Scobey *et al*., 2013). Both viruses have been found to originate from bats and the transmitted into intermediate mammalian host civets in the case of SARS-CoV and camels in the case of MERS-CoV and eventually infected humans (Song *et al*., 2019). The emergence of MERS-CoV in Middle Eastern countries in 2012 further illustrated the ability to infect people, with a combined death toll of 1616 people, from infection with these β coronaviruses (Rhaman and Sarkar, 2019; Shereen *et al.*, 2020). Reports of pneumonia with an unfamiliar etiology coming out of Wuhan China, in 2019, rapidly lead to the identification of a novel β coronavirus with high infection rates in the first 50 days (Li *et al*., 2020; Wang *et al*., 2020). Early infections were limited to people who had contact with the Huanan seafood market of Wuhan, which sells animals noted for being zoonotic reservoirs of β coronaviruses (Riou and Althaus, 2020; Shereen *et al*., 2020). Spread of the disease between people through coughing, sneezing and aerosols that enter the body through inhalation resulted in infection of people who had no contact with the Huanan market (Shereen *et al.*, 2020; Phan *et al*., 2020).

The global pandemic of coronavirus disease 2019 (COVID-19) was first reported on 31 December 2019 by the World Health Organization (WHO) country office following a cluster of pneumonia cases in Wuhan City, Hubei Province of China with 1,844,683 confirmed cases and 117,021 deaths globally by 14th April 2020 (World Health Organization, 2020). To characterize the novel coronavirus, bronchoalveolar lavage fluid and throat swabs were collected from nine patients who had visited the Wuhan seafood market during the initial outbreak. Special pathogen-free human airway epithelial (HAE) cells were used for virus isolation. The collected samples were inoculated into the HAE cells through the apical surfaces. HAE cells were monitored for cytopathic effects and supernatant was collected to perform real-time (RT)-PCR assays. Apical samples were collected for next-generation sequencing after three passages. The whole-genome sequences of SARS-CoV-2 were generated by a combination of Sanger, Illumina, and Oxford nanopore sequencing (Lu *et al.* 2020). Phylogenetic

analysis has revealed that bats might be the source of SARS-CoV-2 (Andersen *et al*., 2020). To date, on 11 January 2024, WHO has reported over 701M cases of coronavirus with over 6,97M deaths reported worldwide and it is continuing to grow with the rising number of cases worldwide.

When the SARS epidemic emerged initially, almost all early indicator patients had been exposed to animals before developing the disease. After the agent that was causing SARS was identified, SARS-CoV and/or anti-SARS-CoV antibodies were found in masked palm civets (Paguma larvata) and animal handlers in a marketplace (Guan *et al.*, 2003: Kan *et al*., 2005; Tu *et al*., 2004; Wang *et al*., 2004; Xu *et al*., 2004; Song *et al*., 2005). The genome sequences of SARS-CoVs from market civets were found to be almost identical to the genomes of human SARS-CoVs (Song *et al*., 2005; Chinese, 2004). However, the two genes show major variation.

The first variable region is located in the S gene (Figure 2). The S gene encodes a surface protein, the spike protein, which is a homotrimeric glycoprotein complex that is essential for infectivity. This complex consists of two subdomains. The SARS-CoV S protein is functionally divided into two subunits, denoted S1 and S2, which are responsible for receptor binding and fusion with the cellular membrane, respectively (Masters & Perlman, 2013). The S1 contains a receptor-binding domain (RBD) with high binding/interaction strength for mammalian angiotensin-converting enzyme 2 (ACE 2) and is further divided into the amino-terminal domain (S1-NTD) and the carboxy-terminal domain (S1-CTD). The S1-CTD functions as the receptor-binding domain (RBD) and is responsible for binding ACE2 and entering cells (Li *et al*., 2003; Babcock *et al.*, 2004: Wong *et al*., 2004). Two amino acid residues in the RBD, 479 and 487, were identified to be very important for ACE2-mediated SARS-CoV infection and critical for virus transmission from civets to humans (Li *et al*., 2005; Qu *et al*., 2005).

The second major location of variation is the accessory gene, orf8. Because of the SARS spread, the SARS 2002–2003 outbreak could be divided into three phases, with the early phase characterized by a limited number of localized cases, followed by a middle phase during which a super spreader event occurred in a hospital, and finally the late phase of international spread (Chinese, 2004). The viral genomes from early-phase patients contain two genotypes of orf8, one with a complete orf8 (369 nucleotides) and the other containing an 82-nucleotide deletion. By contrast, viral genomes from late-phase patients and most of the genomes from middle-phase patients contain a split orf8 (orf8a and orf8b) owing to a 29-nucleotide deletion; two exceptions were found in middle-phase genomes, one containing an 82-nucleotide deletion in orf8 and the other with the whole orf8 deleted. The human

isolates from 2004 and all civet SARS- CoV genomes have a complete orf8 except one civet strain with an 82-nucleotide deletion (Chinese, 2004). These data indicate that the orf8 gene was adapted during animal-to-human transmission during the SARS epidemic. A limited functional analysis showed that the ORF8a protein is essential for SARS-CoV replication in her Vero-E6 cells, but has a role in regulating endoplasmic reticulum stress, inducing apoptosis, and inhibiting interferon responses in host cells. It was suggested that it may have worked (Hu *et al*., 2017; Le *et al*., 2007; Oostra *et al*., 2007; Wong *et al.*, 2018; Sung *et al*., 2009; Chen *et al*., 2007). Whether and how these adaptations contributed to the virulence of SARS-CoV is not fully understood.

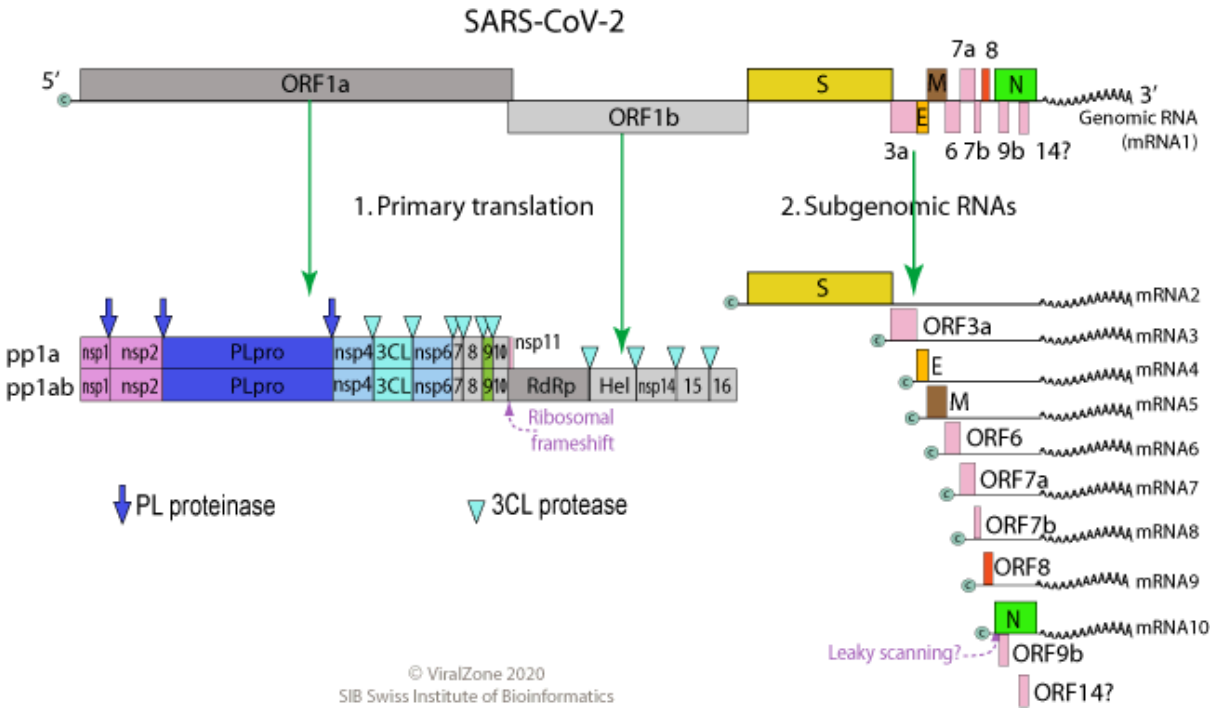


Figure 2: The genome of SARS-CoV-2 shows transcription sites and protein-coding domains. Source: https://viralzone.expasy.org/resources/nCoV_genome_bis.png

1.3 Morphology and Genome Organization of SARS-CoV-2

To resolve the structure of SARS-CoV-2, Park *et al.* (2020) isolated SARS-CoV-2 from the nasopharyngeal and oropharyngeal and the samples were inoculated on Vero cells. To identify SARS-CoV-2, inoculated cells were prefixed using 2% paraformaldehyde and 2.5% glutaraldehyde, and transmission electron microscopy was performed. The structure of SARS-CoV-2 was then observed by examining infected cells 3 days after they were infected. Electron microscopy revealed the coronavirus-specific morphology of SARS-CoV-2 with virus particle sizes ranging from 70 to 90 nm observed under a wide variety of intracellular organelles, most specifically in vesicles (Park *et al.* 2020). Due to the high sequence similarity observed, the structure of SARS-CoV-2 was then speculated to be the same as SARS-CoV (Kumar *et al*. 2020). As a novel beta coronavirus, SARS-CoV-2 shares 79% genome sequence identity with SARS-CoV and 50% with MERS-CoV. The genetic similarity between SARS-CoV-2 and SARS-CoV could help in developing possible treatments that are still unavailable, for a disease that has devastated the world. The surface viral protein spike, membrane, and envelope of coronavirus are embedded in a host membrane-derived lipid bilayer encapsulating the helical nucleocapsid comprising viral RNA (Finlay *et al.* 2004). The structure of spike (Yan et al. 2020) and protease of SARS-CoV-2 (Zhang *et al*. 2020) has been resolved, which provides an opportunity to develop a newer class of drugs for the treatment of COVID-19.

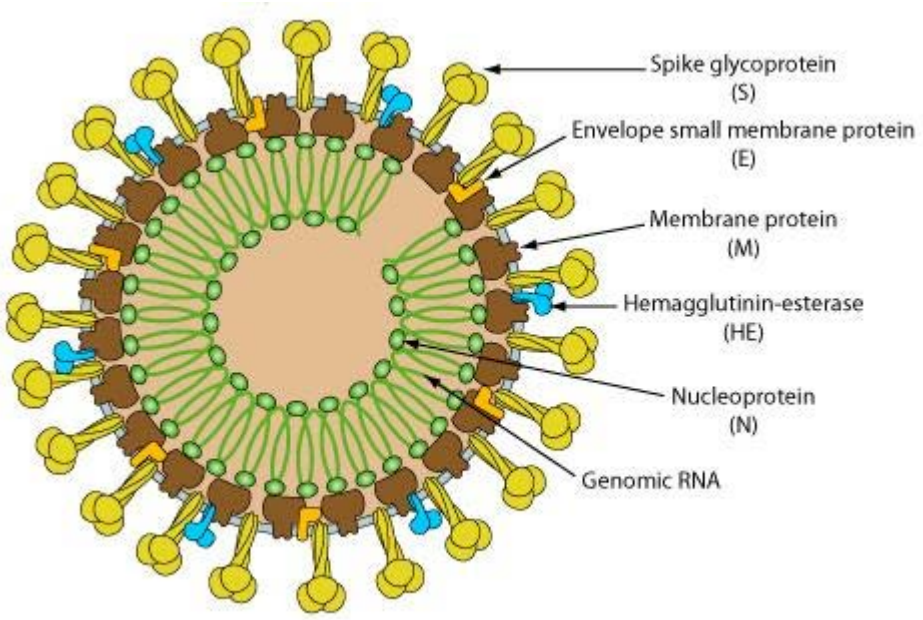


Figure 3: General coronavirus schematic as retrieved from Biowiki (http://ruleof6ix.fieldofscience.com/2012/09/a-new-coronavirus-should-you-care.html).

SARS-Cov-2, much like other coronaviruses, is a spherical enveloped particle (120 -160 nm in diameter) containing +ssRNA (Figure 3), associated with a nucleoprotein within a capsid consisting of matrix proteins (Woo et al., 2010). The viral envelope has club-shaped (homotrimer) spike glycoprotein protrusions, with some coronaviruses additionally containing hemagglutinin-esterase (HE) protein on the surface (De Haan *et al.,* 1998; Walls *et al.,* 2020). The mature proteins of SARS-CoV-2 are responsible for genome maintenance and viral replication (van Boheemen et al., 2012). Coronaviruses have the largest genomes (26–32 kb) among all RNA virus families and comprise 6-11 open reading frames (ORFs) encoding 9680 amino acid polypeptides (Guo *et al.*, 2020). Each viral transcript has a 5′-cap structure and a 3′ poly (A) tail (Lai, M.M. & Stohlman, S.A., 1981). The six functional open reading frames (Figure 2) are arranged in order from 5’ to 3’: replicase (ORF1a/ORF1b), spike (S), envelope (E), membrane (M), and nucleocapsid (N) and accessory proteins like ORFs.

1.3.1 The Spike Protein (S)

The spike (S) proteins protrude from the viral surface and give the coronavirus family the characteristic namesake crown-like appearance under electron microscopy (Woo *et al.*, 2010). This type 1 transmembrane protein with a signal peptide is used for receptor binding and viral entry and has the most variable sequence in the coronavirus genome (Belouzard et al., 2009; Woo et al., 2010). S determines the host range and cell tropism and is the main target of neutralizing antibodies during infection (Du *et al.*, 2009). S proteins consist of two functional subunits, with the S1 subunit responsible for binding to the host cell receptor and the S2 subunit responsible for the non-covalent perfusion conformation for the fusion of the viral and cellular membranes (Kirchdoerfer *et al.*, 2016; Walls *et al.*, 2016). Cleavage by furin-like host proteases between the S1 and S2 subunits during biogenesis occurs in some coronaviruses (Bosch *et al.,* 2003). After virion uptake by the target host cell, endo-lysosomal proteases at the S2' cleavage site cleave the S protein further to result in the activation of coronavirus S proteins (Burkard *et al.*, 2014; Walls *et al.,* 2016). The domains used within the S1 subunit to recognize attachment and entry receptors are dependent on each viral species (Walls *et al.*, 2020). SARS-CoV and SARS-related coronaviruses (SARSr-Cov), such as SARS-CoV-2, interact directly with angiotensin-converting enzyme 2 (ACE2) via SB to enter the targeted host cell (Kirchdoerfer *et al*., 2018; Song *et al.,* 2018; Walls *et al*., 2020). Walls *et al.* (2020) report that the affinity of SARS-CoV-2 SB for human ACE2 (hACE2) is high, which would explain the efficient transmission of the virus between people (Walls *et al.*, 2020; Wan *et al.,* 2020) - especially when contrasted with SARS-CoV which interacts weakly and has a lower pathogenicity and transmissibility (Li *et al*., 2005). This would make hACE2 a functional receptor for the virus, making it a promising drug target - as aforementioned, it is already attacked by antibodies in attempts to neutralize infection. Findings by Walls *et al*. (2020) of a furin cleavage site at the S1 and S2 subunits, which has already been noted above as a feature found only in some coronaviruses, set SARS-CoV-2 apart from SARS-CoV and SARSr-Cov. Disruption of the furin cleavage motif by Walls *et al.* (2020) resulted in moderate disruption of SARS-CoV-2 S-mediated entry into VeriE6 and BHK, with an associated warning that this might have the adverse effect of reducing the cell specificity of SARS-CoV-2.

1.3.2 The Envelope (E) protein

Envelope membrane (E) proteins are a group of relatively small viral proteins that help in the assembly and release of the virions (Fehr A.R., Perlman S., 2015). Among the structural proteins of the SARS-CoV-2, E protein is considered a potential drug target. The E protein is relatively small (75 aa) and plays a significant role in viral morphogenesis and assembly (Li S. *et al.*, 2020). The E protein is known to act as a viroporin that assembles into host membranes, forming protein-lipid pores involved in ion transport. The sequences of the E protein for all four strains are highly conserved regions among the BAT-CoV, SARS-CoV, and SARS-CoV-2 while exhibiting a slight variation in the sequence of the MERS-CoV envelope proteins.

1.3.3 Membrane (M) protein

M proteins are 222 amino acid-long structural proteins that function in concurrence with E, N, and S proteins and play a major role in RNA packaging. The conserved stretch of amino acids suggests a common architecture for these proteins. M proteins are the most abundant viral proteins of CoVs that are involved in providing a distinct shape to the virus. The MSA profile of the M protein shows higher sequence conservation among BAT-CoV, SARS-CoV, and SARS-CoV-2. However, a considerable variation in the sequence of the M protein of the MERS-CoV strain was observed. The presence of three transmembrane domains is a distinct feature of M proteins.

1.3.4 Nucleocapsid (N) protein

Nucleocapsid proteins (N) play an important role in the packaging of viral RNA into ribonucleocapsids. The protein of SARS-CoV-2 is highly conserved across CoVs, sharing ~90% sequence identity with that of SARS-CoV. It mediates viral assembly by interacting with the viral genome and M protein, which are helpful in the augmentation of viral RNA transcription and

replication. Thus, N proteins are considered potential drug targets. The N proteins bind to viral RNA through their ~140 amino acid long RNA-binding domain in their core in a “bead on a string” manner. High levels of conservation are seen in the N protein MSA profiles from BAT-CoV, SARS-CoV, and SARS-CoV-2. Given the high sequence similarity of the N protein, it is possible that antibodies directed against the SARS-CoV N protein would also likely recognize the SARS-CoV-2 N protein. A similar pattern has been observed for the MERS-CoV strain, where regions of slight sequence variations suggest its divergence in the evolutionary process.

1.3.5 Haemagglutinin esterase (HE)

As mentioned above, the HE gene is a viral protrusion present in some coronaviruses. The Betacoronavirus subgroup A (Human coronavirus OC43; Bovine coronavirus; Porcine hemagglutinating encephalomyelitis virus; Equine coronavirus) all have an HE gene that encodes a glycoprotein with neuraminate O-acetyl-esterase activity, downstream of ORF1ab and upstream of the S gene (Woo *et al.*, 2010). The presence of this gene in only subgroup A of betacoronaviruses suggests a heterologous recombination event occurred in the ancestors of this subgroup with the influenza C virus (Luytjes *et al.*, 1988).

1.4 Key features and entry mechanisms of SARS-CoV-2

The life cycle of SARS-CoV-2 in host cells begins with the binding of spike (S) protein to angiotensin-converting enzyme 2 (ACE2) as the cell receptor. The cellular serine protease TMPRSS2 has the ability to cleave the spike protein between its S1 and S2 domains in close proximity to the ACE2 receptor. This process triggers the fusion of the viral membrane with the plasma membrane (Inhibitor II: Camostat) via the endocytosis pathway (Hoffmann *et al*., 2021). Then SARS-CoV-2

releases the genome RNA to enter the host cell. The genomic RNA is then translated into viral replicase polyproteins pp1a and pp1ab, which are cleaved by the papain-like protease (PLpro, Nsp3) and 3C-like protease (3CLpro, Nsp5) (two proteases that are crucial for virus replication) to form functional non-structural proteins (NSPs) such as Helicase or the RNA replicase–transcriptase complex (RdRp). PLpro cleaves at its LXGG recognition sites at nsp1, nsp2, and nsp3, while Mpro cleaves the remaining downstream non-structural proteins (nsp4-16). The replication of coronavirus involves ribosomal frameshifting during the translation process and generates a series of sub-genomes mRNA by discontinuous transcription that encodes for relevant viral proteins and eventually translates into related viral proteins. The viral proteins and genome RNA of SARS-CoV-2 subsequently assembled into viral particles in the endoplasmic reticulum (ER) and Golgi apparatus and then transported through vesicles and released out of the cells (Sheeren *et al.,* 2020).

The processing of the polyproteins (pp1a and pp1ab) is mainly through the activity of the 3C-like protease (3CLpro) and papain-like protease (PLpro) (Thiel *et al.,* 2003). The essential function of the proteases in the lifecycle of coronaviruses has made them a target for the development of antiviral drugs targeted at viral replication. The papain-like protease (PLpro) active site consists of a catalytic triad. PLpro functions by cleaving ISG15, a two-domain Ub-like protein, and Lys48-linked polyUb chains. Hence, their main function lies in the processing of the viral polypeptide into functional proteins, which further deubiquitinase and dampen host anti-viral reactions by hijacking ubiquitin (Ub), an enzyme playing a pivotal role in host defense mechanisms.

The targeting of PLpro for the treatment of viral infection from coronaviruses has been rare when compared to the main protease because the structure of this membrane-associated enzyme has remained elusive to researchers for some time (Ratia *et al.,* 2006). Attempts to resolve this and produce viable drugs that evade the issues of toxicity and lack of specificity that often plague cysteine protease drug inhibitors have included attempts at elucidating the structure of PLpro (Ratia *et al*., 2006; Barreto *et al.*, 2005). Purification of the catalytic domain of PLpro revealed the deubiquitination and de-ISGylating enzyme activity of the enzyme, with the role of this activity remaining unclear (Barreto *et al.*, 2005; Lindner *et al*., 2005; Devaraj *et al.*, 2007). The crystal structure solved by Ratia et al. (2006) for SARS-CoV PLpro of nsp3 revealed an intact catalytic triad, a zinc-binding domain, and an N-terminal ubiquitin-like domain. The catalytic triad is suggested to be Cys1651-His1812-Asp1826 by Barretto *et*

al. (2005). This structural information, although sparse when compared to the main protease, provides a basis for drug targets aimed at PLpro.

Mpro, also termed 3CL protease, is a 33.8-kDa cysteine protease that mediates the maturation of functional polypeptides involved in the assembly of replication-transcription machinery (Wang H. *et al*., 2016). Mpro digests the polyprotein at no less than 11 conserved sites, starting with the autolytic cleavage of this enzyme itself from pp1a and pp1ab. In addition, Mpro has no human homolog and is highly conserved among all CoVs (Yang *et al.*, 2006). These above features make it an attractive drug target against CoVs. The Mpro consists of 306 amino acids and has a high structural and sequence resemblance to that of the SARS-CoV Mpro. SARS-CoV-2 Mpro monomer comprises three domains (i.e., N-terminal domain-I, N-terminal domain-II, and C-terminal domain-III), with domains 1 and 2 forming the chymotrypsin structure and the third domain consisting of α-helices – the substrate or inhibitor binding site is found between the second and third domains (Hsu *et al.*, 2005; Pillaiyar *et al*., 2016). The Mpro active site consists of the catalytic dyads C145 and H4. According to Ziebhur *et al.* (2000), the catalytic dyad of the main protease consists of conserved His and Cys residues. The cysteine residue of the Cys-His dyad undergoes nucleophilic attack on the reactive atom of the substrate, while the histidine residue helps to stabilize the intermediate state. Around this dyad, Mpro forms a conserved binding pocket that is composed of four subsites (S1′, S1, S2, and S4),well accommodating the substrate (Xue *et al*., 2008).

Figure 4: Crystal structure of free SARS-CoV-2 M pro solved at 1.75 Å resolution (PDB entry: 6Y2E (Zhang *et al.*, 2020a)).

The main protease of coronaviruses is a potential drug target since it is responsible for the maturation of itself and other important polyproteins (Ziebuhr *et al.*, 2000). SARS-CoV-2 has 14 open reading frames (ORFs). The M pro (nsp5), encoded by the major ORF1ab, cleaves two overlapping polyproteins (pp1a and pp1ab) into 16 non-structural proteins, which are important for viral replication and maturation (Paul, 2006; Ziebuhr *et al*., 2000; Chen et al., 2020b; Gordon *et al.*, 2020). In addition, it plays a significant role in virus entry to host cells, where inhibition of this enzyme halts the viral entry and subsequent infection (Jain and Mujwar, 2020). These important functions of the viral protease enzyme itself are an interesting therapeutic target for curbing coronavirus-associated diseases (Thiel *et al.*, 2003; Naqvi *et al.*, 2020).

1.5 Inhibitors targeting the main protease and papain-like protease

High-throughput assays have been effectively used for large-scale screening of existing drugs to identify potential antiviral leads for SARS-CoV-2. From a virtual structure-based, high-throughput screening of a library of roughly 10,000 chemicals, carbofur and ebselen that inhibit SARS-CoV-2 infection of Vero cells were discovered (Jin *et al*., 2020). Similarly, Apilimod, MDL-28170, and ONO 5334 that inhibit SARS-CoV-2 were identified by profiling a library of 12,000 clinical-stage or Food and Drug Administration (FDA)- approved small molecules (Riva *et al.*, 2020). Inhibiting protease activity is how the medication combination lopinavir-ritonavir prevents and treats HIV infection (Nishiga *et al.*, 2020). Lopinavir showed in vitro activity against SARS-CoV and has been effective in improving the clinical outcome of MERS in nonhuman primates (Chan *et al.*, 2015). In addition, viral and host-factor-targeting agents, combined with drugs that directly target viral enzymes, could lead to a therapeutic regimen to treat COVID-19 (Gordon *et al.*, 2020). Camostat mesylate, which inhibits the plasma membrane-associated host serine protease, TMPRSS2, has been shown to block the SARS-CoV cell entry mechanism (Zhou *et al.,* 2015). Antivirals such as Remdesivir, Favipiravir, and Galidesivir, targeting RdRP, have shown inhibitory activities against SARS-CoV-2 (Sheahan et al., 2020). Remdesivir was granted emergency use authorization for SARS-CoV-2 from the U.S. FDA on May 1, 2020 (Eastman *et al*., 2020). Although remdesivir can shorten infection times and may have clinical benefits in patients with severe COVID-19, it did not significantly improve survival (Grein *et al.*, 2020; Wang *et al*., 2020). An oral RdRP inhibitor, Molnupiravir (MK-4482, EIDD-2801) was found effective in patients early in the course of their illness (Fischer *et al*., 2021). Molnupiravir has been authorized by the FDA and the United Kingdom Medicines and Healthcare Products Regulatory Agency (MHRA) to treat mild-to-moderate COVID-19.

1.6 PURPOSE OF STUDY

1.6.1 PROBLEM STATEMENT AND HYPOTHESIS

Countries are experiencing a return to climbing infection and mortality rates from COVID-19 as of December 2023, which has placed global health services and economies under unprecedented strain when the pandemic was at its peak. Initially, antivirals and antibiotics were used as treatments, either combined or alone. Repurposed drugs (which are drugs that were originally designed for other diseases) such as remidesivir, chloroquine, and interferon β have had variable success in blocking SARS-CoV-2 replication (Pushpakom et al., 2019). Other antiviral drugs that have shown good results in clinical trials are ab *initio-designed* drugs based on structure characterization. Even though efforts to identify possible SARS-CoV-2 drug targets and discover the first agents to modulate them have culminated and the vaccines gave hope in controlling the pandemic, the emergence of new variants of the wild-type strain of SARS-CoV-2 that could hamper the effectiveness of the developed vaccines proves the main reason behind this study, which is the urgent need to develop first effective drugs to treat COVID-19 patients. The mutation of SARS-CoV-2 into the detected variants demands a continued look into treatment strategies. Computational techniques offer an approach for conducting high-throughput studies that can direct wet work research to efficiently direct resources. In this study, the papain-like and main protease of SARS-CoV-2 which are potential drug targets from the latest published data were identified, and the initiation of the development of therapies against COVID-19 on these targets was attempted through molecular docking of covalent and non-covalent inhibitors and molecular dynamic studies using compound databases.

1.6.2 AIM

This study aims to use high-throughput computational-based approaches to search for acrylonitrile-based covalent and non-covalent inhibitors of SARS-CoV-2 drug targets.

1.6.3 OBJECTIVES

1. Identification and retrieval of SARS-CoV-2 protease crystal structures from NCBI.

2. High-throughput screening (HTS) of compounds with acrylonitrile functional groups from the ZINC database that result in covalent and non-covalent bonding with the SARS-CoV-2 main protease.

3. Covalent and non-covalent molecular docking of screened compounds with the SARS-CoV-2 main protease and evaluation and identification of the most promising compounds.

4. Non-covalent molecular dynamic simulations of:
    - Protein
    - Protein-bound conjugates

Further analysis of these is also needed to determine more comprehensively the properties of promising potential inhibitors of the main protease.

5. Covalent molecular dynamics of protein-bound conjugates, to determine any changes in the structure of the main protease due to covalent binding

1.7 METHODOLOGY OVERVIEW

1.7.1 Protein Preparation

The crystallographic structures for 3CL pro of SARS-CoV-2 were retrieved from the Protein Data Bank (PDB) (Bernstein *et al*., 1977). To select the most representative structure for experiments, many factors have been considered/checked, including PDB structure validation, resolution, secondary structure match, and whether chains in the structure are complete or not. For covalent docking, the structure was viewed and cleaned using Protein Preparation on Schrodinger Maestro, and for non-covalent docking, it was cleaned using GROMACS commands.

1.7.2 Construction and identification of compounds for targeted compound libraries.

Ligands containing acrylonitrile functional groups for “warheads" were constructed and cleaned using RDkit libraries (Csizmadia, 1999; Landrum, 2013). The Zinc Database (Irwin and Shoichet, 2005) was downloaded locally, and a structure search for the acrylonitrile “warhead” was performed using RDKit libraries using a custom Python script, and hits were managed within Jupyter Notebooks (Perkel, 2018). The initial 403804 ZINC compounds with acrylonitrile functional groups were filtered down using a combination of RDKit library quantitative estimation of drug-likeness (QED), further molecular properties, and ligand efficiency (LE) to construct compound libraries containing more viable compounds for covalent and non-covalent inhibition of the main protease. The binding energy per atom of a ligand to its binding partner, such as a receptor or enzyme, is measured as "ligand efficiency” (Kuntz *et al.*, 1999). Hopkins et al. made an extension to Kuntz’s *et al.* concept, and they defined ligand efficiency (LE) numerically as the quotient of ΔG and the number of non-hydrogen atoms of the compound:

$$LE = \Delta g = (\Delta G) / N$$

where ΔG = -RTlnKi and N is the number of non-hydrogen atoms. A ratio of two related efficiency indices has been used, i.e., surface efficiency index (SEI)/ binding efficiency index (BEI).
BEI is the binding efficiency index, which correlates potency to molecular weight on a per kDa scale, and SEI is the surface efficiency index, which tracks potency gains concerning changes in polar surface area (PSA), referred to as 100 $Å^2$ (Abad-Zapatero, 2007).

1.7.3 Molecular Docking

The filtered subset of acrylonitrile-containing ZINC compounds were docked both non-covalently and covalently to the 6XHM SARS-CoV-2 main protease.

The covalent docking was performed using Maestro Schrödinger (Release, 2022-2). Schrödinger Knime workflows were used to filter covalent docking results through prime energy (total energy) and Prime MM-GBSA (Prime Molecular Mechanics-Generalized Born Surface Area) total energy, with RDKit and Maestro used to visualize the most viable ligand-protein complexes (Li *et al*., 2011).
Non-covalent docking was performed using Autodock Vina, controlled by batch system commands on a high-performance computing cluster for high throughput screening.

1.7.4 Molecular Dynamics

For molecular dynamics on covalently bound ligands, systems were prepared from the System Builder panel, and the parameters used were as follows; the solvent model was predefined SPC, boundary conditions were within the orthorhombic box shape, and the box size calculation method was Buffer with distances of 10.0 and angles of 90.0. Further, the force field used was OPLS4.
The systems were solvated and neutralized within Schrodinger Maestro (Release, 2017), with job scripts for dynamics created by Maestro. These job scripts were submitted to a high-performance computing cluster for molecular dynamics using Desmond (Desmond, 2021).

Protonation of the 6XHM protease was performed using the H++ server before simulations (Gordon *et al.*, 2005). The amber96 force field was used for the MD simulations, including an apoprotein simulated in an aqueous environment. Antechamber and AmberTools were used for generating AMBER topologies of ligands for use in GROMACS (Lindahl and Hess, 2001; Da Silva *et al.*, 2012; Mack *et al.*, 2010). Non-covalent ligand-protein complex simulations required that the topologies first be generated for the ligand and the complex systems, respectively, to construct topologies and parameters for the covalently bonded complex through combining missing parameters. A 50 ns simulation was performed at a 2 fs time step, with trajectories visualized using Visual Molecular Dynamics (VMD) and XMGRACE, and the stability assessed according to the Root Mean Square Fluctuation (RMSF), Root Mean Square Deviation (RMSD) and radius of gyration (Rg) (Humphrey *et al*., 1996).

CHAPTER TWO: MOLECULAR DOCKING

2.1 INTRODUCTION

Computational and molecular modeling tools are much like experiments that help in understanding the molecular aspects of biological systems. Molecular docking is a computational procedure performed on structure-based rational drug design to identify the correct conformations of small molecule ligands and also to estimate the strength of the protein-ligand interaction, usually between one receptor and one ligand. It is one of the computational-based approaches that have been widely used to discover novel hits for various therapeutic targets. The docking programs and software that are most commonly used include Autodock (Morris *et al*., 2009), Autodock Vina (Trott & Olson, 2010), GOLD (Jones *et al.*, 1997), and FlexX (Rarey *et al*., 1996). The aforementioned docking programs and many other methods that are similar to them focus on the docking between two molecules through non-covalent interactions or using other knowledge-based scoring functions to characterize these non-covalent interactions (Rarey *et al.*, 1996). However, some drugs do not bind non-covalently to the active site, namely covalent drugs. The presence of both drugs that bind covalently and non-covalently to the active site is one of the motivations behind this study.

In recent reports, the interface between computational approaches and experiments has been highlighted as an important tool in drug discovery with the increasing interest in the design of covalent inhibitors. In this chapter, covalent and non-covalent docking are used as tools that can help in understanding covalent interactions between acrylonitrile-based inhibitors and the main-protease (Mpro) of SARS-CoV-2, a method that has potential clinical use; because the inhibition of viral protease can decrease the assembly of mature viral particles. In recent reports, many antiviral drugs have been developed against viral infections via targeting proteases. For example, in a study by Lv *et al*., 2015, HIV-1 protease inhibitors (tipranavir, darunavir, amprenavir, lopinavir, saquinavir, atazanavir, indinavir, ritonavir, and nelfinavir) and hepatitis C virus (HCV) NS3/4A protease inhibitors (boceprevir, telaprevir, ritonavir, asunaprevir, paritaprevir, grazoprevir, glecaprevir, voxilaprevir, and sofosbuvir) (de Leuw and Stephan, 2017) were found to be amongst the FDA approved drugs.

Computational chemistry in drug discovery has allowed for quick access to therapeutic agents by making it possible to understand and predict the structural details of chemical interactions (Gschwend *et al.,* 1996). Molecular docking, which is a structure-based virtual screening method as mentioned

before, is used mainly for identifying features that are responsible for specific biological interactions (hit identification) and to predict modifications that can improve potency (lead optimization) when the structure of a target and its active site are available (Kitchen *et al*., 2004). This is possible through the prediction of ligand conformation and pose within a targeted binding site. This is to find the best conformation at which the ligand is bound in a fitted pose as determined by the scoring function. Like other virtual screening methods, molecular docking is a more direct and rational drug discovery approach with lower costs for effective screening (Moitessier *et al.*, 2008; Meng *et al.*, 2011). Along with the availability of the protein structure, the location of the ligand binding site must be identified (Campbell *et al.*, 2003; Laurie and Jackson, 2006). Knowing the location of the binding site, rather than blind docking, increases the docking efficiency (Meng *et al.*, 2011).

2.2 HIGH THROUGHPUT VIRTUAL SCREENING

High-throughput screening (HTS) is an important drug discovery process that allows automated evaluation of large numbers of chemical and/or biological compounds for a specific biological target and to identify hits from compound libraries that may become leads for medical chemistry optimization. High-throughput screening methods are widely used in the pharmaceutical industry, taking advantage of robotics and automation to test the biological or biochemical activity of drugs faster than usual. The main goal of HTS is to identify, through compound library screenings, candidates that affect the target in the desired way, so-called "hits" or “leads”. In this study, the ZINC database was used to identify hits from compound libraries in a virtual screening approach. After the identification of hits, the determination of drug-likeness for the viable compounds is critical, which results in a decrease in the number of viable compounds (Darvas *et al*., 2000). An *in silico* approach to predicting drug-likeness is the use of Lipinski's rule of 5, which tries to predict the oral availability of a compound by the number of hydrogen bond donors, hydrogen acceptors, molecular weight, and its lipophilicity (Lipinski *et al.*, 1997). Methods to define drug-likeness have been developed, with Bickerton et al. (2012) providing a quantitative metric for assessing drug desirability, which they have termed the quantitative estimate of drug-likeness (QED) (Xu and Stevenson, 2000; Ohno *et al.*, 2010).

ZINC is a free database of commercially available compounds that contains over 21 million compounds in ready-to-dock, 3D formats, available at the URL http://zinc.docking.org. Molecules in

ZINC are annotated by molecular properties that include molecular weight, number of rotatable bonds, calculated log P, number of hydrogen-bond donors, hydrogen-bond acceptors, chiral centers, chiral double bonds (E/Z isomerism), polar and apolar desolvation energy (in kcal/mol), net charge, and rigid fragments. The database contains 'lead-like' molecules, having a molecular weight in the range 150 to 350 with calculated log $P < 4$, number of hydrogen bond donors ≤ 3, and number of hydrogen-bond acceptors ≤ 6. ZINC provides several search criteria, such as molecular property constraints, ZINC codes, vendor-based, and molecular substructure searches. ZINC uses a modified filter_light.txt parameter file to filter out undesirable molecules (Irwin and Shoichet, 2005). This filtering step uses OpenEye's implementation of LogP, which also uses Wang's algorithm as the first step for processing molecules into the database (Wang *et al.*, 1997).

As many different properties that are important for the development of a new drug cannot be evaluated by compound library screening, high-throughput virtual screening processes do not identify drugs; for instance, HTS processes cannot evaluate toxicity and bioavailability. As mentioned above, the main role of high-throughput screening assays is to identify "leads" and provide suggestions for their optimization. The HTS assay results instead provide the starting point for further steps in the drug discovery pipeline, like drug design, and for understanding the interaction or role of a particular biochemical process.

Acrylonitrile, which is a reactive Michael acceptor, was used for the design of ligand databases from the ZINC database just to do an extensive reading into their activity as targeted covalent inhibitors of serine, threonine, and cysteine proteases. The complete ZINC database was downloaded locally in a Jupyter notebook, and a high-throughput screening (HTS) using a Python script using RDKit libraries was executed. The SMILES of acrylonitrile were used for the identification of ligand databases. The Python script was written in such a way that it excluded unwanted ligands with SMILES that do not match with acrylonitrile.

Using the RDKit Chem module, the acrylonitrile-based ligands were all saved in a respective directory and converted to sdf format. All the acrylonitrile-based ligands that were saved were then added to a list named "molecules," and the total number of compounds screened from the ZINC database was 403804, visualized using 'Draw' from the RDKit Chem library.
Molecular properties for all the acrylonitrile-based ligands were retrieved using a Python script utilizing the rdkit.Chem.QED module, and a data frame that contained all the data was created. The

molecules were filtered to get those that are more viable as leads for drug discovery. The molecules were filtered using ligand efficiency because they are known to quantify the molecular properties, particularly size and lipophilicity, of small molecules that are required to gain binding affinity to a drug target (Hopkins *et al.,* 2014).

Although BEI and SEI for each ligand have the same potency according to their definitions (as pKi or equivalent), there is no correlation between the two variables, hence they are seen as separate factors. The definitions of these new variables should be compared to the accepted wisdom in the field, which is encapsulated in Lipinski's "rule of five," since these criteria are more commonly used as "rules-of-thumb" or filters than as a rigorous mathematical framework to optimize the drug discovery process. BEI provides a continuous numerical scale for one of Lipinski's variables (MW) and implicitly (via PSA), and SEI provides an analogous scale for all the others related to solubility and favorable PK. As mentioned before, efficiency indices only give three crucial variables (potency, MW, and PSA) a numerical framework by combining them into two (BEI and SEI) and providing comparable and continuous numerical scales for ranking, contrasting, and optimizing their values in a straightforward, two-dimensional "optimization" plane (Abad-Zapatero and Metz, 2005).

The molecules that had a ratio of BEI/SEI between 0.755 and 1.3 were chosen as good candidates, and further work could be done on them. The molecules were filtered down further by how they bind with cysteine.

Substructure searching using SMILES enabled us to find the acrylonitrile (subgraph) in the ZINC molecules (graph) as a full graph theoretic search rather than simple string matching in SMILES. This returned 403804 hits overall. The positive hits were output into separate directories, converted to SDF, and read using Python. The visualization of these systems was done through the RDKit Draw.MolsToGridImage procedure. To retrieve molecular properties for all the positive hits, a Python script was used based on the RDKit Chem module.

2.2.1 RESULTS OF ZINC DATABASE SEARCH

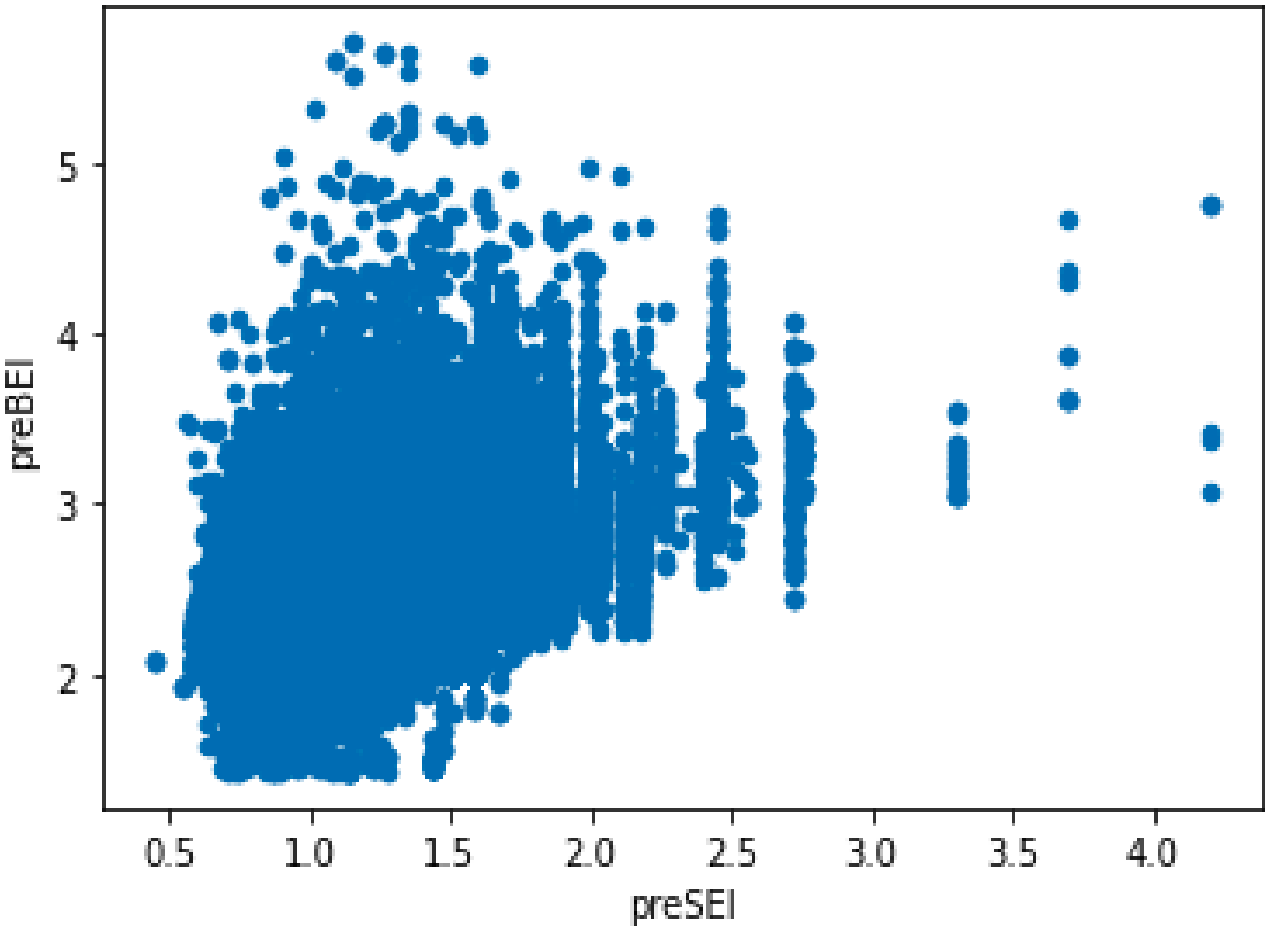


Figure 5: Graphical representation of all 403804 positive hits from the ZINC database.

Given that the molecules with the best properties will have both a large BEI and SEI, and that both of these may be estimated from docking scores, the ratio of BEI/SEI is therefore not dependent on docking scores. Figure 8 shows the plot of “preBEI” vs. preSEI," where both metrics simply further require the docking score to be properly estimated. The 403804 molecules were filtered according to the ratio of efficiency indexes, i.e., BEI/SEI. The 2582 molecules had a ratio that ranged between 0.755 and 1.3, and they were considered good compounds and selected for further work. As mentioned above, efficiency indices give only three crucial variables (potency, MW, and PSA) a numerical framework by combining them into two (BEI and SEI) and providing comparable and continuous numerical scales for ranking, contrasting, and optimizing their values in a straightforward, two-dimensional "optimization" plane. Figure 7 shows the same preBEI/preSEI plot but for a subset of filtered molecules. It is clear in this figure that the ratio is close to 1 for these compounds, which is the most desired place in terms of this metric.

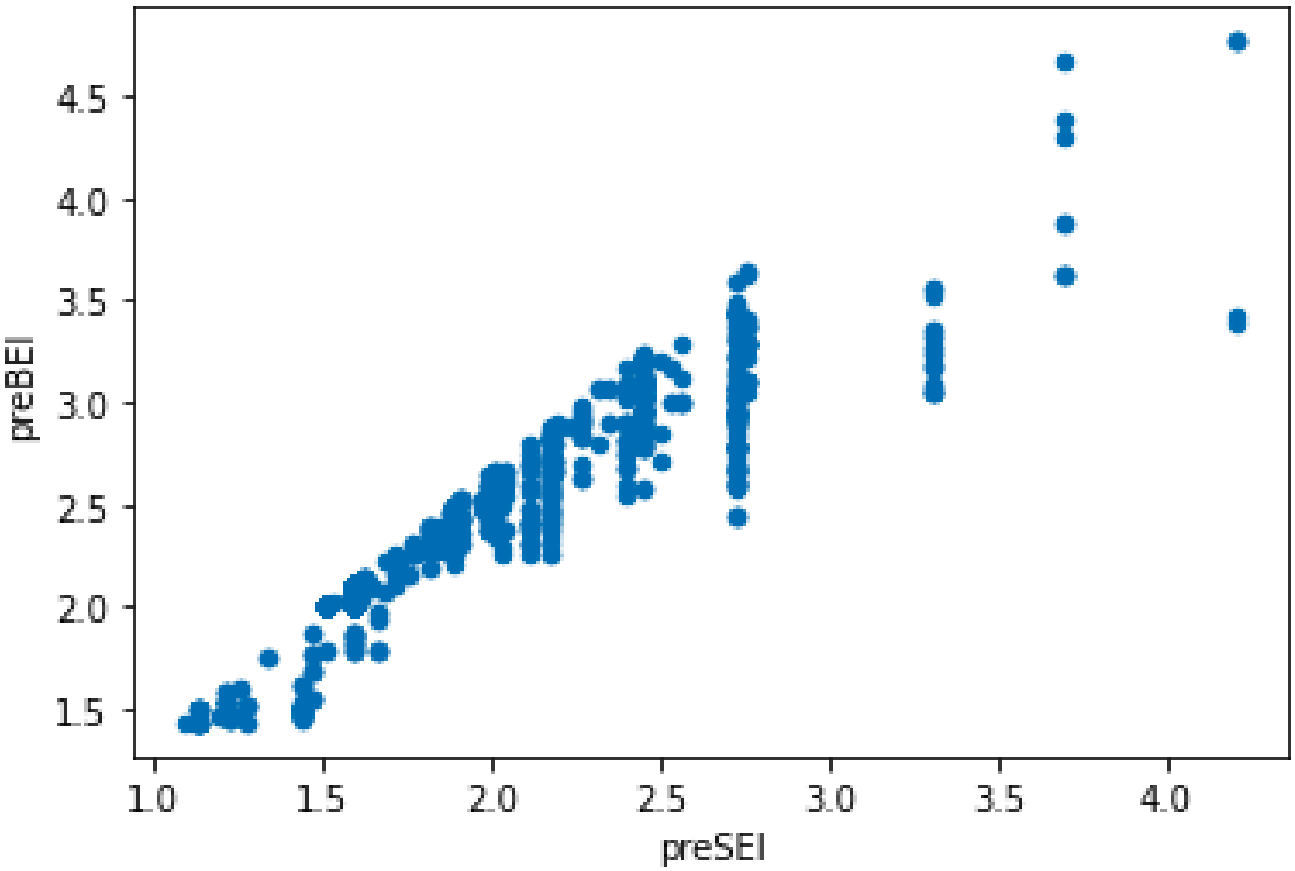


Figure 6: 2582 filtered molecules from the 403804 molecules in the ZINC database.

The compounds' SEI increases as they move to the right in the optimization plane, which causes polarity to decrease, improving the compound's drug-like properties and increasing BEI. Effective selection of the appropriate chemical groups would probably increase the likelihood of a successful clinical candidate while minimizing attrition. It has been discussed how important very high BEI values are in identifying "nuisance" substances or irreversible inhibitors by using the protein tyrosine phosphatase 1B (PTP1B) (Abad-Zapatero and Metz, 2005; Abad-Zapatero et al., 2006).

2.2.2 FURTHER FILTERING OF RESULTS

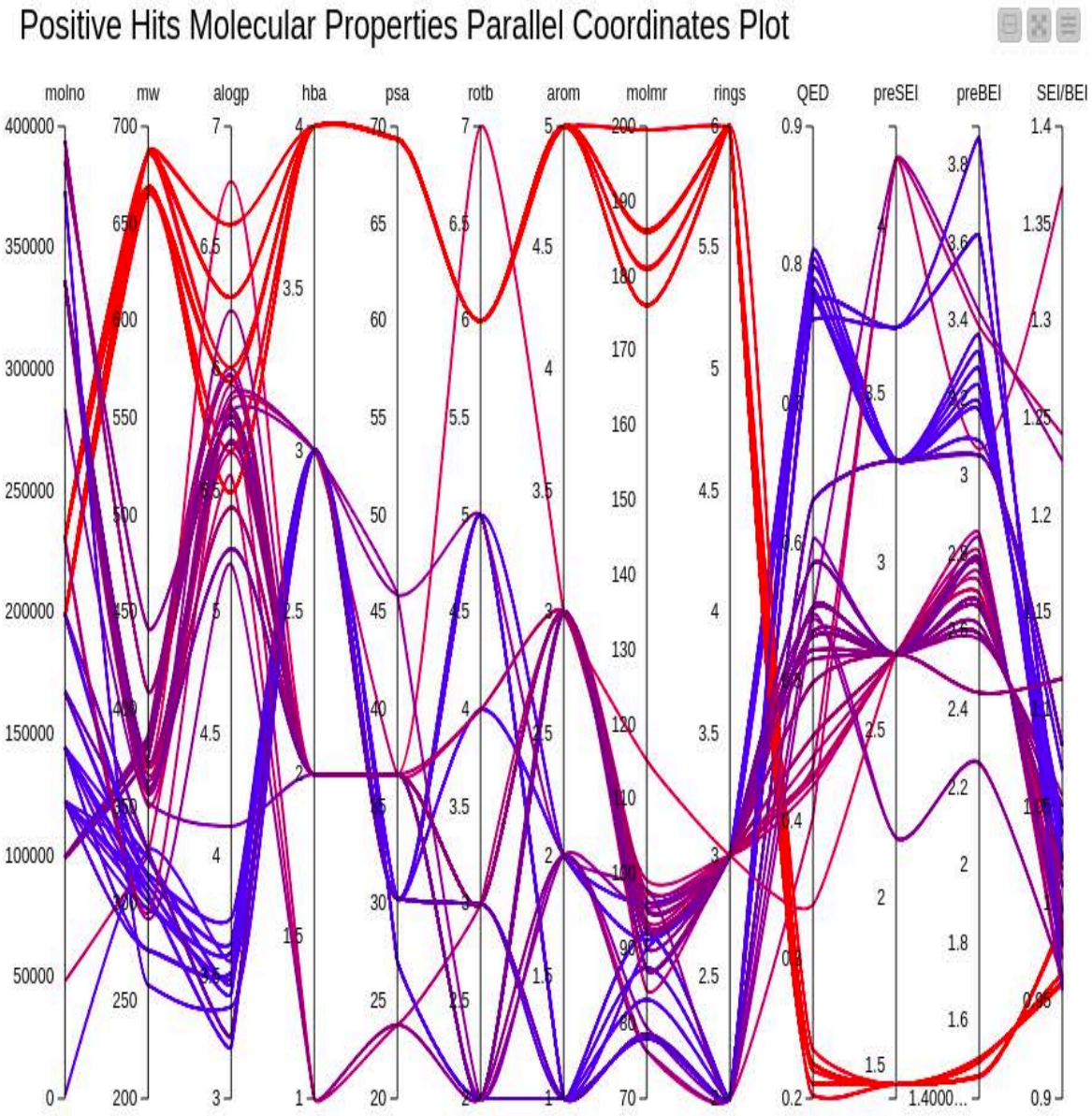


Figure 7: Parallel coordinate plot of filtered positive hits's molecular properties.

Figure 7 shows the results of the calculation of all physicochemical properties, carried through to the final set of hits (2582 molecules).

Further filtering was performed by increasing the BEI:SEI ratio filter to 0.95 from 0.755 because relatively small changes (~1) in SEI or BEI reflect considerable changes in the fingerprint of the compounds they represent in the optimization plane, analogous to the Richter scale used to express the magnitude of seismic occurrences. As stated, the ratio BEI:SEI is proportional to PSA/MW and independent of the pKi values (Abad-Zapatero, 2007).

2.3 INTERACTIONS IN BIOLOGICAL SYSTEMS

Recent studies have shown a growing interest in the design of drugs that form a covalent bond with the target proteins, with almost 30% of drugs that are on the market targeting enzymes that are known to act by covalent inhibition (Robertson, 2005; Robertson, 2007). Covalent inhibition is defined as a mechanism in which compounds that are small inactive their targeted protein receptors reversibly or irreversibly. Firstly, an inhibitor forms a reversible association with the target protein, bringing the inhibitor's chemical warhead into proximity with a specified reactive amino acid residue of the enzyme. Then, a covalent link between the enzyme's reactive component and the inhibitor is formed (Mah *et al.,* 2014; Ray & Murkin, 2019).

The covalent drugs have a much stronger binding strength with the targets because of the covalent linkage formed between Ligand, which is electrophilic, and the nucleophilic target, and that is the reason for the stronger potency while maintaining a small molecule size, which is pharmaceutically favored. The benefit of covalent interactions with the target protein is the prolonged duration of the biological effect. Nevertheless, covalent inhibitors are associated with toxicity because it is not easy to dissociate if off-target binding happens, which is why highly specified selectivity profiles of the covalent drugs are needed. An example of a covalent drug is aspirin, which was initially put on the market over a century ago. Aspirin covalently changes cyclooxygenase (which is an enzyme that is responsible for the formation of prostanoids) by inducing the acetylation of a serine residue that is located right in the active site (Chen and Marnett, 1989; Lecomte *et al*., 1994; Roth *et al.*, 1975; Wells and Marnett, 1992).

2.4 MOLECULAR DOCKING

A computational understanding of covalent and non-covalent docking is increasingly essential to comprehending how covalent inhibitors might be employed to address selectivity and potency issues as a result of the recent boom in drug discovery. Many different approaches have been used to perform both covalent and non-covalent docking of the inhibitors to the target protein, although covalent docking programs in many cases only predict the binding energy between a nucleophilic receptor and an electrophilic ligand. Programs such as Autodock, Autodock/Vina, Moldock, Gold, Glide, and Schrödinger's CovDock perform molecular docking investigations, in addition to having functions that predict the binding energies and ranking the docked compounds based on the binding affinity of the ligand-receptor complex (Goodsell *et al.*, 1996; Trott and Olson, 2010; Thomsen and Christensen, 2006; Verdonk *et al*., 2003; Zhu *et al*., 2014; Huang and Zou, 2010). AutoDock Vina improves the efficiency and accuracy of docking when compared to Autodock while still performing the same ligand-protein binding posing searches. Autodock Tools is available for preparing files, choosing the search area, and viewing results, but the manual selection of atom types for grid maps and selection of search parameters required in Autodock is no longer necessary as Vina calculates its grid map (Ferreira *et al.*, 2015).

In covalent binding, the ligand initially binds through non-covalent interactions with the receptor in an optimal pose and then reacts to form the covalent bond of the ligand-protein complex (Kumalo *et al.*, 2015).

Covalent inhibitors have some unique advantages. For example, covalent warheads can target a rare residue of a particular target protein that is non-conserved and, as a result, lead to the development of highly selective inhibitors. Further, covalent inhibitors can be effective in targeting proteins with a shallow binding cleavage, which will lead to the development of novel inhibitors with increased potency than non-covalent inhibitors (Smith *et al.*, 2009). Zhang *et al.* conducted a study reporting covalent docking using GOLD version 4.0 together with molecular dynamics (MD) simulations (Amber Molecular Dynamics Package version 8.0) to explore the binding mode of peptide aldehyde inhibitors as anti-tumor drugs. The results from this study contributed to the understanding of the mechanism and structure-activity relationship of the peptide aldehyde inhibitors; this may provide useful information for rational drug design (Zhang *et al*., 2009).

The discovery of covalent agents that can cause irreversible and full inhibition of drug targets has resulted in an advancement in the field of covalent targeting, shifting away from non-covalent therapeutics (Singh *et al*., 2011; Bauer, 2015). This allowed researchers a better understanding of covalent targeting's inhibitory mechanisms (Mar *et al*., 2014; De Vita, 2021). Covalent docking has been used in a variety of drug discovery strategies and has proven to be a useful tool for simulating covalent interactions between inhibitors and their biological targets, although there is always room for improvement. To address these issues, it is necessary to revisit several key elements like speed, ligand sampling, accuracy, and protein flexibility and develop new, superior algorithms.

In docking simulations, the goal is to find the best receptor-bound ligand pose using a scoring function best defined by having the lowest binding conformation in this study (Heberlé and de Azevedo, 2011). The energy scoring function is important for evaluating predicted ligand conformations, particularly for the differentiation of correct poses or binders from inactive compounds, which is a crucial aspect of molecular docking (Kitchen *et al.*, 2004; Meng *et al*., 2011). Ligand conformations of low energies are considered to be suitable binding modes that represent a favorable interaction. Free-energy simulation techniques quantitatively model protein-ligand interactions and predict binding affinities, which is impractical for large numbers of ligand-protein complexes (Kollman, 1993; Simonson, 2002; Kitchen *et al.*, 2004). Scoring functions, in contrast, estimate rather than calculate binding affinities, along with adopting assumptions and simplifications without accounting for various physical phenomena that determine molecular recognition (Kitchen *et al*., 2004; Meng *et al.*, 2011). Scoring functions can be divided into force field-based, empirical, and knowledge-based functions (Kitchen *et al*., 2004). Docking programs such as Autodock and Glide search for the best ligand-protein complex (Goodsell *et al.*, 1996; Friesner *et al.*, 2004).

2.4.1 NON-COVALENT AND COVALENT DOCKING SIMULATIONS

The acrylonitrile-containing ZINC compounds were docked both non-covalently and covalently to the 6XHM SARS-CoV-2 main protease. For covalent docking, the ligands in each acrylonitrile-based compound library were imported into Maestro and prepared using LigPrep, and the receptor was prepared using the Protein Wizard (Sastry *et al*., 2013; Tools, 2020). The reactive residue was set to CYS 145, the center of the box set to CYS 145 was selected visually, and the box size was

approximated for all ligands based on the assumption that they were of similar size. A Michael addition was selected for the reaction type, with the core, constraints, and torsional constraints left unaltered. The SMARTS representation for the Michael addition used for covalent docking is shown in Figure 8. The covalent docking simulations were run at the CHPC by saving the ligands in 'batches' (rows) of.maegz files to collate dockings into manageable timeframes within the context of the scheduler.

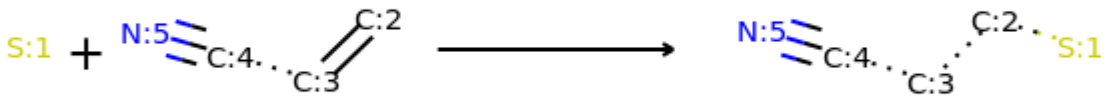


Figure 8: SMARTS representation for the Michael addition used for covalent docking.

For non-covalent docking, the preparation of ligands and the receptor was performed using AutoDock MGLtools, in particular the prepare_receptor4.py for the receptor and prepare_ligand4.py scripts. The prepare_receptor4.py command removes non-polar H atoms and adds polar H atoms where they should exist; it also removes water molecules and ligands. The prepare_ligand4.py command prepares the ligand in terms of providing the torsion tree. These ligands may be preprocessed in terms of conformational searching or optimization using Gaussian. The next step was the preparation of Vina configuration files for each ligand through the use of a Python script. The Vina configuration files provide the Autodock Vina program with all the information it requires for the docking procedure. The output files were then split with the use of a Python script in terms of best to worst docking poses using vina-split. The files with the best docking poses are then used for further work, which is molecular dynamics for this study.

The 2582 molecules were taken forward to both covalent and non-covalent docking workflows using Glide Covalent Docking and Autodock Vina; respectively, we identified 11 of the best-docked systems, which are described in the following discussion. The choice of these 11 systems was based not only on the favorable energy of non-covalent docking but also on the distance of non-covalent docking to the active site, particularly the CYS145 residue used in covalent docking. The thinking was that favorable

non-covalent docking close to CYS145 was necessary for there to be a possibility of a covalent bond being formed. A KNIME workflow aided this decision. The scatter plot produced just for the 11 best systems is shown in Figure 9 below.

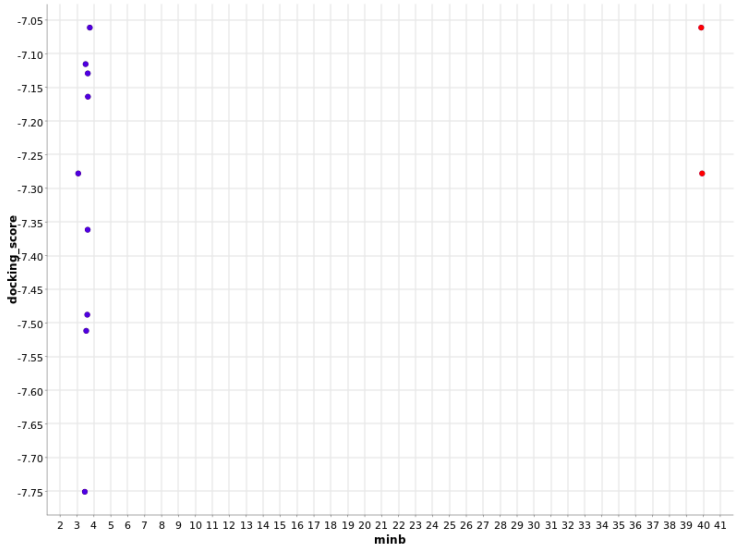


Figure 9: Graph of Vina docking score vs the minimum distance to CYS145 for the docked ligand

In Figure 9, the 11 ligands chosen are shown - the red indicates that the ligand is close to CYS145 for chain B rather than chain A. This graph is focused on the ligands with the best distance/docking score combination, so most of the 2582 ligand points are not visible. Although non-covalent docking drove the decision process, the covalent docking results are important and are described first below.

2.4.1.1 COVALENT DOCKING RESULTS

All of these 11 ligands have a distance of close to 3Å in the non-covalent docking and a vina docking score in the range of -7.05 to -7.75. Since all of these fall within the criteria of interest the covalent docking pose was explored first to identify any salient features of the binding that may raise attention when considering compounds as lead compounds. The non-covalent poses are discussed after this as support for the primary covalent discussion.

The first ligand of interest was ligand 48356. The covalent docked pose is shown in Figure 12. Note that the covalent binding in these diagrams is sometimes mistaken by software to be an unfavorable bump in the two-dimensional plot. However, this confirms the covalent binding to CYS145.

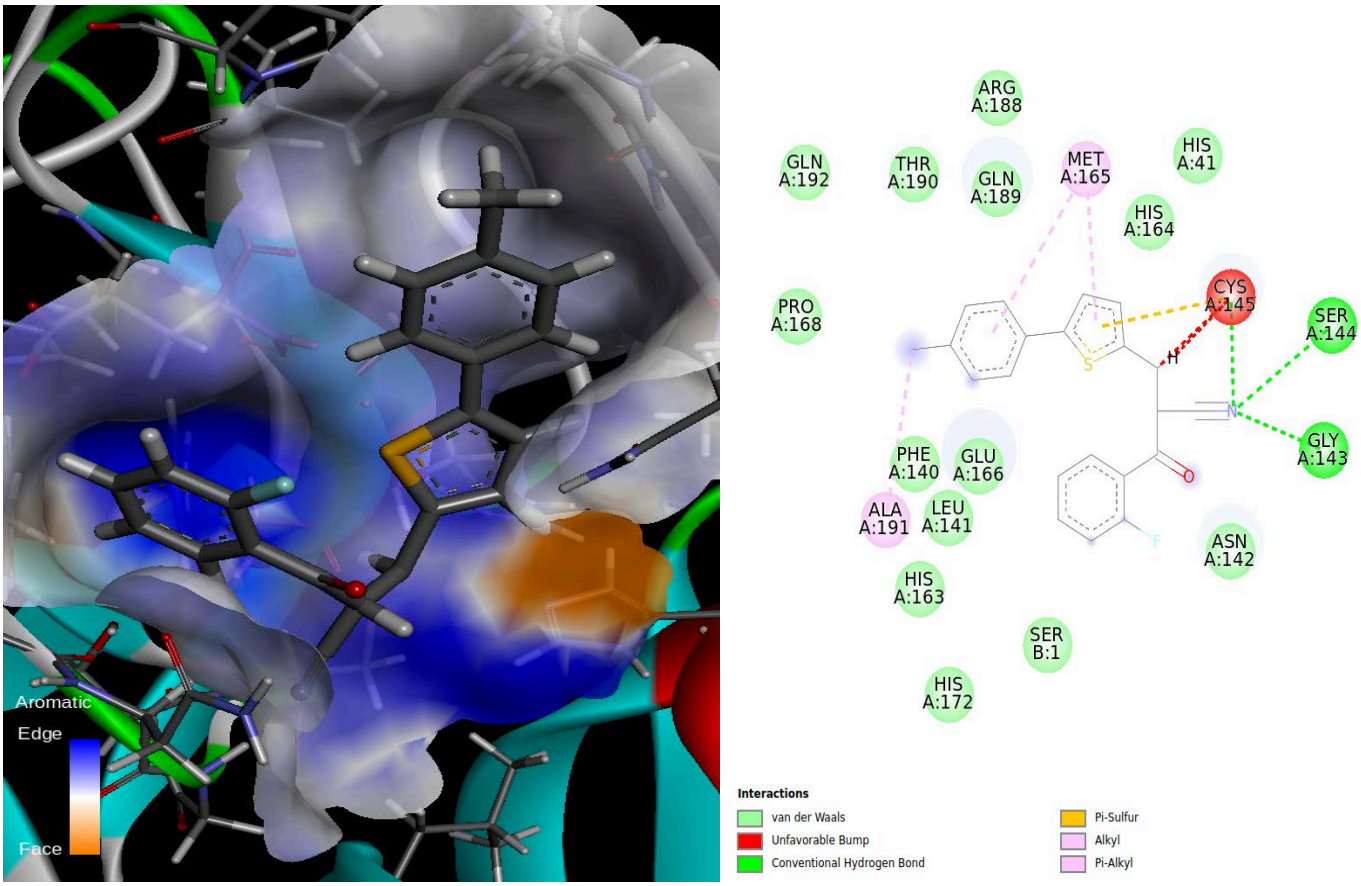


Figure 10: DiscoveryStudio generated a 3D (left) and 2D (right) diagram of Ligand 48356 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional

hydrogen bonding, the red indicating unfavorable bump interactions, the orange indicating Pi-Sulfur, the pink colors showing alkyl and Pi-Alkyl interactions, respectively. The bonds are shown with dots from the circle to the ligand; the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

In Figure 10, for this covalently bound ligand-receptor complex, there are several non-bonded van der Waals interactions, an Alkyl and Pi-alkyl interaction with the ligand, several hydrogen bonds with the catalytic dyad CYS-145 (HIS-41 is the other residue in this dyad), and protein residues SER-144 and GLY-143. In the 2-dimensional representation, there appears to be an unfavorable bump interaction of cysteine with the ligand; however, as mentioned, in this case, the ligand is covalently bound to this cysteine residue, so it is expected to be much closer than in a non-covalently bound case. In cases where this unfavorable bump is observed, but in non-covalent molecular docking, it may be due to an "induced fit" problem that generates these unfavorable interactions. This highlights the necessity of progressing to molecular dynamics to enable the easing of such interactions.

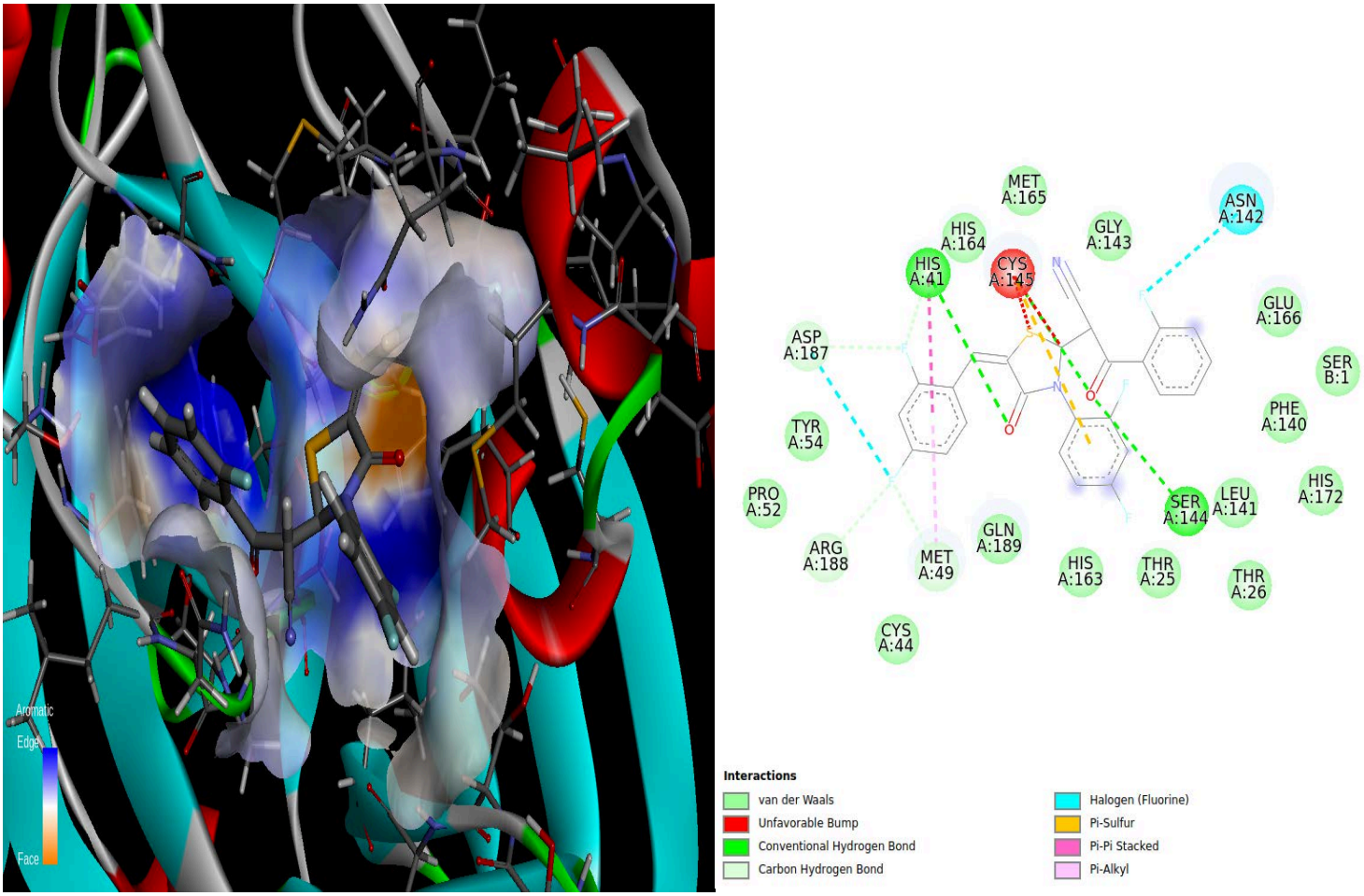


Figure 11: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 117238 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the red indicating unfavorable bump interactions, the light lime showing carbon-hydrogen bond interactions, blue showing halogen interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

In Figure 11, we see again the cysteine 145 binding to the ligand, in this case ligand 117238. This ligand contains two halogens (Fluorine) and it is interesting to see the identification of a halogen bond to ASN-142 and ASP-187. There are several non-bonded van der Waals interactions. The catalytic dyads CYS-145 and HIS-41 appear to have a conventional hydrogen bond with the ligand, which indicates the stability of the complex.

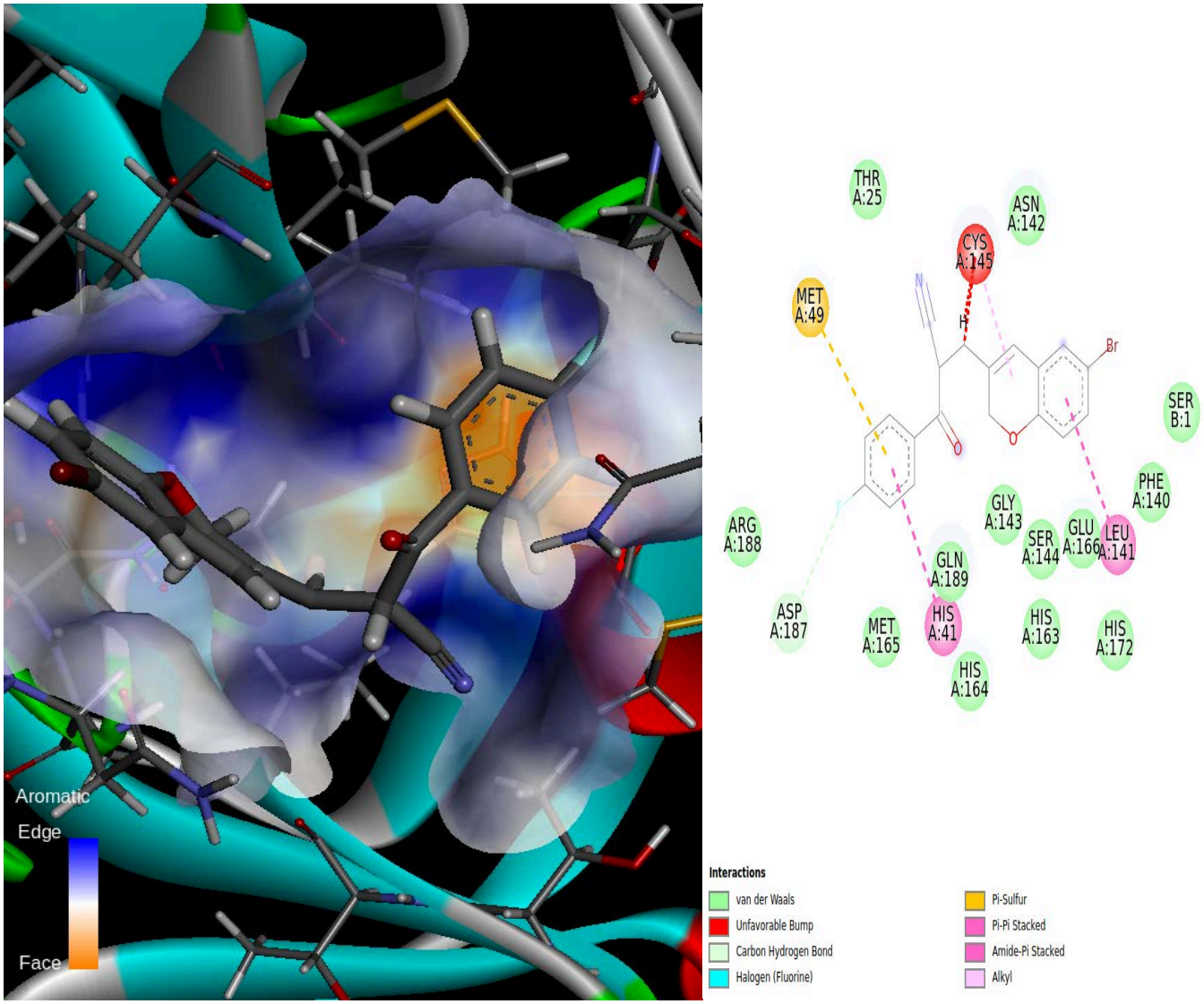


Figure 12: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 337222 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the red indicating unfavorable bump interactions, the light lime showing carbon-hydrogen bond interactions, blue showing halogen (fluorine) interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

In Figure 12, again there seems to be an interaction between the cysteine 145 and the ligand 337222 that appears to be unfavorable. As mentioned above, the ligand is covalently bound to this cysteine residue and so it is expected to be much closer than in a non-covalently bound case. The catalytic dyad CYS-145 and HIS-41 appears to have an alkyl interaction with the ligand together with the residue LEU-141 which is part of an oxyanion loop in the binding pocket that maintains the correct conformation of the pocket, with changes to the loop alerting a possible enzymatic inactivity (Li et al., 2016).

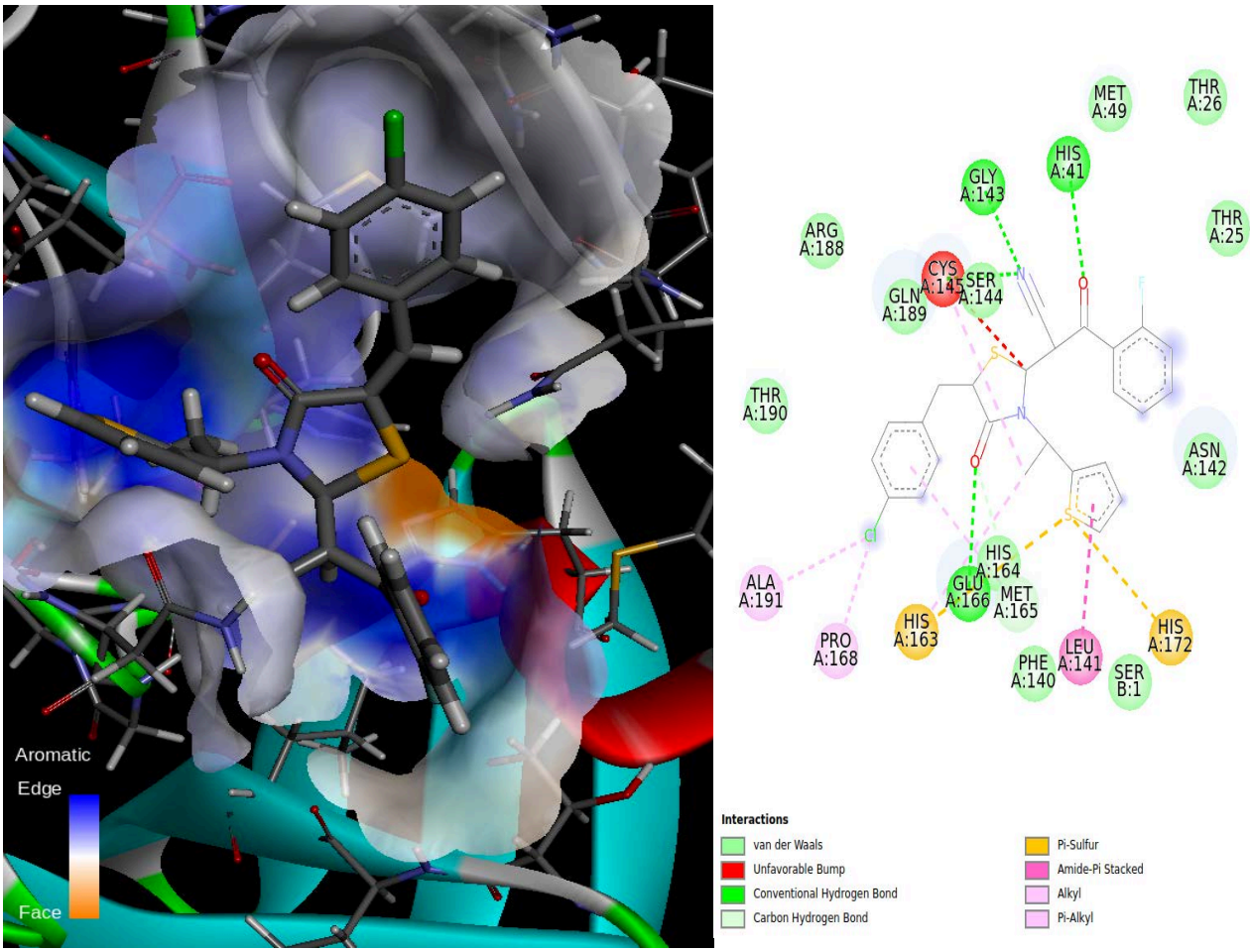


Figure 13: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 387305 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the

2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the red indicating unfavorable bump interactions, the light lime showing carbon-hydrogen bond interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

In Figure 13, we see again the cysteine 145 binding to the ligand, in this case ligand 387305. This bump is a positive thing in covalently bonded receptor-ligand interactions. The residue LEU-141 which as mentioned above is part of an oxyanion loop in the binding pocket that maintains the correct conformation of the pocket, with changes to the loop alerting a possible enzymatic inactivity (Li et al., 2016). The residues HIS-172 and GLU-166 appear to have Pi-Sulfur and conventional hydrogen bonding with the ligand respectively, and these residues are of importance because they are believed to make up the active site's opening for substrates (Yang et al., 2003). The catalytic dyad HIS-41 appears to have a conventional hydrogen interaction with the oxygen of the ligand.

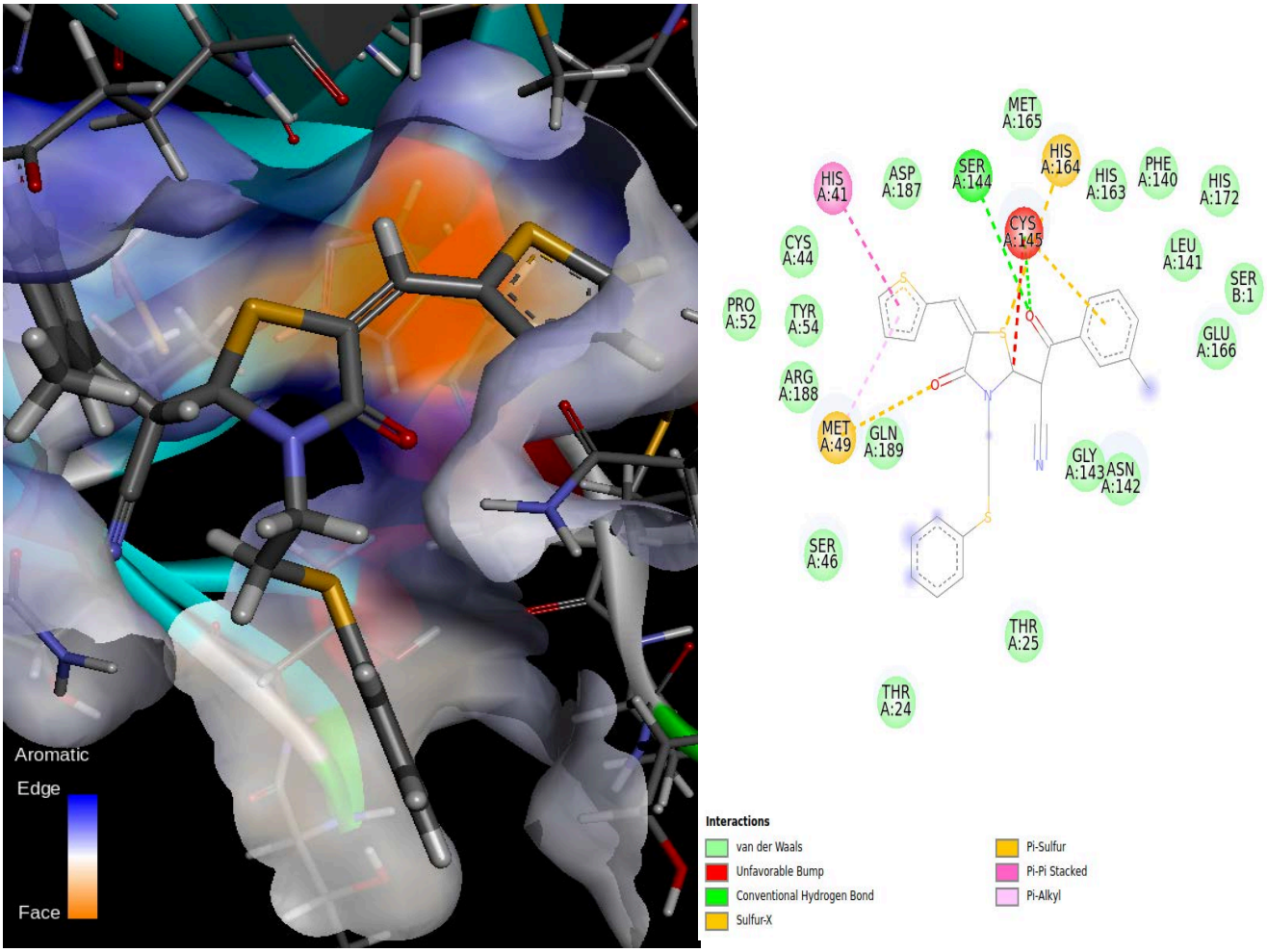


Figure 14: DiscoveryStudio generated 3D (left) & 2D (right) diagrams of Ligand 396939 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the red indicating unfavorable bump interactions, orange indicating Pi-Sulfur, the pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

In Figure 14, the cysteine 145 is again observed to be binding to the ligand, in this case ligand 396939, even here this appears to be an unfavorable interaction but as mentioned before that is not the case. There are several non-bonded van der Waals interactions. The catalytic dyad HIS-41 has an alkyl interaction with the ligand. The residue MET-49 has a Pi-Alkyl interaction and a Pi-Sulfur interaction

with the ligand and this residue is one of the essential residues implicated in substrate binding that is found in the protein's binding pocket, as LEU-141 mentioned above, this MET-49 residue is believed to make up the active site's opening (Yang et al., 2003).

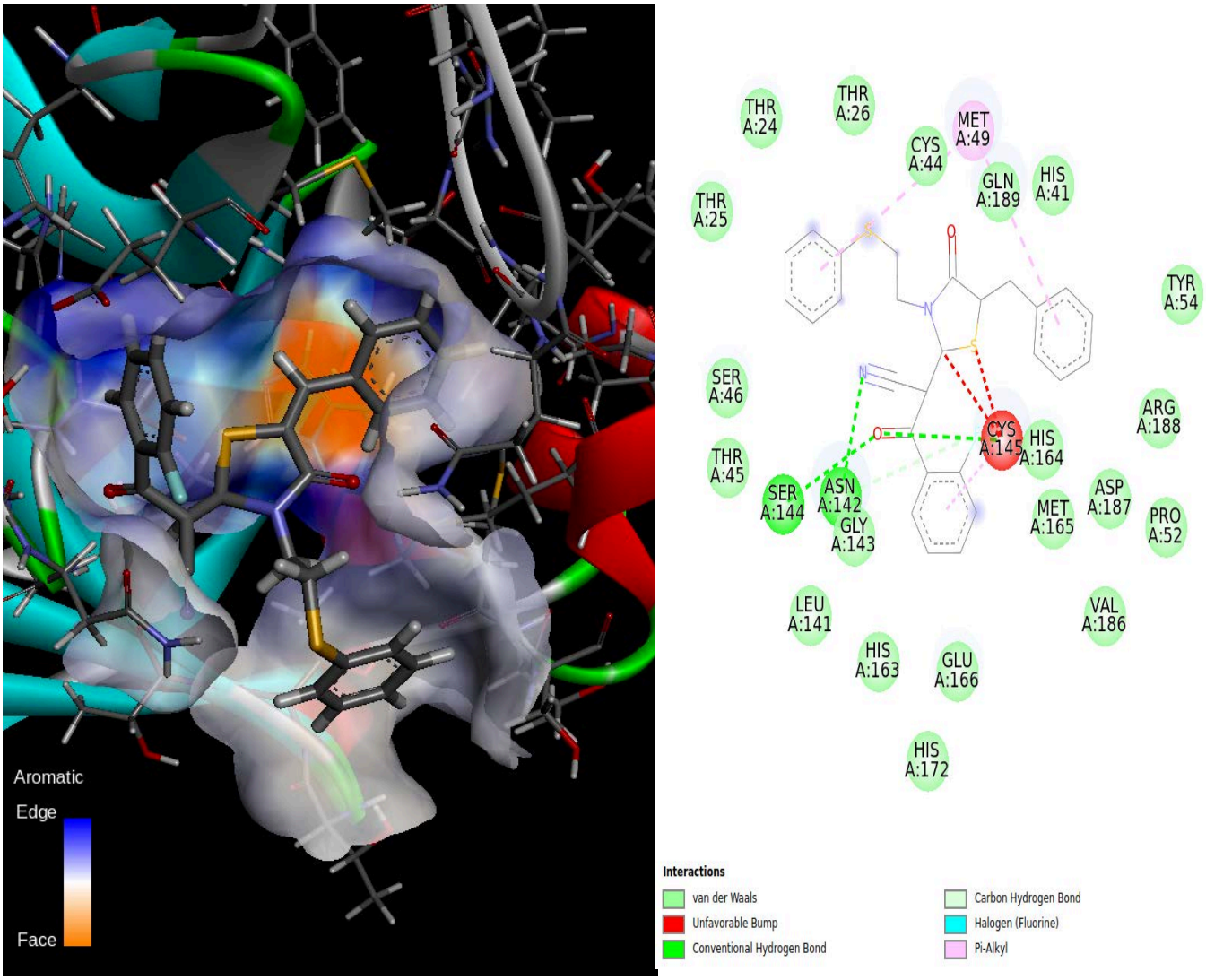


Figure 15: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 397136 covalently docked to the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the red indicating unfavorable bump interactions, orange indicating Pi-Sulfur, the

pink colors showing Alkyl and Pi-Alkyl interactions respectively. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

In Figure 15, there also appear to be several non-bonded van der Waals interactions. The cysteine 145 seems to be bonded with the sulfur atom of the ligand. The essential residues MET-49 and ASN-142 appear to have Pi-Alkyl and conventional hydrogen bond interactions with ligand 397136, respectively.

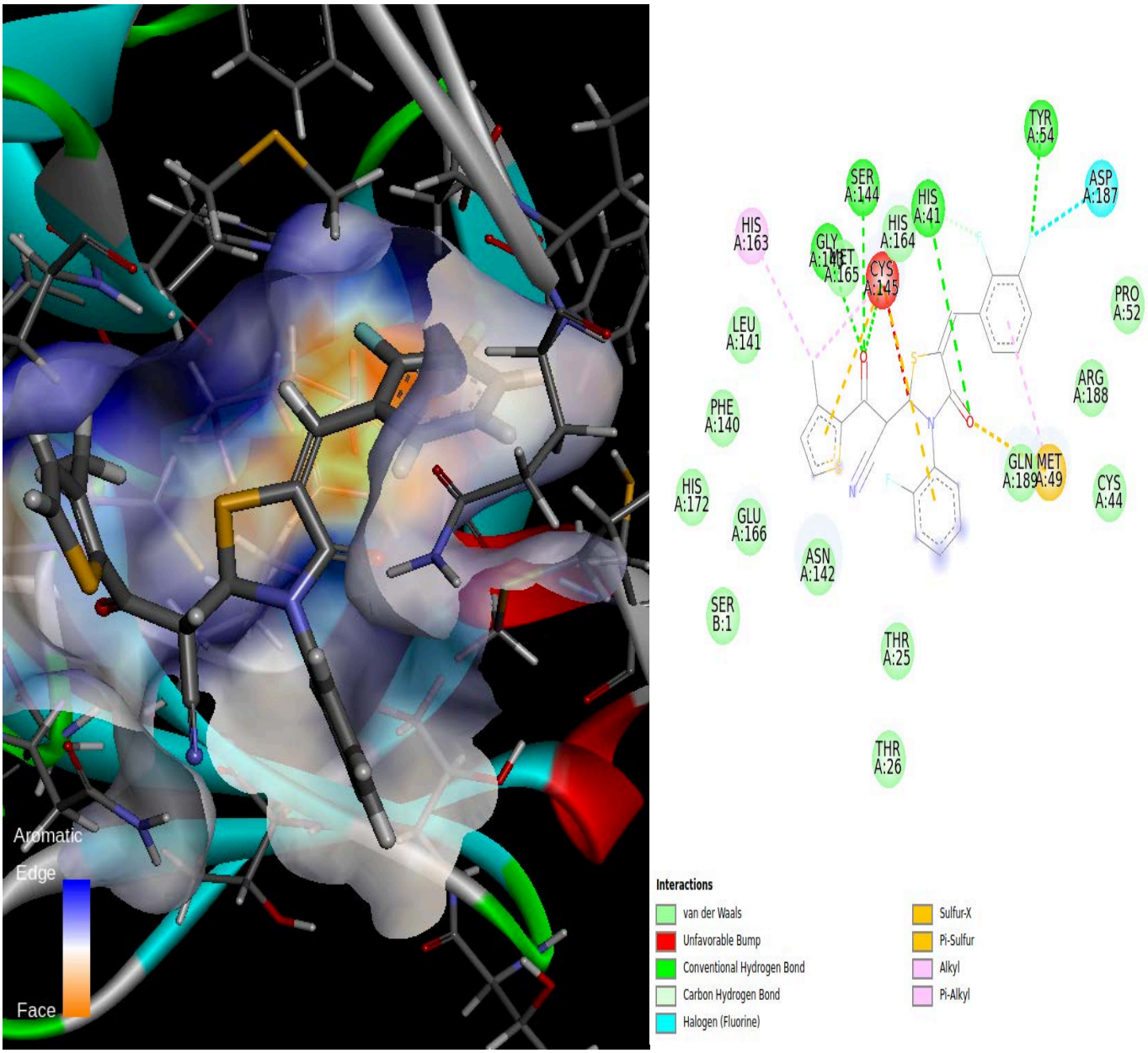


Figure 16: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 397730 covalently docked to the receptor.

In Figure 16, we see again the cysteine 145 binding to the ligand, in this case ligand 397730. The catalytic dyad HIS-41 has a conventional hydrogen bond with the oxygen atom of the ligand. Several non-bonded van de Waals interactions are observed and a halogen interaction with residue with ASP-187, this is the second halogen interaction observed in covalently bound complexes.

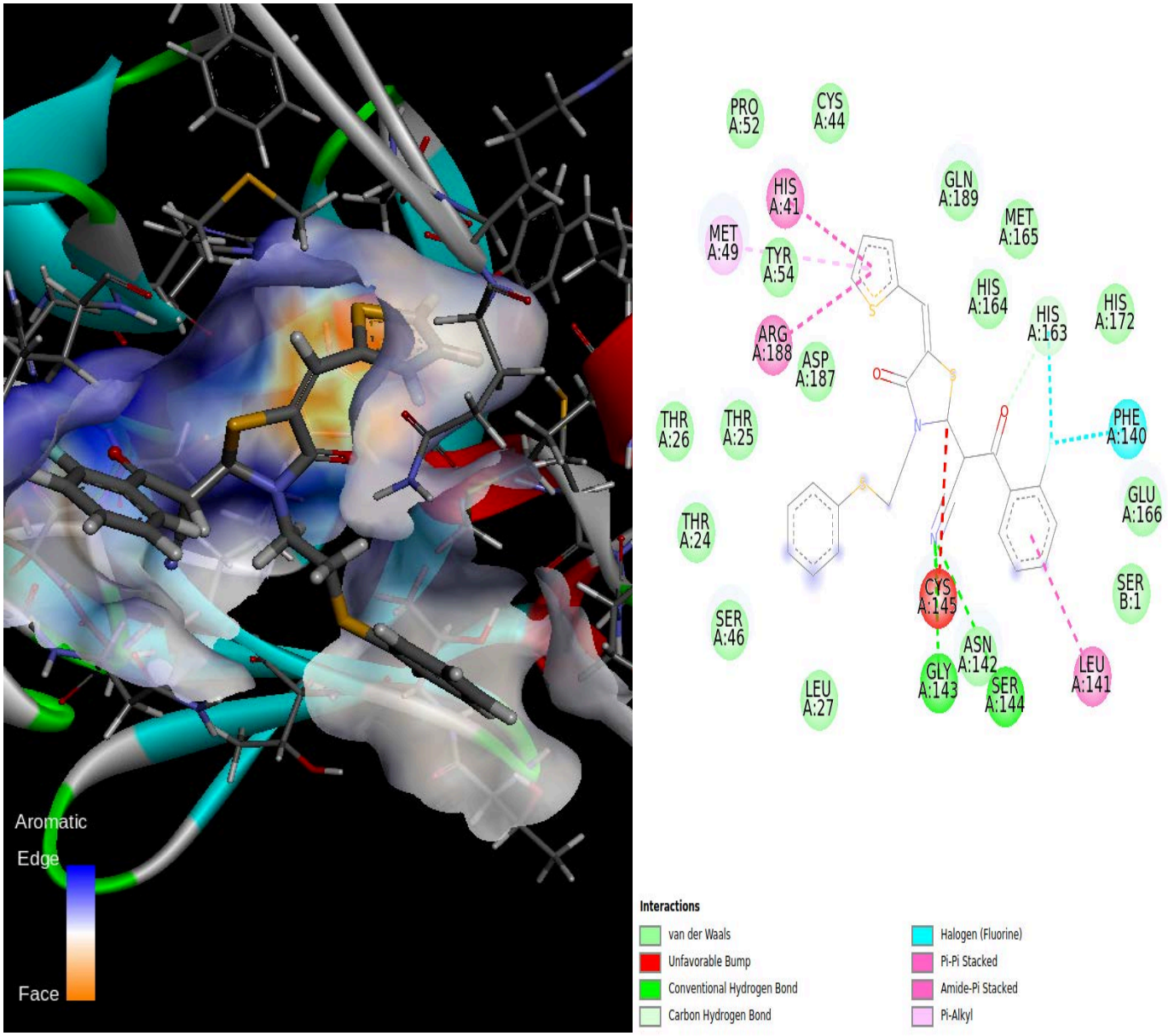


Figure 17: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 402091 covalently docked to the receptor

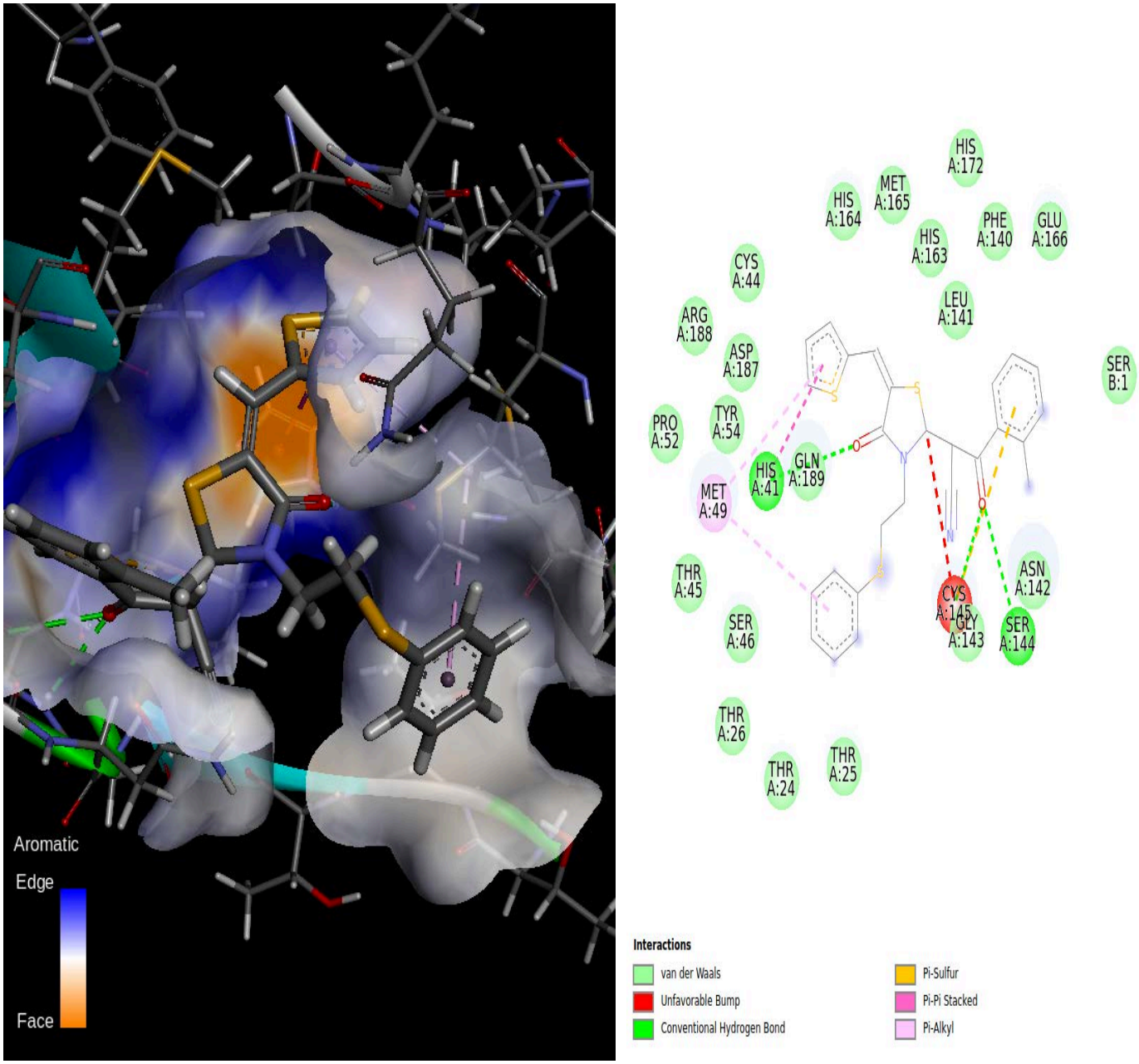


Figure 18: DiscoveryStudio generated 3D (left) & 2D (right) diagram of Ligand 403456 covalently docked to the receptor

In Figures 17 and 18, the cysteine 145 is again observed to be interacting with the ligands 402091 and 403456 respectively. In Figure 19 there appears to be a halogen interaction with residue PHE-140 and

HIS 163. Halogen interactions are only found in three complexes in the covalent molecular docking complexes. In both figures, as with the rest, there appear several non-bonded van der Waals interactions.

2.4.1.1.1 DISCUSSION OF COVALENT DOCKED SYSTEMS.

This set of 11 covalently docked systems provides some features of interest. The docked systems show a range of stabilizing interactions across a broad range of types of interaction. For example ligands 402091, 117238 and 397730 show halogen bonding between the receptor and the ligand, which is crucial to medical chemistry as halogenation of drugs, generally, improves both selectivity and efficacy toward protein active sites. On the other hand, several ligands 402091, 396939, 387305 and 337222 have π-stacking interactions. This range of binding interactions across these systems is expected to infer some resilience among these compounds when exploring mutations in the receptor. Should one compound lose a particular interaction in a mutant protease, it is possible that another will not lose its separate interactions to different residues.

In this section, we felt it important to detail the interactions to inform both decisions on the choice of ligands (should this study proceed to *in vitro* work) but also to compare when molecular dynamics is used to simulate the complex in a buffered solution.

2.4.1.2 NON-COVALENT DOCKING RESULTS

The essential residues implicated in substrate binding are found in 3CLpro's binding pocket. These residues, along with the CYS-145 and HIS-41 catalytic dyad, THR-45, MET-49, PHE-140, ASN-142, ASP-187, ARG-188, GLN-189, MET-165, HIS-172, and GLU-166, are believed to make up the active site's opening for substrates (Yang *et al.*, 2003). The residues SER-139, PHE-140, and LEU-141 are part of an oxyanion loop in the binding pocket that maintains the correct conformation of the pocket, with changes to the loop alerting a possible enzymatic inactivity (Li *et al*., 2016). The stabilization of the binding pocket is reliant on the distance between CYS-145 and HIS-41, where a hydrogen bond must be maintained for stability and another thing that is of importance is the hydrophobic packing of HIS-163 and PHE-140 (Huang et al., 2016; Li *et al*., 2016).

In the non-covalent docking, it was felt important to bear this structure in mind to better understand the interactions between the ligands chosen and the protein. Further, these particular ligands were chosen based on the proximity of the ligand to CYS145; as such the positioning of the ligand is favorable in terms of the subsequent generation of the covalent bond. However additional interactions will further stabilize the ligand within the active site during the bond formation to CYS145. Below we detail the exact binding orientation and interactions for these 11 ligands bound in a non-covalent pose (from Autodock Vina).

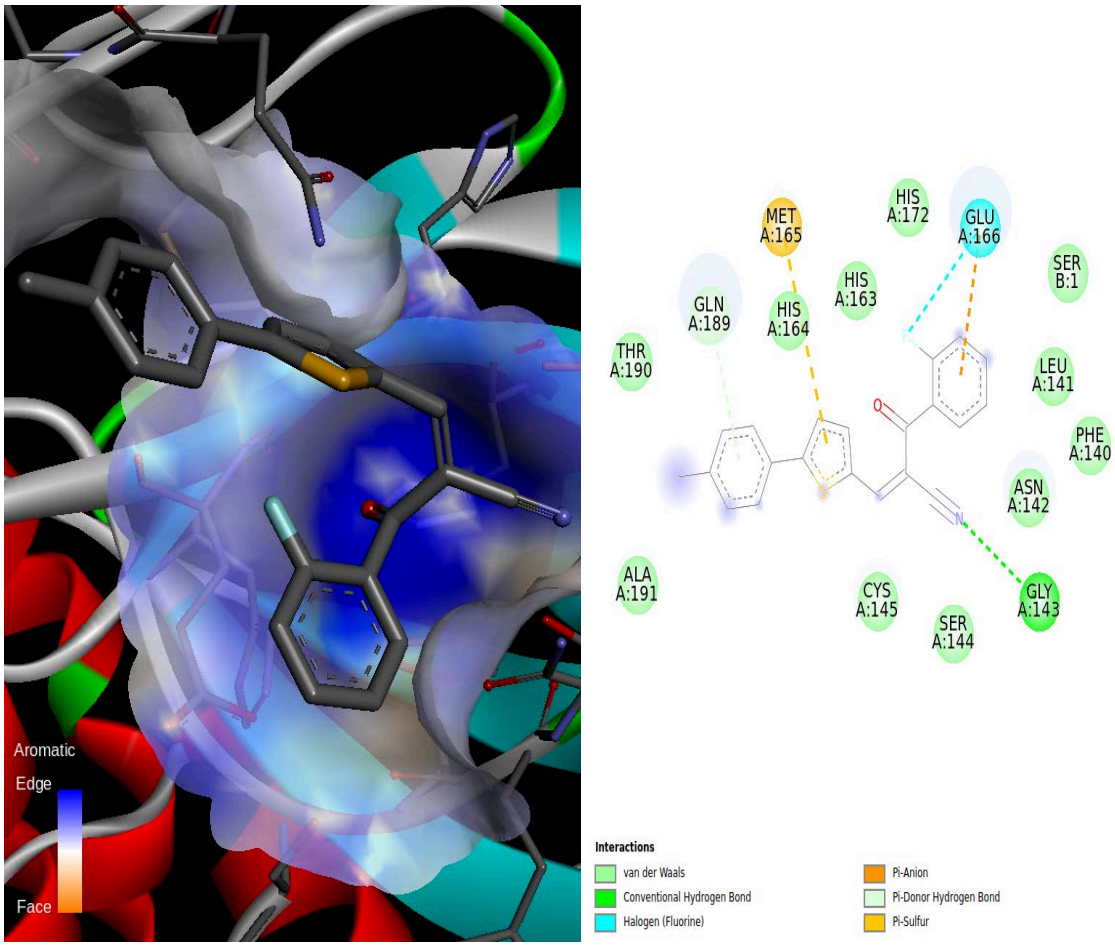


Figure 19: DiscoveryStudio generated 3D (left) & 2D (right) diagram of ligand 48356 non-covalently docked to chain A of the receptor. The protein residues are represented with circles with light green colors (on the 2D diagram on the left) showing van der Waals interactions, the dark green showing conventional hydrogen bonding, the blue showing Halogen interactions, darker orange indicating Pi-Anion, the lime color indicating Pi-Donor Hydrogen Bond and the light orange indicating Pi-Sulfur interactions. The bonds are shown with dots from the circle to the ligand, the non-bonded interactions (atoms interacting neither non-covalently nor through hydrogen bonds) are shown without the dots.

In Figure 19, the ligand-receptor complex for ligand 48356, there are several van de Waals interactions, with protein residues CYS A:145, SER A: 144, ALA A:191, THR A:190, HIS A:163, HIS A:172, SER B:1, LEU A:141, PHE A:140, and ASN A:142. This interaction with CYS145 is expected since the ligands were chosen based on proximity to this residue. There is a Halogen and a Pi-Sulfur interaction of protein residue GLU A:166 and the Fluorine and Sulfur of the ligand. In almost all non-covalently bound ligand-receptor complexes generated from this study, it is observed that there is a conventional hydrogen bonding between the protein residue GLY A:143 and the Nitrogen (N) atom of the ligands. Complexes are stabilized by hydrogen bonding and electrostatic interactions and the formation of hydrogen bonds by GLY A:143 shows that this residue is crucial for keeping ligands in the binding pocket stable (Gohlke and Klebe, 2002).

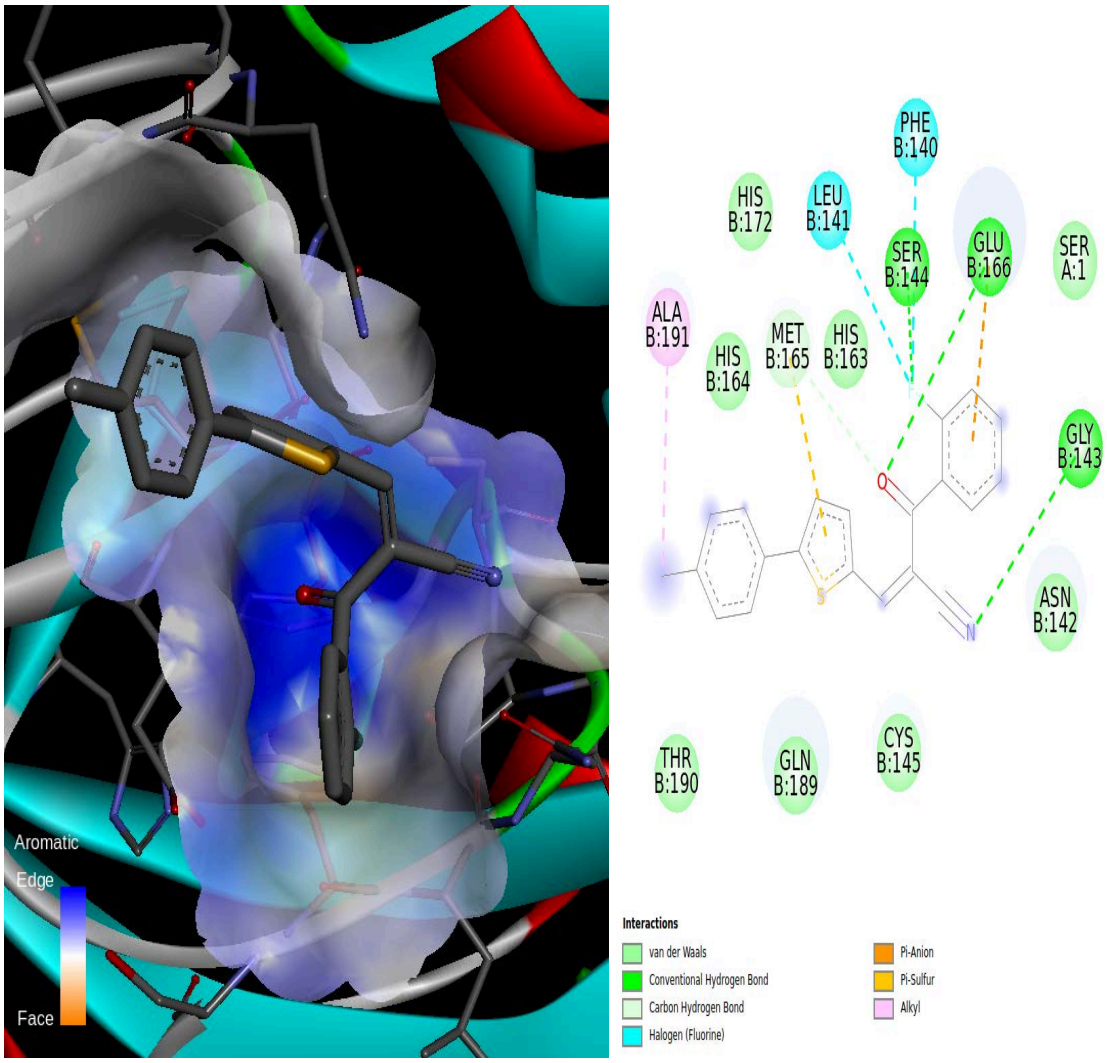


Figure 20: A 3D (left) & 2D (right) diagram of ligand 48356 non-covalently docked to chain B of the receptor

Figure 20 shows the same ligand bound, but now to chain B. As mentioned before, there is a conventional hydrogen bonding between the protein residue GLY B:143 and the Nitrogen (N) atom of the ligands. Complexes are stabilized by hydrogen bonding and electrostatic interactions and the formation of hydrogen bonds by GLY B:143 shows that this residue is crucial for keeping ligands in the binding pocket stable (Gohlke and Klebe, 2002). There are several hydrogen bonds formed in figure 20 ligand-receptor interactions, those being with protein residues SER B:144, GLU B:166, and GLY B:143. This indicates that this is a highly stable complex.

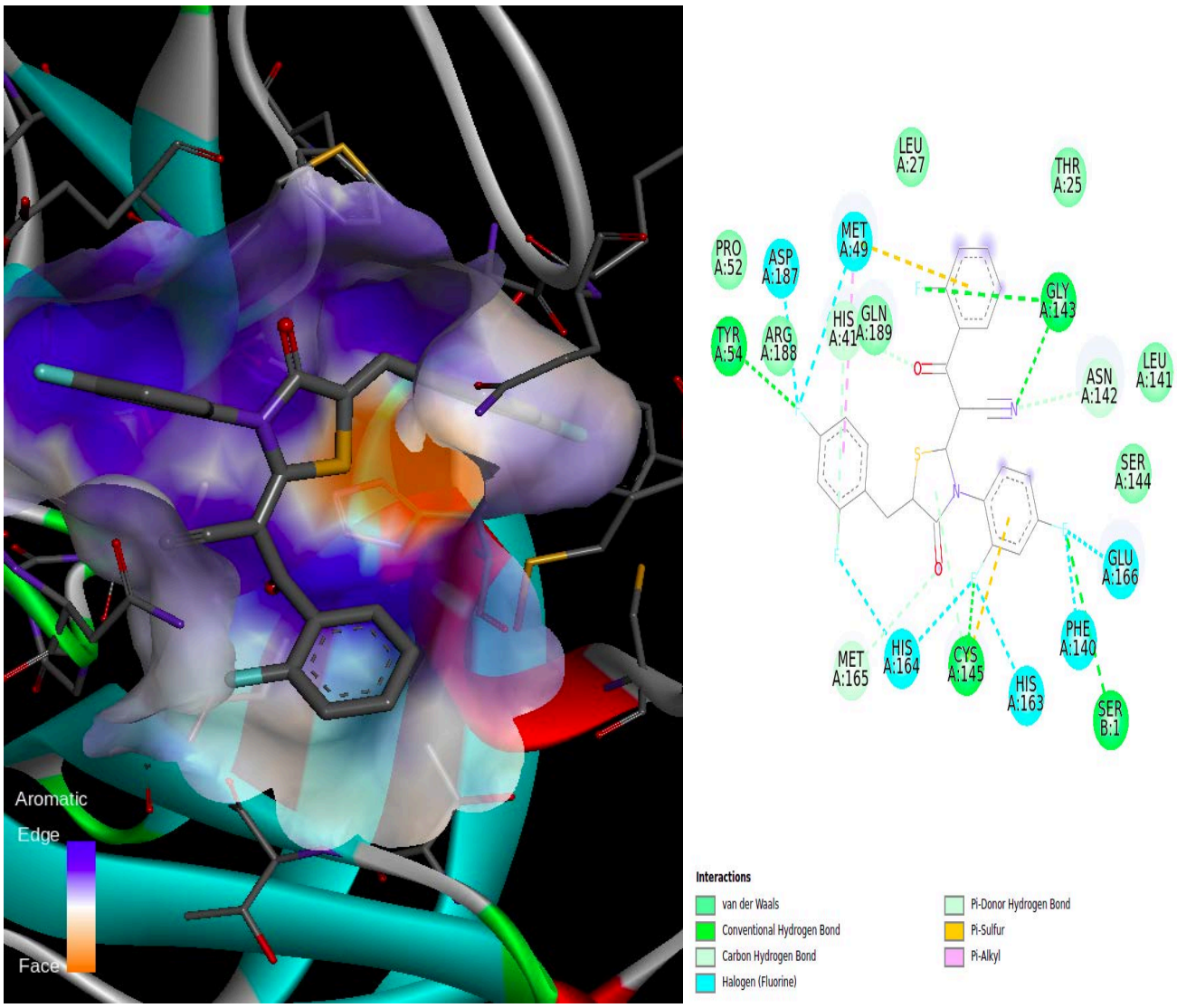


Figure 21: A 3D (left) & 2D (right) diagram of ligand 117238 non-covalently docked to chain A of the receptor.

Turning to ligand 117238, in Figure 21 for this ligand-receptor interaction complex, the catalytic dyad CYS A:145 has a conventional hydrogen bond with the fluorine of the ligand 117238 and a Pi-Sulfur interaction and this is also seen in Figure 22 (binding to chain B). However, in Figure 22 this cysteine 145 has an additional Pi-Alkyl interaction with the ligand. In Figure 23 (Chain A) the HIS A:41 catalytic dyad has a carbon-hydrogen bond interaction with the ligand and a non-bonded interaction in Figure 22. The conventional hydrogen bonding with the catalytic dyad and GLY A:143, TYR A:54, and SER B:1 is evidence of the stability of this complex.

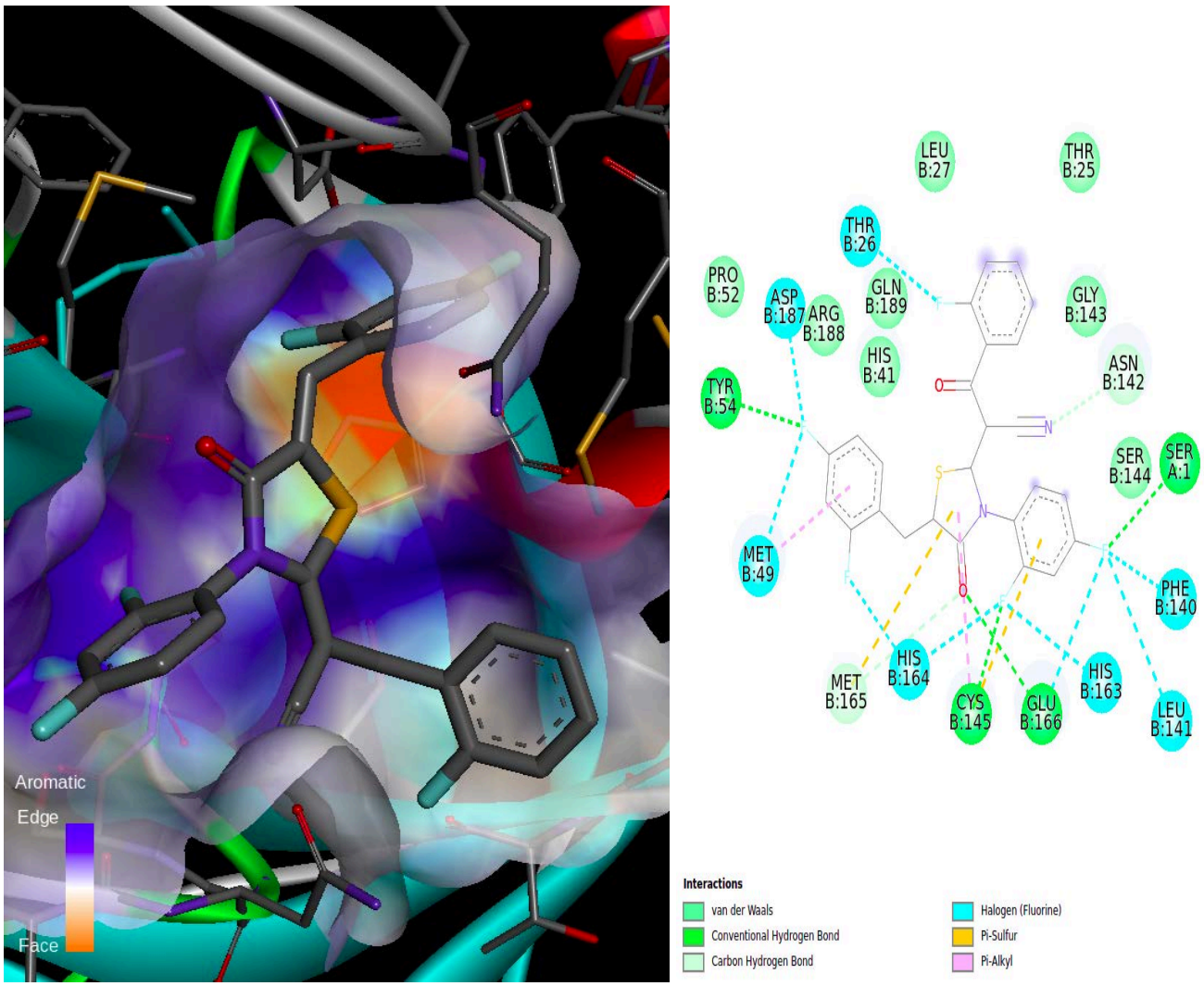


Figure 22: A 3D (left) & 2D (right) diagram of ligand 117238 non-covalently docked to chain B of the receptor.

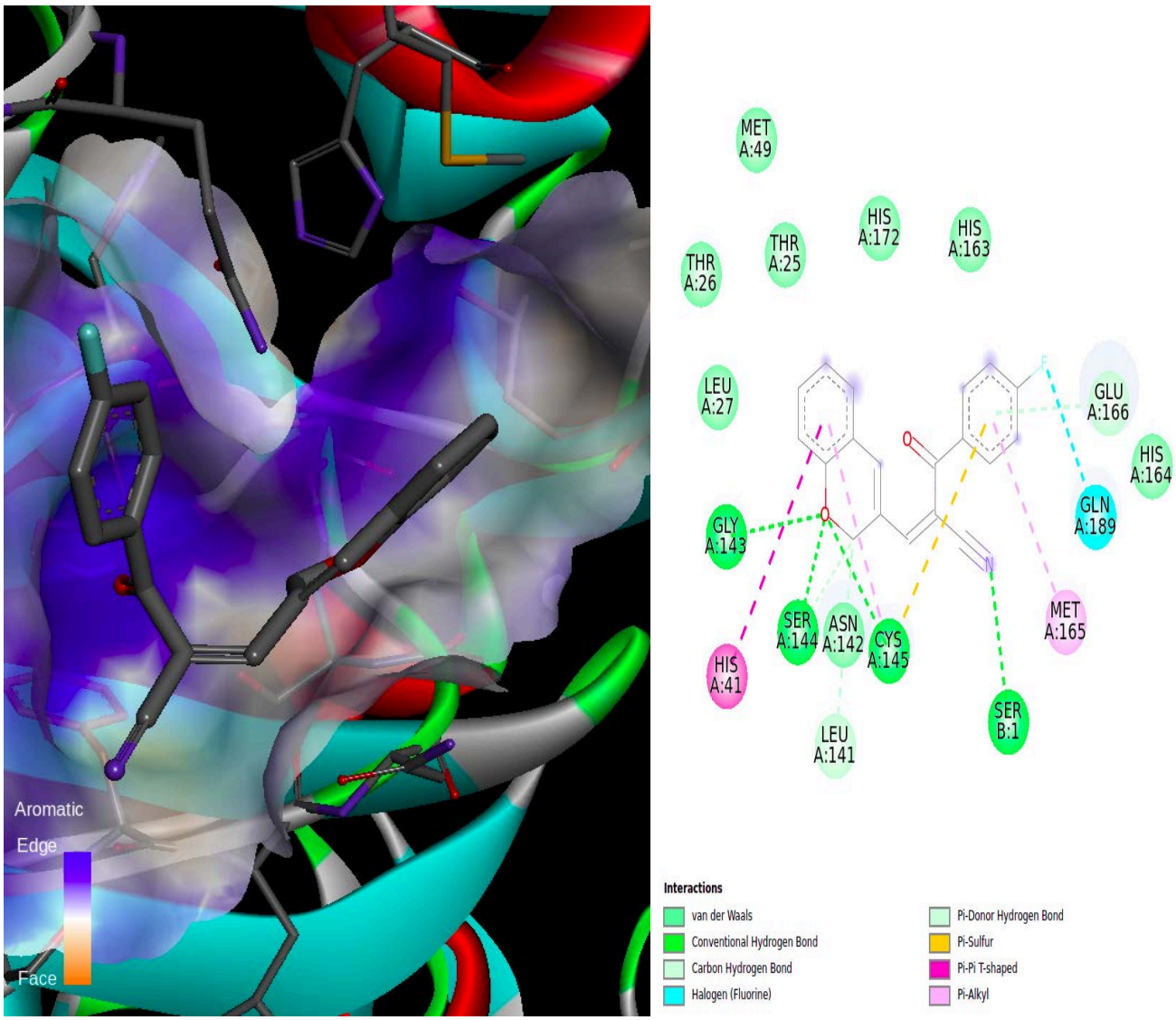


Figure 23: A 3D (left) & 2D (right) diagram of ligand 337222 non-covalently docked to chain A of the receptor.

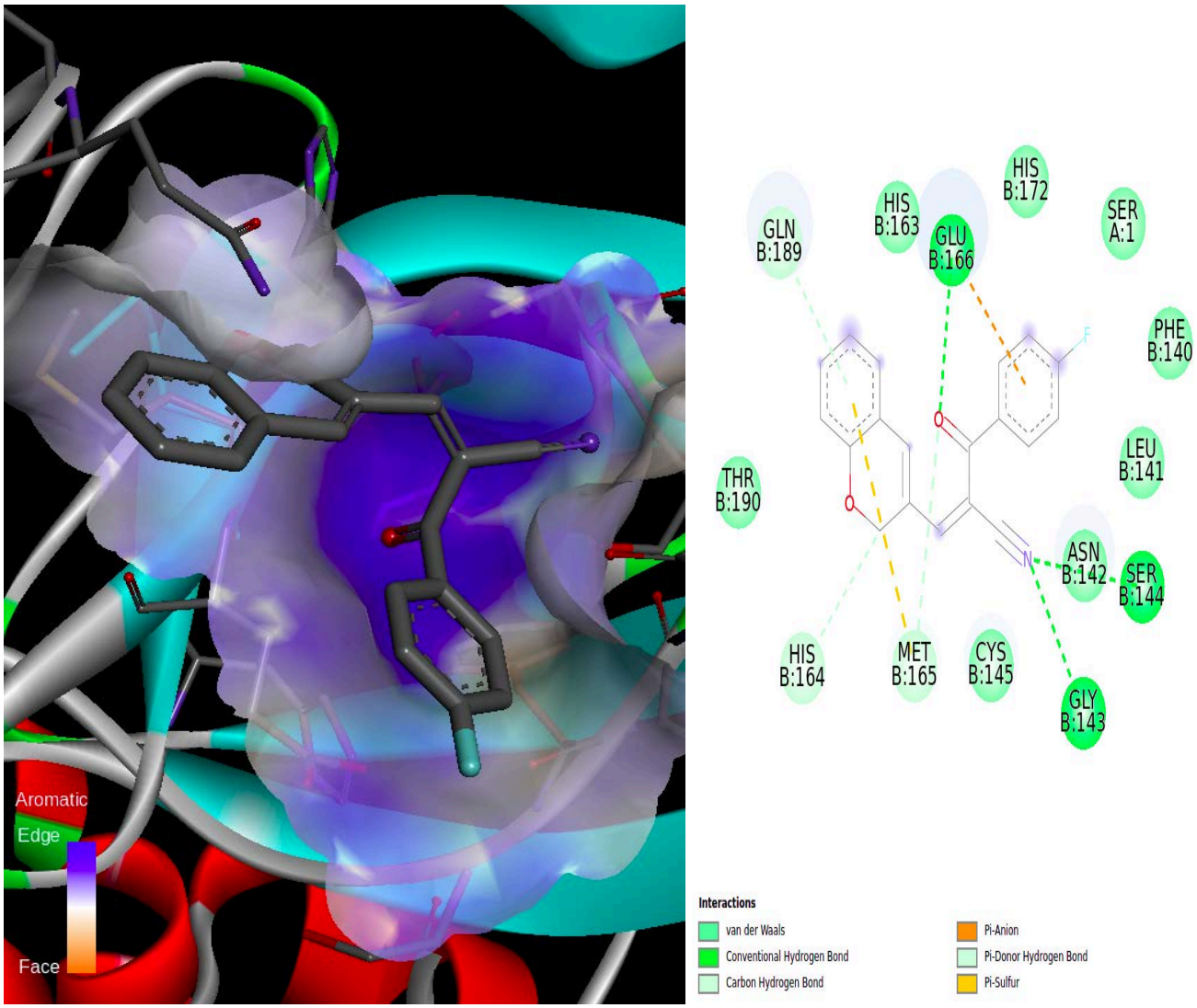


Figure 24: A 3D (left) & 2D (right) diagram of ligand 337222 non-covalently docked to chain B of the receptor.

In Figures 23 and 24, the interaction between chain A and chain B of the protein with ligand 337222 is observed respectively. In Figure 23, there appear to be favorable hydrogen bonds between the protein residues SER-144 and GLY-143. In comparison with chain B (Figure 24), chain A (Figure 23) seems to have more significant interactions i.e. the catalytic dyad CYS-145 and HIS-41 are interacting with the ligand, with HIS-41 having Pi-Pi T-shaped interactions with chain a of the ligand and CYS-145 having conventional hydrogen bonding with the oxygen atom of the ligand, a Pi-alkyl interaction and a

Pi-Sulfur interaction. A halogen interaction in Figure 23 is also observed. It can be concluded that the ligand 337222 interacted better with chain A compared to chain B.

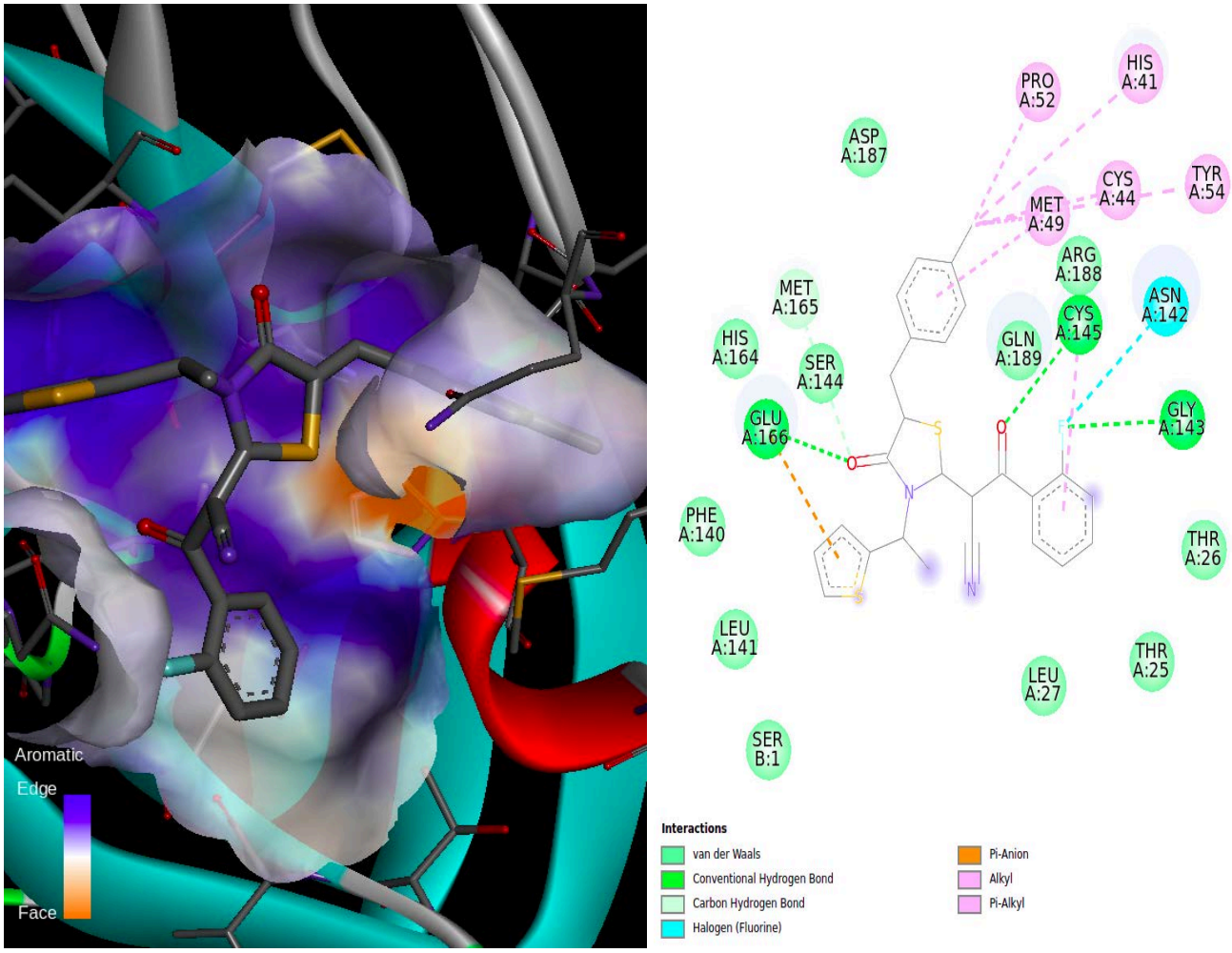


Figure 25: A 3D (left) & 2D (right) diagram of ligand 387305 non-covalently docked to chain A of the receptor.

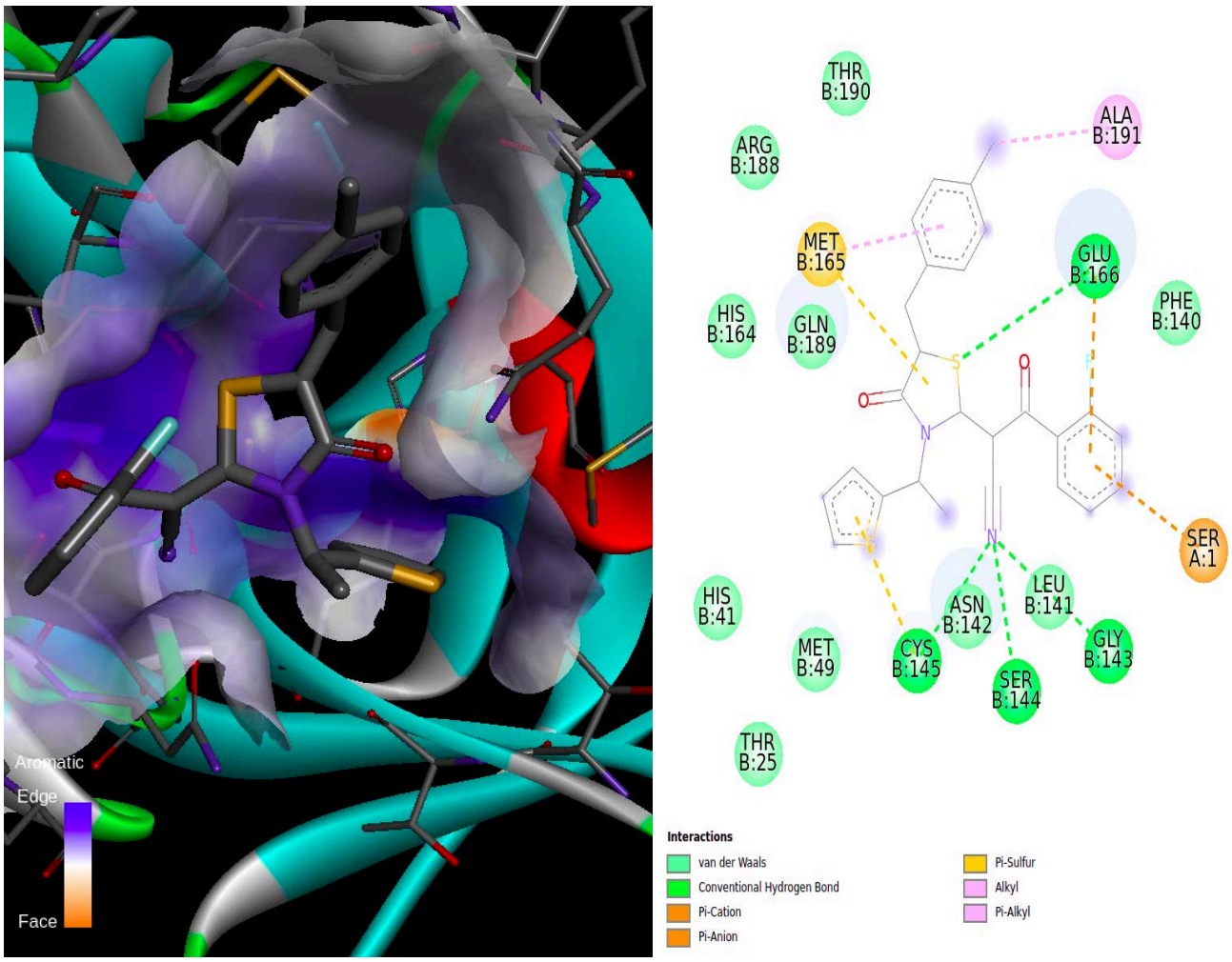


Figure 26: A 3D (left) & 2D (right) diagram of ligand 387305 non-covalently docked to chain B of the receptor.

In Figures 25 and 26 ligand 387305 is non-covalently bonded to the receptor chains A and B. In Figure 25, there appear to be several Alkyl and Pi-Alkyl interactions with the ligand and this is interesting as this is the only complex with this many Alkyl and Pi-Alkyl interactions. There is a nice halogen interaction of ASN-142 and a conventional hydrogen bonding interaction with the fluorine atom of the ligand. The cysteine 145 has two interactions with the ligand, the first being a conventional hydrogen bond with the oxygen atom of the ligand and the second being a Pi-Alkyl interaction. There is a nice hydrogen bond of residue GLU-166 with the oxygen atom of the ligand and a Pi-Anion interaction. In Figure 26, the cysteine 145 has a conventional hydrogen bond interacting with the ligand and Pi-Sulfur

interaction, the HIS-41 has a non-bonded van der Waals interaction. These complexes seem to be stable and hence ligand 387305 appears to be a good inhibitor.

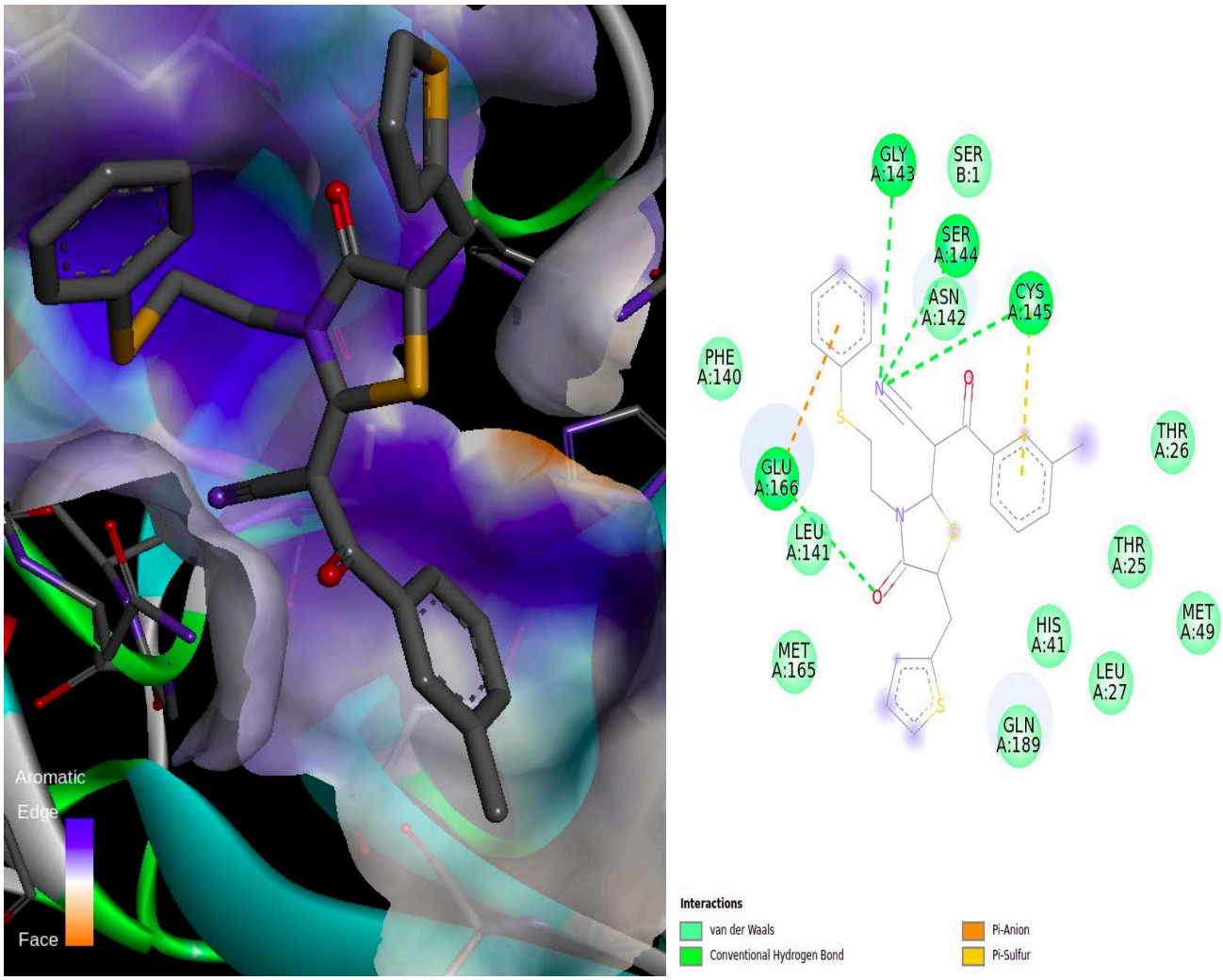


Figure 27: A 3D (left) & 2D (right) diagram of ligand 396939 non-covalently docked to chain A of the receptor.

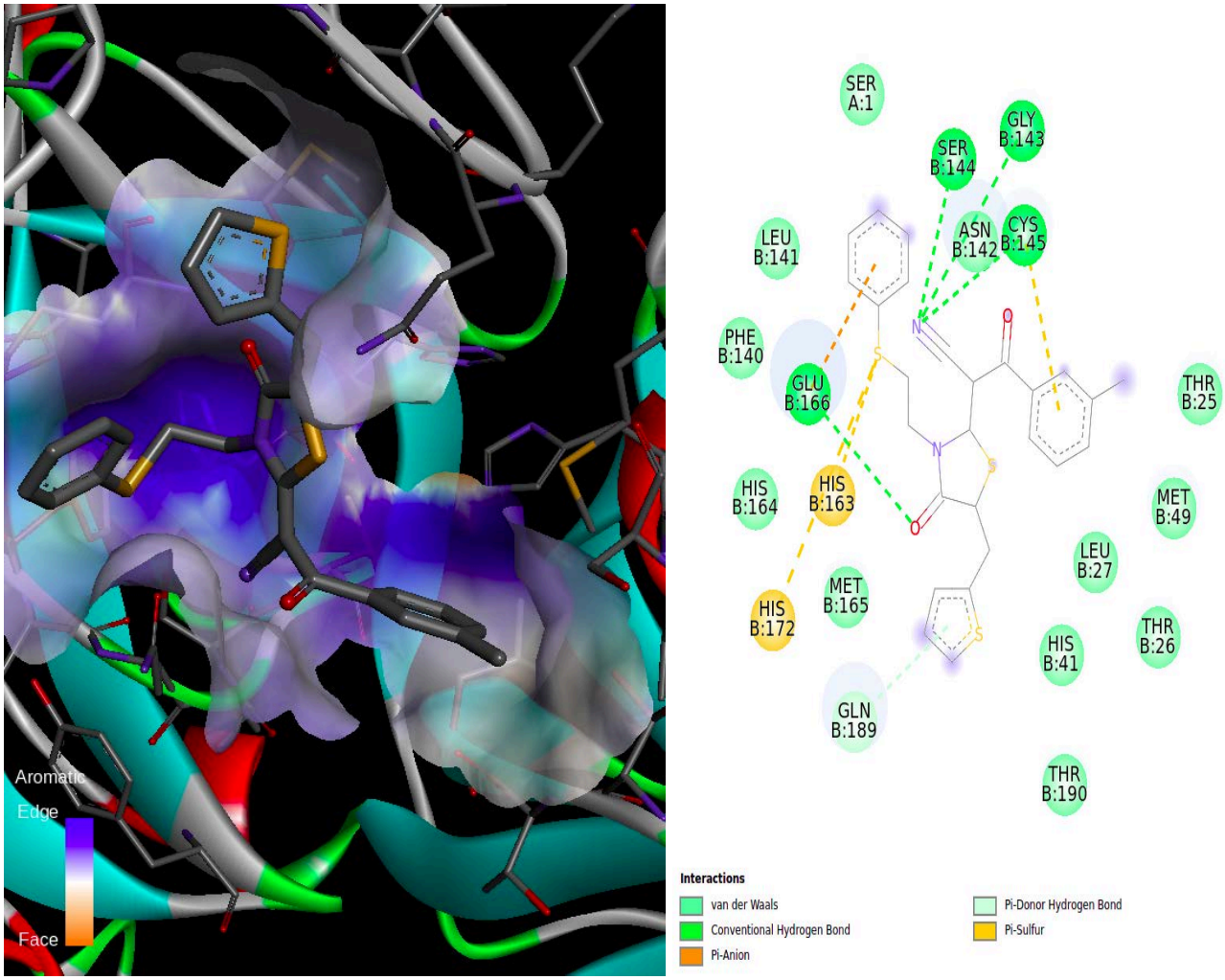


Figure 28: A 3D (left) & 2D (right) diagram of ligand 396939 non-covalently docked to chain B of the receptor.

In Figures 27 and 28, the ligand 396939 is non-covalently bonded to chain A (Figure 27) and chain B (Figure 28) of the receptor. The interactions in these figures are almost identical, with cysteine 145 having a conventional hydrogen bond and Pi-Sulfur interaction with the ligand, GLY-143 and SER-144 having hydrogen bonds with the ligand, and GLU-166 having hydrogen bond and Pi-Anion interactions with the ligand in both figures. The only difference in the complexes is the Pi-Sulfur interaction of protein residues HIS-163 and HIS-172 with the sulfur atom of the ligand.

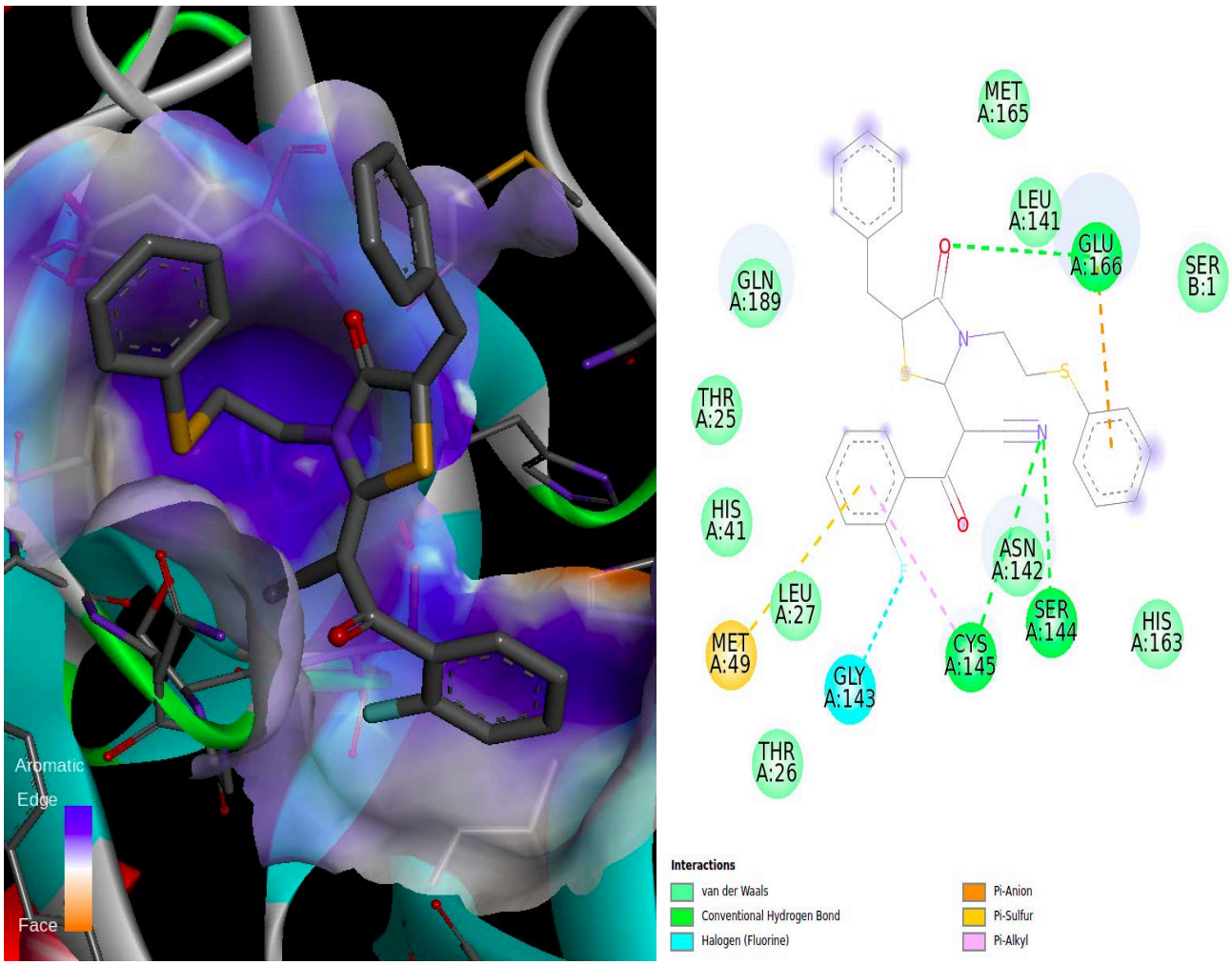


Figure 29: A 3D (left) & 2D (right) diagram of ligand 397136 non-covalently docked to chain A of the receptor.

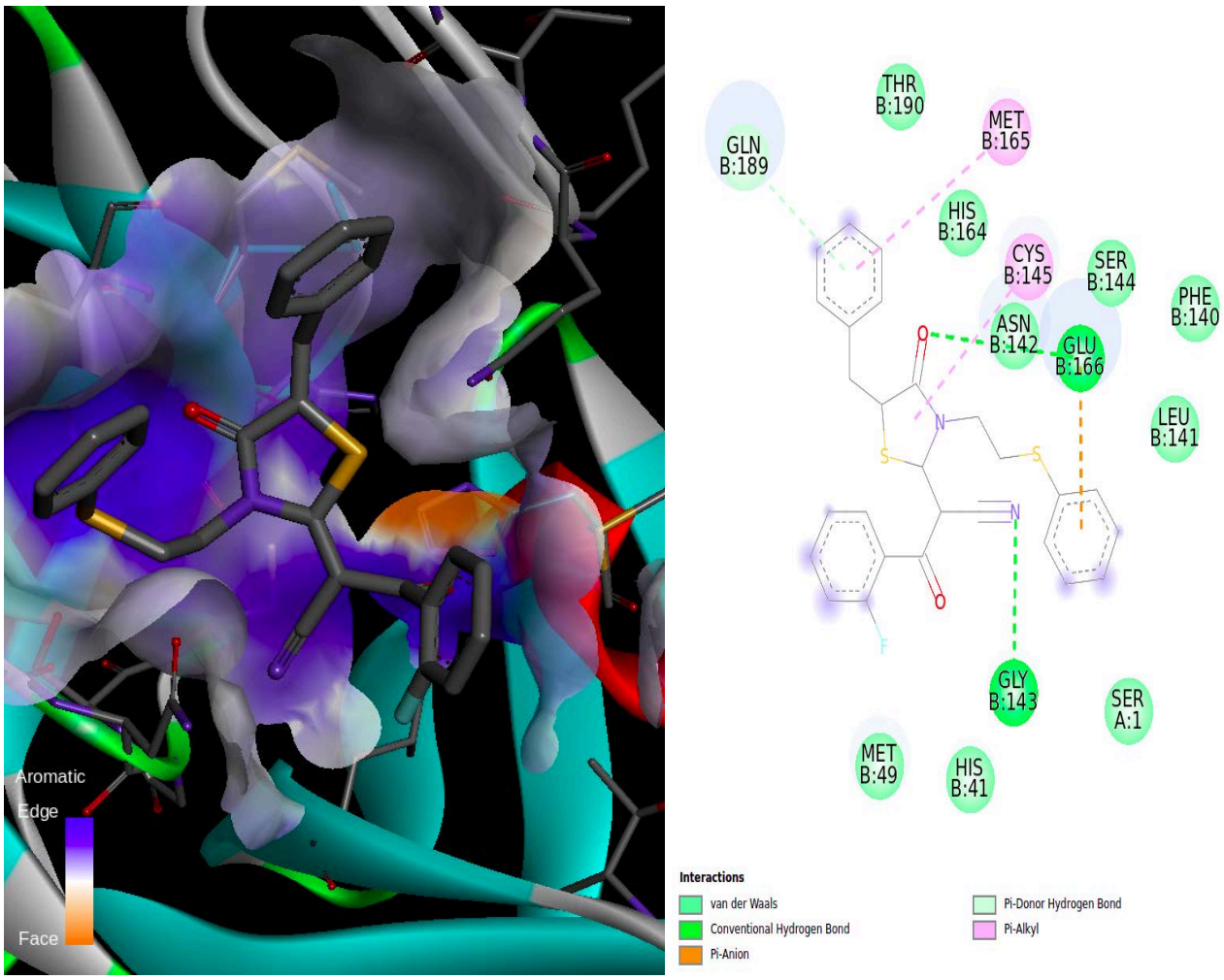


Figure 30: A 3D (left) & 2D (right) diagram of ligand 397136 non-covalently docked to chain B of the receptor.

In Figures 29 and 30, the ligand 397136 is non-covalently bonded to chain A (Figure 29) and chain B (Figure 30) of the receptor. The cysteine 145 in Figure 29 has a hydrogen bond with the nitrogen atom of the ligand and a Pi-Alkyl interaction with the ligand. There are several interactions with the ligand, halogen interaction of GLY-143, a Pi-Sulfur interaction of MET-49, and hydrogen bond interaction of SER-144 and GLU-166. In Figure 31, the CYS-145 and MET-165 have a Pi-Alkyl interaction with the ligand. In both figures, there are several non-bonded van der Waals interactions.

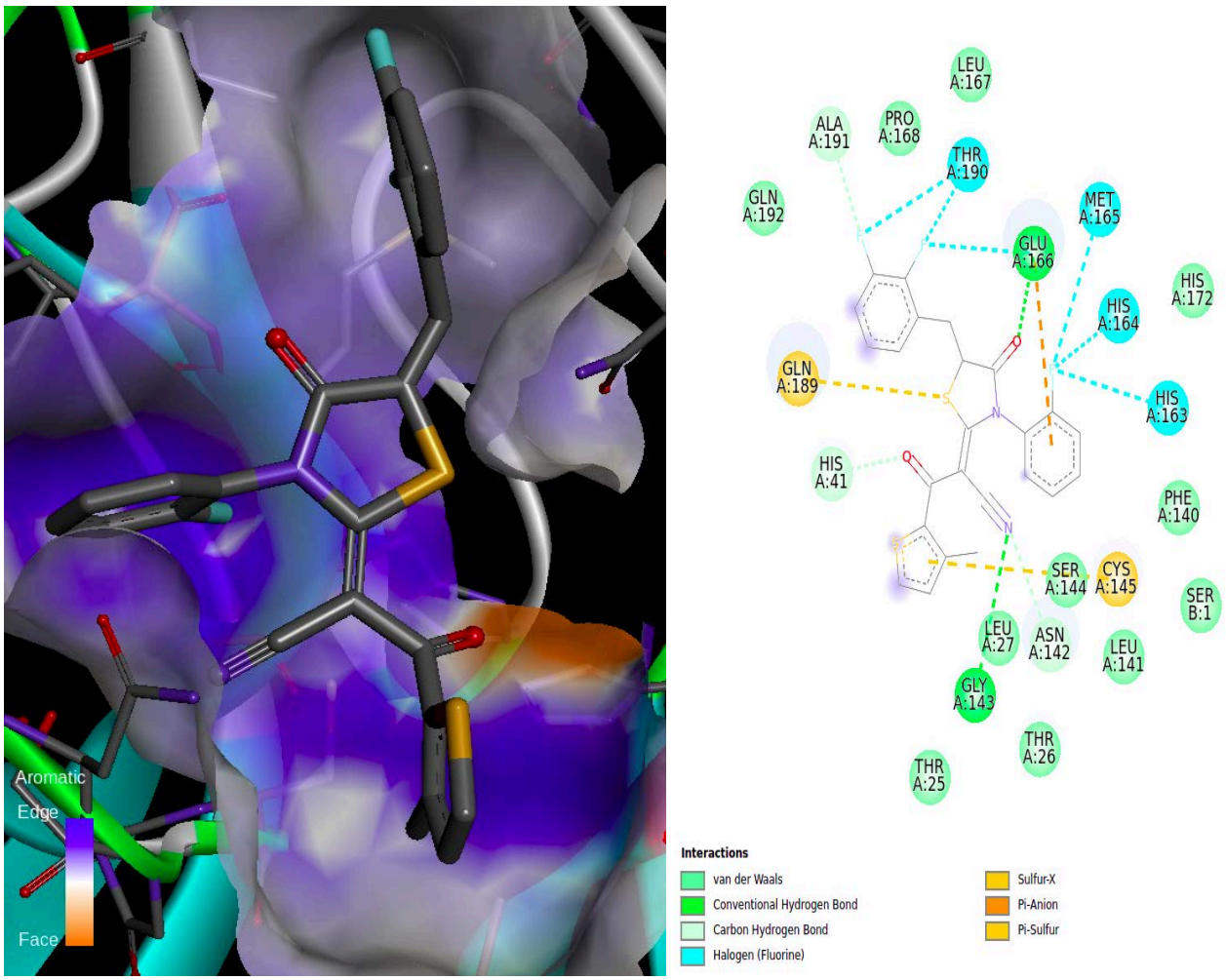


Figure 31: A 3D (left) & 2D (right) diagram of ligand 397730 non-covalently docked to chain A of the receptor.

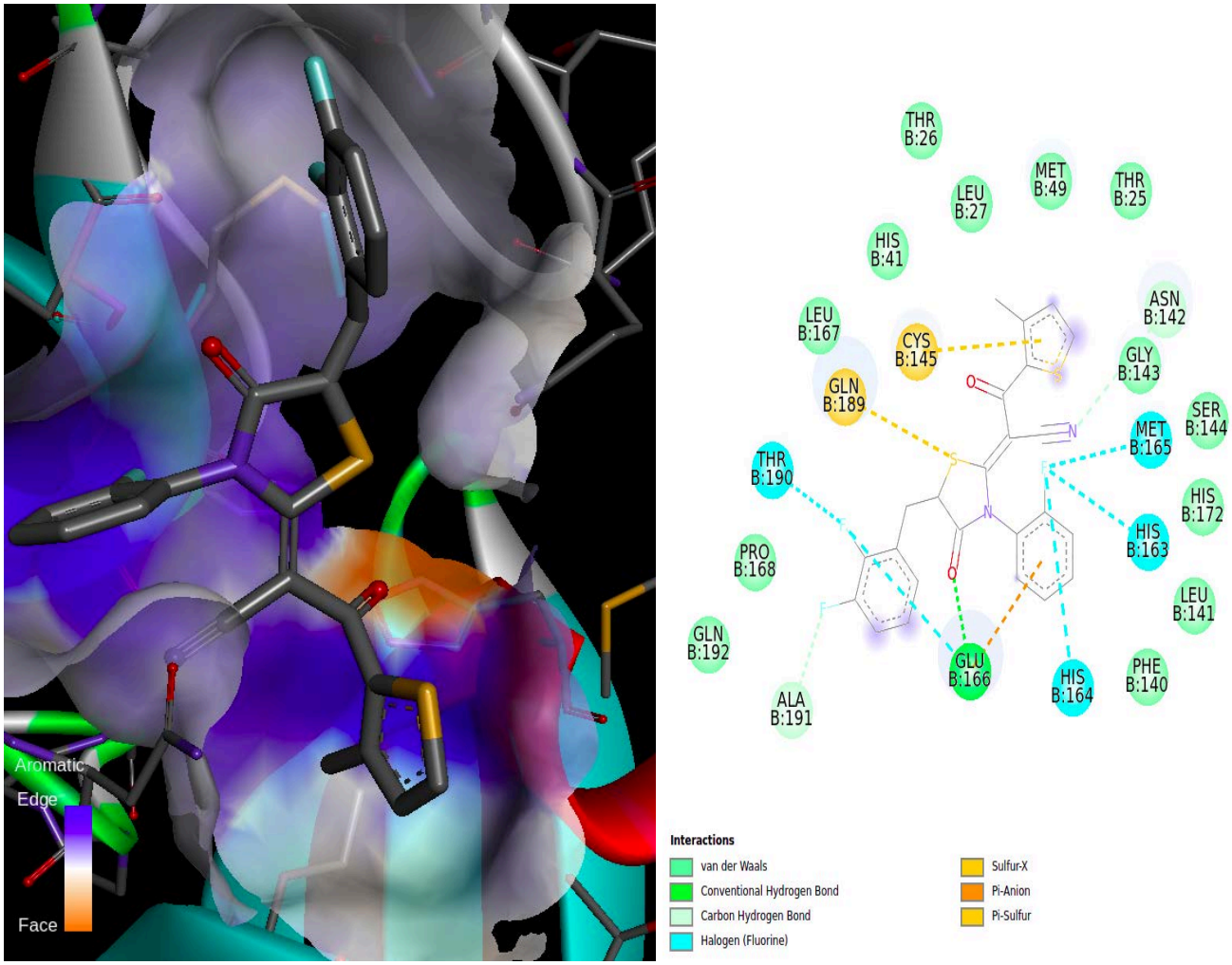


Figure 32: A 3D (left) & 2D (right) diagram of ligand 397730 non-covalently docked to chain B of the receptor.

In Figures 31 and 32, the ligand 397730 is non-covalently bonded to chain A (Figure 31) and chain B (Figure 32) of the receptor. In both complexes, the catalytic dyad CYS-145 has a sulfur interaction with the ligand. In both complexes, the protein residues MET-165, HIS-163, and HIS-164 have a halogen interaction with the ligand. Observing the interactions in both complexes, it seems that the interactions are almost identical.

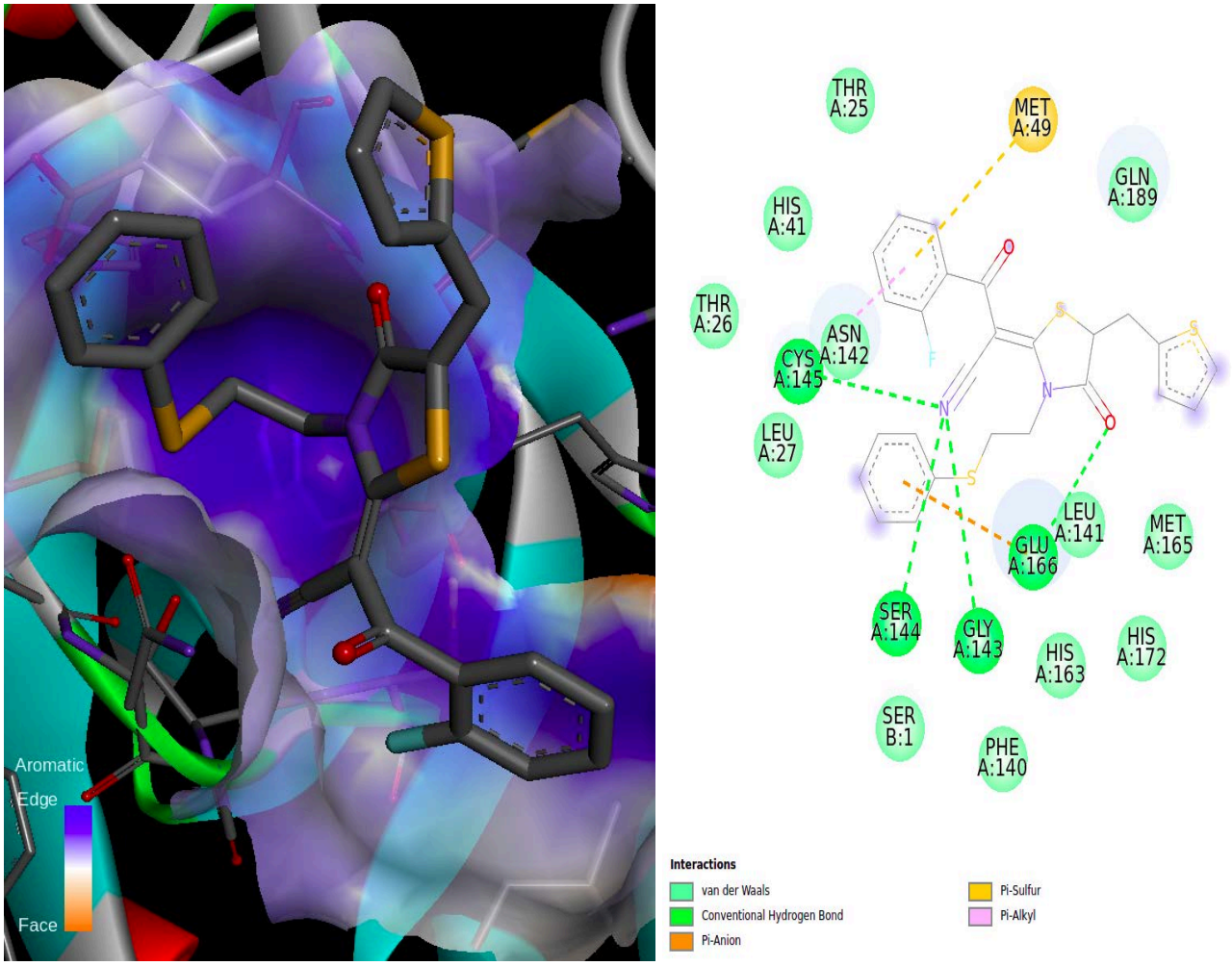


Figure 33: A 3D (left) & 2D (right) diagram of ligand 402091 non-covalently docked to chain A of the receptor.

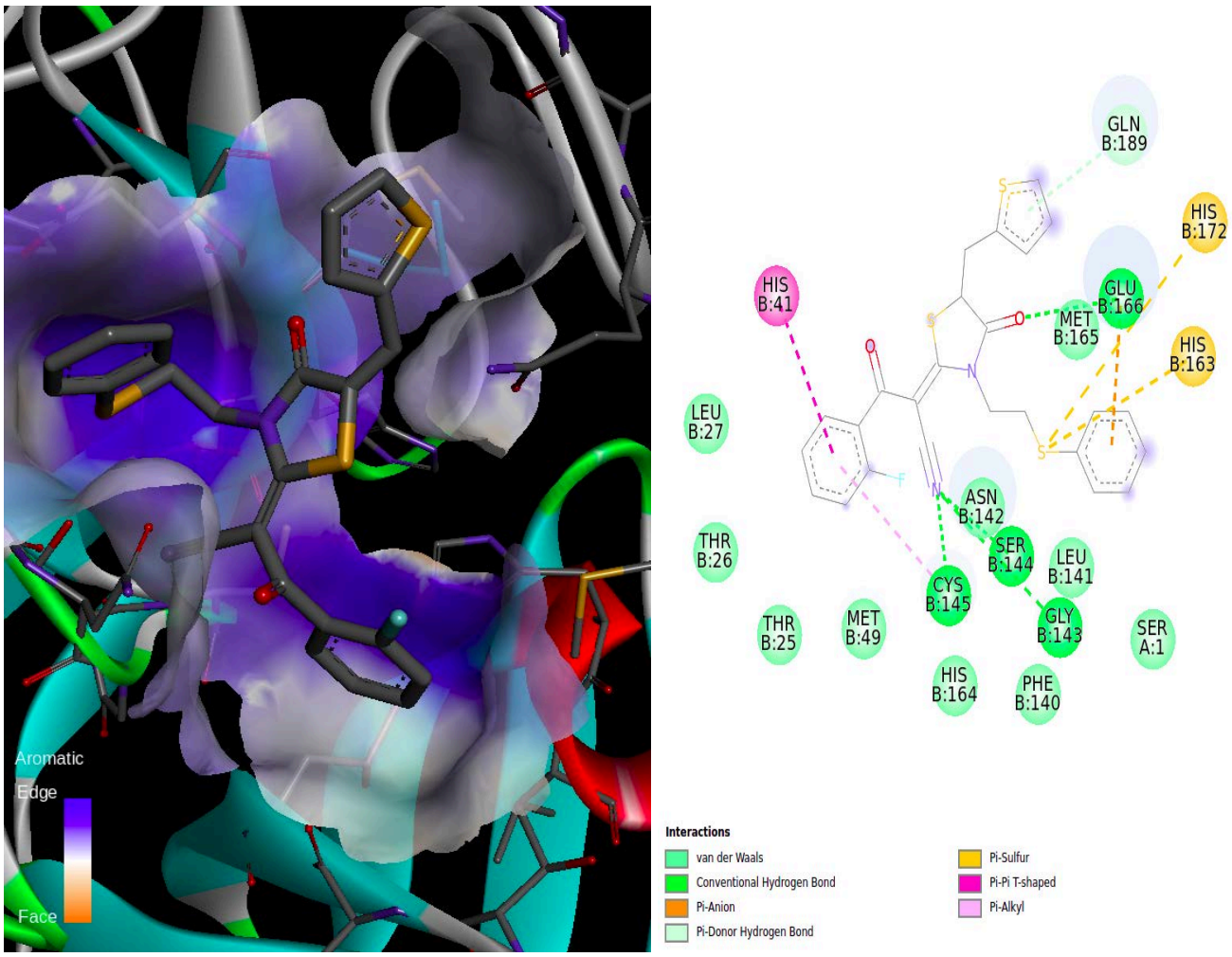


Figure 34: A 3D (left) & 2D (right) diagram of ligand 402091 non-covalently docked to chain B of the receptor.

In Figures 33 and 34, the ligand 402091 is non-covalently bonded to chain A (figure 35) and chain B (Figure 34) of the receptor.

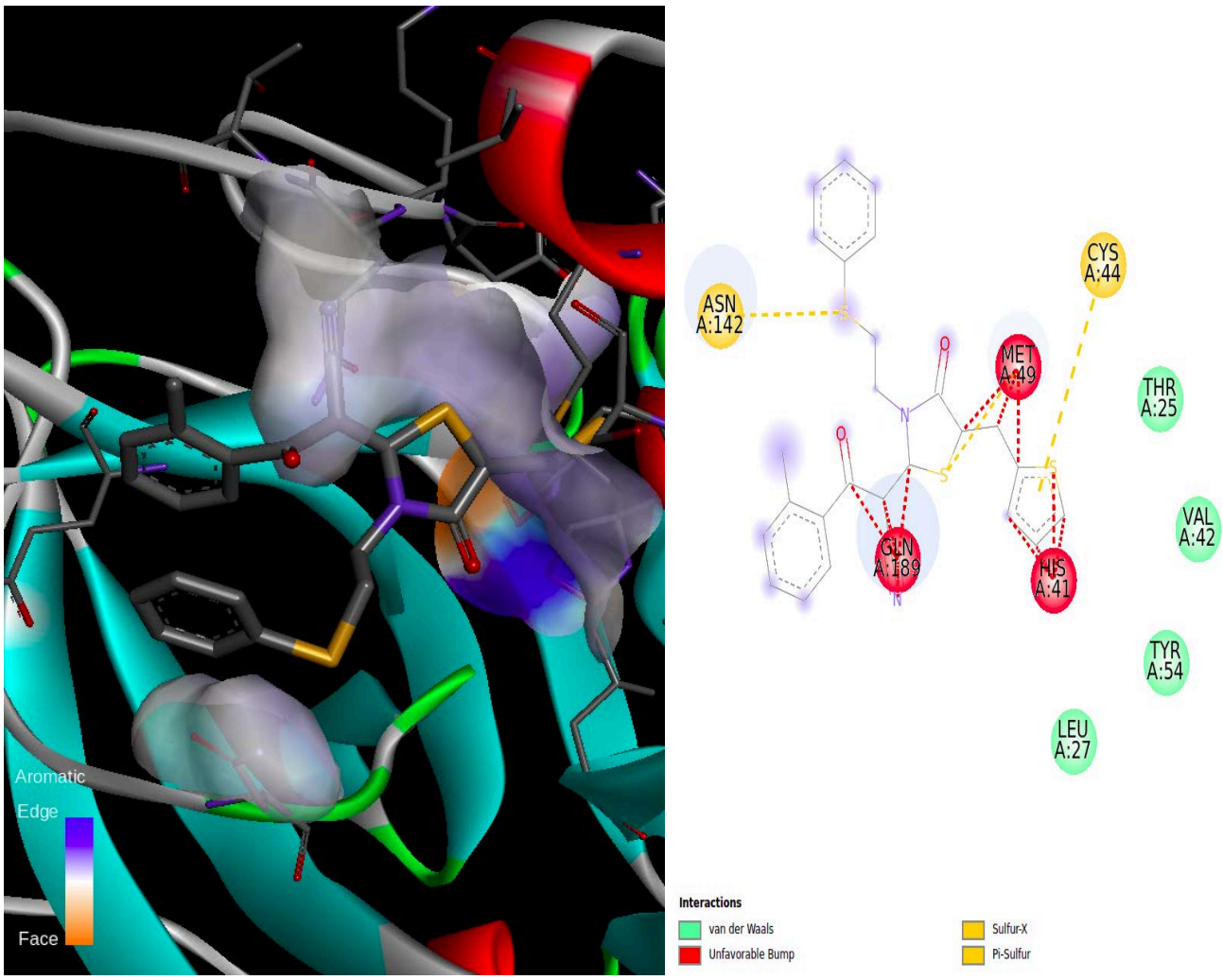


Figure 35: A 3D (left) & 2D (right) diagram of ligand 403456 non-covalently docked to chain A of the receptor.

Figure 35 for ligand 402456 has several unfavorable bump interactions with the ligand with protein residues GLN-189, MET-49, and the catalytic dyad HIS-41. As mentioned before, unfavorable bump interactions in non-covalent molecular docking may indicate that the compound is not a good inhibitor of the protein, this may be due to an "induced fit" problem that generates these unfavorable interactions. However, molecular docking studies alone are not enough to conclude the potency of the inhibitor. Further work for this ligand may be required such as molecular dynamics or wet lab experiments like enzymatic assays, characterization, etc. There are several van der Waals interactions in this complex with protein residues forming Sulfur-X and Pi-Sulfur interactions with Sulfur atoms of the ligand, respectively.

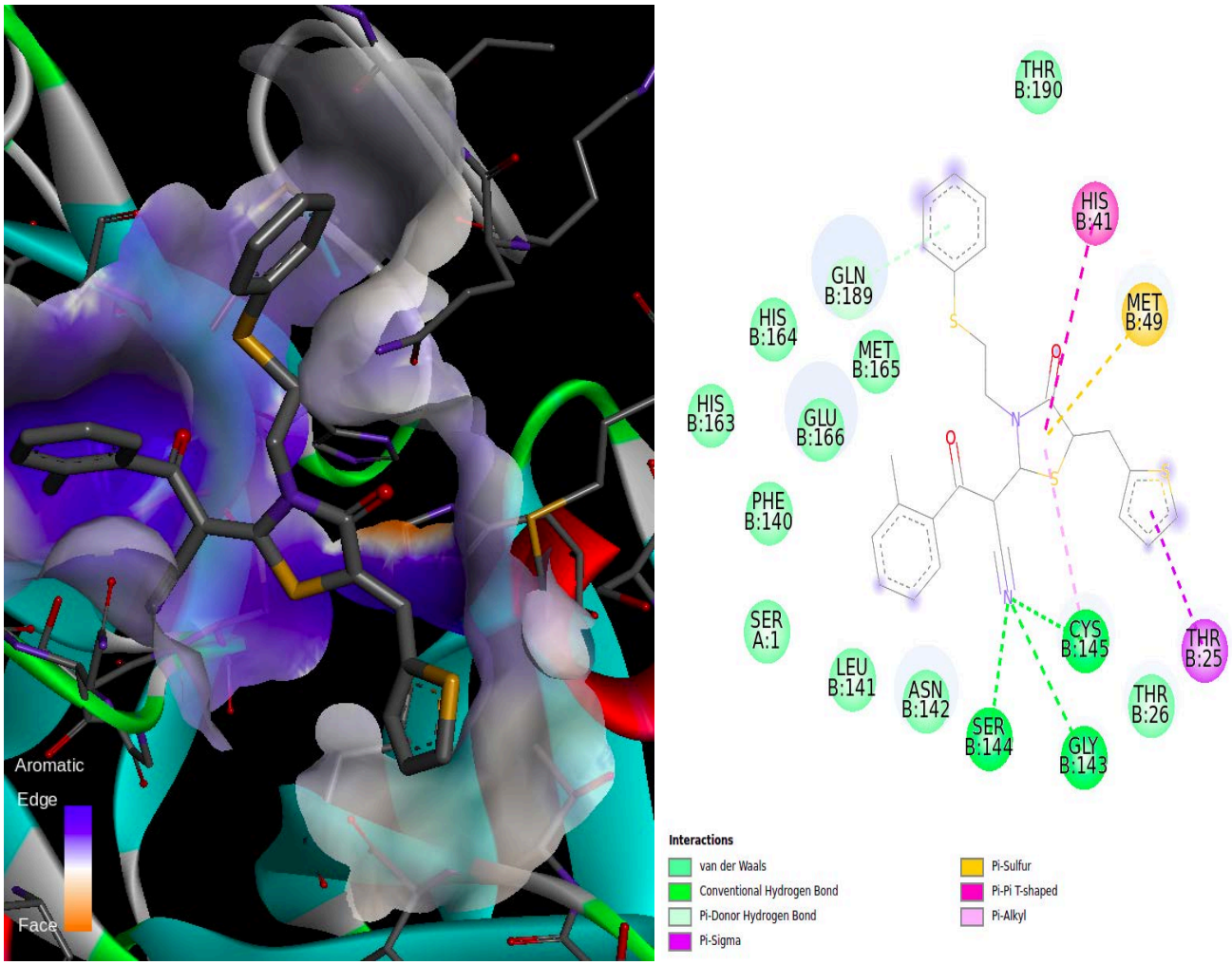


Figure 36: A 3D (left) & 2D (right) diagram of ligand 403456 non-covalently docked to chain B of the receptor.

In ligands that are non-covalently attached, the GLY-143 residue nonetheless frequently forms hydrogen bonds with ligand atoms to stabilize the ligands. This observation together with the additional hydrogen bonds from the catalytic dyad and other residues justifies the stability of the complex. In Figure 36, the catalytic dyad CYS-145 forms hydrogen bonds with the N atom of the ligand and a Pi-alkyl interaction with the Sulfur atom.

2.4.1.2.1 DISCUSSION OF NON-COVALENT DOCKED SYSTEMS.

There are some interesting features of this non-covalent docking both within and without the context of the non-covalent docking. Firstly, all systems are in relative close access ~3Å of CYS145. With ligand 48356 the positioning of the nitrile and the hydrogen bonding to GLY143 is common to both the non-covalently docked (both the docking to chain A and to chain B) and the covalently docked systems. An interaction with GLY143 appears in several systems. For example with ligands 396939 and 402091 the positioning of GLY143 is consistently visible in all docking experiments. For ligands 117238, 397730 and 337222 the commonalities between covalent and non-covalent docked systems are not clear; with 117238 and 397730 individually both non-covalent dockings to chain A and B are similar, which is not the case for 337222. For ligand 387305 the GLU166 hydrogen bonding interaction is important across all non-covalent and the covalent dockings. SER144 is another important residue with consistent contacts with ligand 397136 in all settings, and with ligand 403456 except in the docking to chain A. An argument can now contextually be made supporting most systems from consistent non-covalent through to covalent docking, with perhaps exceptions in ligands 117238, 397730 and 337222. Ultimately the test of ligand stability within the active site (in non-covalently bound systems) is the additional needed step, and, further of more peripheral interest is the behavior and effect of the bound ligand during dynamics of the covalent systems.

CHAPTER SUMMARY

This chapter describes how 2.3 million screening compound analogs and over 250 000 purchasable compounds from the ZINC database were screened for acrylonitrile "warhead" functionality, which made them suitable for use as covalent and non-covalent inhibitors of the SARS-CoV-2 3CL pro main protease. The screening yielded 403804 chemical compounds. These molecules were screened based on chemical features that assess the compounds' drug-likeness, with a particular emphasis on ligand efficiency. Python scripts were used to filter the 403804 compounds down to the 11 best ones. Covalent and non-covalent docking studies were performed on these 11 compounds; the non-covalently docked compounds showed generally encouraging binding energies and distances from the reactive residue. These substances bonded to the 3CL pro receptor both covalently and non-covalently, according to molecular docking analysis. In contrast, in non-covalently attached ligands, the GLY-143 residue frequently forms hydrogen bonds with ligand atoms to stabilize the ligands. This observation, along

with the additional hydrogen bonds from the catalytic dyad and other residues, supports the stability and selection of the non-covalently docked complexes.
Given that the ligand is covalently bound to this cysteine residue, it is expected to be much closer than in a non-covalently bound case, despite the appearance of an unfavorable bump interaction between the ligand and the cysteine in the 2-dimensional representations of covalent docking studies. In cases where this unfavorable bump is observed in non-covalent molecular docking, it may be due to an "induced fit" problem that generates these unfavorable interactions. This highlights the necessity of progressing to molecular dynamics to enable the easing of such interactions.

The interactions of all eleven systems are fully characterized for the cases of covalently bound scenarios to one chain (A) and non-covalent binding to both chains (A and B). These covalently and non-covalently docked systems demonstrated the potential for inhibition, which was confirmed by employing molecular dynamics simulations to assess the stability of the ligands at the active site and the effect of ligand binding on receptor conformation.

CHAPTER THREE: MOLECULAR DYNAMICS

3.1 MOLECULAR DYNAMICS SIMULATION

Computational drug discovery can speed up the challenging process of designing and optimizing a new drug candidate (Jorgensen, 2004). Computational structure-based drug design's (SBDD) impact on drug discovery has escalated in the past decade because of the quick development of faster architectures and much better algorithms for high-level computations in a manner that is time-affordable (De Vivo, 2011). Nowadays, molecular dynamics (MD) simulations allow the implementation of SBDD strategies that fully account for the structural flexibility of the overall drug target model system (Durrant and McCammon, 2011; Harvey and De Fabritiis, 2012). To understand how biological molecules function, including the context of molecular complexes, understanding their structure along with their movements and conformational changes as a function of time is essential (Alonso et al., 2006; Karplus and Kuriyan, 2005).

It is now widely accepted that the two major drug-binding paradigms (induced fit and conformational selection) have superseded Emil Fischer's rigid lock-and-key binding paradigm (Boehr *et al.*, 2009; Changeux and Edelstein, 2011; Vogt and Di Cera, 2012), in which a frozen, motionless receptor was thought to accommodate a small molecule without undergoing any conformational rearrangements. According to Fischer *et al.* (2014) and Abagyan and Totrov (2001), accurate prediction of drug binding and associated thermodynamic and kinetic properties depends on the flexibility of the receptor and ligand.

Molecular dynamics simulation is now pushing the frontiers of computationally driven drug discovery in both academia and industry. Molecular dynamics simulations provide an interaction between structure and dynamics by allowing for an exploration into the conformational energy landscape, thus providing detail on individual particle motions over time (Karplus and McCammon, 2002). Given the ever-increasing role played by MD in the evolution of computationally driven drug discovery, in this chapter, molecular dynamics simulations were used to investigate the ligand-receptor interactions of the filtered acrylonitrile compounds that were more viable for drug discovery ligands and the main protease of SARS-CoV-2. Docking studies provided viable structures, and therefore these simulations would additionally incorporate protein flexibility to analyze the stability of the complexes.

3.2 INTRODUCTION

Molecular dynamics (MD) is a computer simulation method that predicts how every atom or other molecular system will move over time, based on a general model of the physics governing inter-atomic interactions (Karplus and McCammon, 2002). The concept behind MD simulations is that by applying Newton's laws of motion and understanding each atom's location within a protein surrounded by water, one can determine the force that each atom is subjected to. Simulations capture the position and motion of every atom at every point in time. For a predetermined amount of time, the atoms and molecules are allowed to interact, providing insight into the system's dynamic "evolution."

The forces between interacting atoms are estimated using a force field, and the system's overall energy is computed. The integration of Newton's laws of motion produces subsequent configurations of the evolving system during MD simulations, producing trajectories that specify the positions and velocities of the particles throughout time. These MD trajectories can be used to determine a wide range of properties, including free energy, kinetics measurements, and other macroscopic values that can be compared with experimental measurements. The approach was developed in theoretical physics in the late 1950s and is currently used in chemical physics, materials science, biomolecular modeling, and, more recently, drug development (Frrenkel and Smit, 2001; Allen and Tisdelsey, 1989).

Although crystallographic evidence strongly supports the critical role that protein flexibility plays in ligand binding, the high cost and time commitment necessary to produce it have prompted scientists to look for computational methods that can anticipate protein motion. Unfortunately, even the finest supercomputers frequently struggle to handle the complicated calculations needed to explain the bizarre quantum-mechanical motions and chemical reactions of huge molecular systems. The goal of molecular dynamics simulations, which were first created in the late 1970s (McCammon *et al*., 1977), is to get around this restriction by simulating atomic motions using straightforward Newtonian approximations that are less computationally demanding.

3.3 MOLECULAR DYNAMICS STEPS

The Molecular Dynamics steps followed in this study were from the GROningen Machine for Chemical Simulations (GROMACS) Tutorial Protein-Ligand Complex by Justin A. Lemkul, Ph.D., Virginia Tech Department of Biochemistry. Using the Centre for High-Performance Computing (CHPC) with GROMACS version 2018.2.

3.3.1 Generation of the Topologies

The protein structure of the SARS-CoV-2 main protease (PDB code 6XHM) was downloaded from the RCSB website. Once downloaded, it was then visualized using PyMol. Water, PO4, and BME were removed from the structure. The protein topology was generated using pdb2gmx. Gmx pdb2gmx reads a. pdb (or. gro) file and some database files, adds hydrogens to the molecules, and then generates coordinates in GROMACS (GROMOS) format or, optionally,. pdb and a topology in GROMACS format. A run input file can then be created by processing these files.

Force field parameters enable the application of MD simulations to a range of drug design studies, such as the investigation of ligand-receptor interactions and the improvement of structure predictions (Hospital *et al*., 2015). The most widely used force fields include AMBER, CHARMM, NAMD, and GROMACS (Wang et al., 2004; Vanommeslaeghe et al., 2010; Phillips et al., 2005; Pronk et al., 2013). The force field that was used in this study is CHARMM36, obtained from the MacKerell lab website, and the latest force field tarball and cgenff_charmm2gmx.py" conversion script that corresponded with Python 2x were also downloaded from the MacKerell lab website.

Since ligands are species that force fields cannot recognize, it is quite difficult to treat ligands properly in molecular simulations. Any new species' force field parameters must be created and verified in a way that is consistent with the original force field. This derivation often takes the form of various quantum mechanical calculations for the OPLS, AMBER, and CHARMM force fields. Automated tools are preferred for GROMOS force fields because the parameterization process is less obvious and relies on the empirical fitting of condensed-phase behavior. In other words, initial charges and Lennard-Jones parameters are determined for each type of atom, checked for accuracy, and then

improved. Although the result—that fluids mimic their real-world counterparts—is immensely satisfying, the derivation process can be tedious and unpleasant.

For each force field, there is software available that can give parameters compatible with various force fields. In this study, the official CHARMM General Force Field server (CGenFF) was used to generate ligand topologies. The docked ligands were converted to .mol2 format and H atoms using OpenBabel in the following way:

babel -ipdbqt filename.pdbqt -omol2 filename.mol2 -h

The reason for this is that a .mol2 file must be provided as input for CGenFF to gather basic atom-type data and bonded connectivity. Several changes were made to the .mol2 file. The .mol2 files were then uploaded to CGenFF, and the server quickly returned a topology in the form of a CHARMM "stream" file (extension.str). The cgenff_charmm2gmx.py script acquired from the MacKerell website was utilized to run simulations in GROMACS because the topology file in CHARMM format cannot be used. Executing the python script used to effect conversions from CGenFF format, the cgenff_charmm2gmx.py script was affected as follows;

python cgenff_charmm2gmx_py2.py LIG filename.mol2 topologyfilename.str charmm36-jul2022.ff

Using GROMACS editconf, .gro files were created for the ligands and then combined with a .gro file that contained the processed, force field-compliant structure of the protein to form a complex, with corrections made to the complex file to reflect the combination.

3.3.2 Solvation

The next step was to define the unit cell and fill it with water. To accurately simulate proteins, one must take into account the aqueous environment of the system under analysis. The TIP3P, TIP4P, SPC, extended SPC/E, and F3C models, which all represent bulk water at ambient temperatures, are frequently used to research biomolecular systems (Jorgensen *et al.,* 1983; Berendsen *et al*., 1987; Levitt

et al., 1997). This study used the transferable intermolecular potential 3P (TIP3P) water model, which is noted for being sufficient for simulations in standard conditions, although it can reproduce neither the temperature-dependence of pure water properties nor hydrophobic hydration for simple solutes (Horn *et al*., 2004; Paschek, 2004). Since most interactions between water and biomolecules involve first- or second-shell hydration and because the absence of the long-range structure was not found to be problematic, the TIP3P model offers an acceptable approach to energetics. If a simulation using this model proves to be unsatisfactory, it is possible to modify the water models in the existing force fields by changing the energy function parameters to take changes in solvent properties into account (Best and Mittal, 2010). The GROMACS commands used to solvate the unit cells were;

gmx_mpi editconf -f complex.gro -o newbox.gro -bt dodecahedron -d 1.0
gmx_mpi solvate -cp newbox.gro -cs spc216.gro -p topol.top -o solv.gro

3.3.3 Adding Ions

Since life does not exist at a net charge, ions need to be added to the system (GROMACS Tutorial). An ions.mdp file was used to construct a .tpr file with grompp using the following command;

gmx_mpi grompp -f ions.mdp -c solv.gro -p topol.top -o ions.tpr

The .tpr file was then passed to the GROMACS genion, by calling gromacs in the following way;

gmx_mpi genion -s ions.tpr -o solv_ions.gro -p topol.top -pname NA -nname CL -neutral

3.3.4 Energy Minimization

With the simulation temperature raised to 300 K under NVT ensemble for 100 ps, energy minimization assures the system has no steric conflicts or geometric mistakes caused by the solvation and addition of

ions before simulation. A system topology file (.tpr) was created by the GROMACS gmx grompp preprocessor using the command below and used by the GROMACS mdrun minimization tool.

gmx grompp -f em.mdp -c solv_ions.gro -p topol.top -o em.tpr

GROMACS mdrun then affected the minimization. A GROMACS .mdp file was edited to control the energy minimization steps (em.mdp). The Steepest Descents converged to Fmax < 1000 in 1574 steps. The Potential Energy was equal to -1.2025008e+06. The Maximum force was equal to 9.4289288e+02 on atom 9280. The Norm of force was equal to 1.5047082e+01.

3.3.5 Equilibration

System equilibration was used to optimize how the solvent interacted with the solute in the system. To equilibrate a protein-ligand complex requires the application of restraints to the ligand, this was done by applying first-position restraints to the topology file.

3.3.6 Production MD

At this stage the system was properly equilibrated at the correct temperature and pressure following the two equilibration phases. Now that the position restrictions have been lifted, production MD was used to gather data using the following command;

gmx_mpi grompp -f md.mdp -c npt.gro -t npt.cpt -p topol.top -n index.ndx -o md.tpr

The mdrun was performed by sending the job to the CHPC cluster using GROMACS version qgromacs_2018-6. The production MD simulation was run for 50ns.

3.3.7 Analysis

Following MD runs, the complex was removed from the PBC simulation box and centered using GROMACS trjconv. The MD runs were analyzed using principal component analysis, RMSD, RMSF, Rg, and hydrogen bonds, with the .xvg viewed using XMGRACE and the PCA viewed using KNIME.

3.4 RESULTS

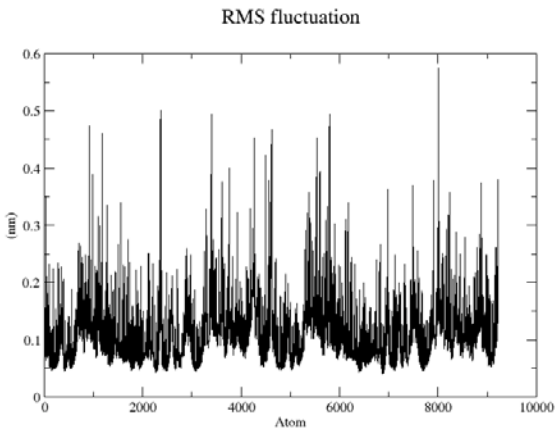
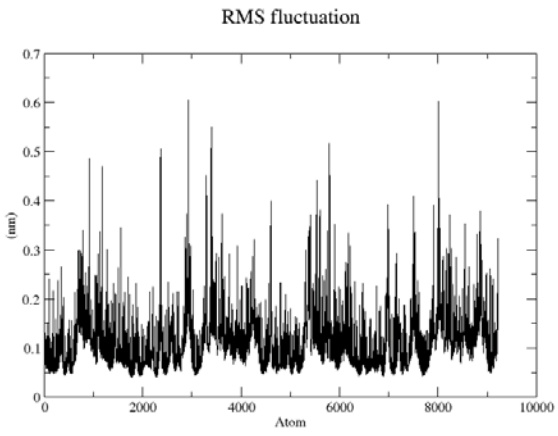
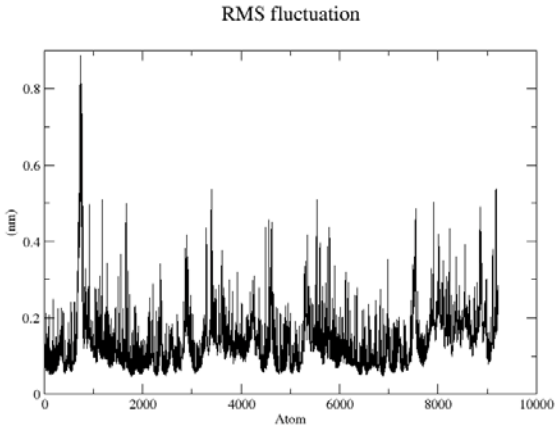
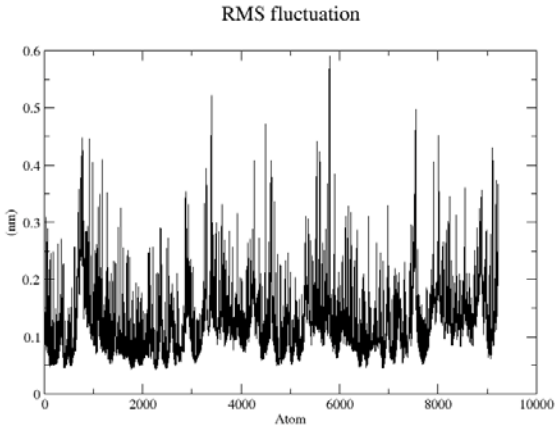
3.4.1 ANALYSIS OF NON-COVALENT MOLECULAR DYNAMICS TRAJECTORIES

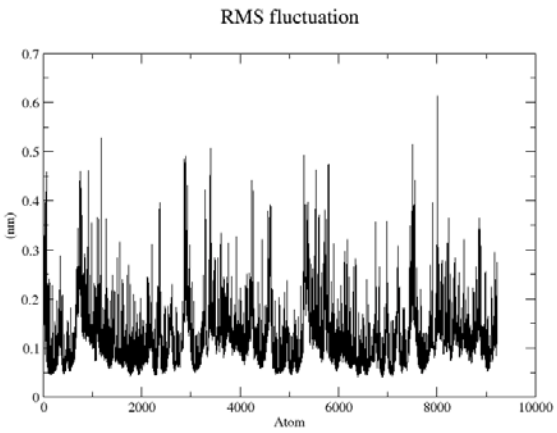
Trajectories produced by molecular dynamics simulations show how a system has changed over time and are subject to both quantitative and visual analysis. Numerous commonly performed analyses include RMSD, Rgyr, RMSF, and others (Kožić and Bertoša, 2024).

3.4.1.1 RMSF

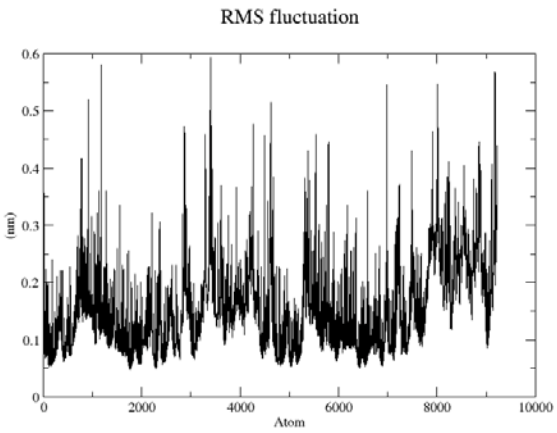
The root mean square fluctuations (RMSF) were calculated on the proteins for each of the trajectories from gromacs to capture the mobility/fluctuation about its position for each atom; it shows the flexibility of regions of the molecule. In Table 1, higher levels of flexibility of the protein are seen by the high peaks displayed in the plots. It is notable in all RMSF results that the protein is stable, with a maximum RMSF of around 0.5nm-0.25nm except for ligand 403456. Ligand 403456 showed very high fluctuations and higher peaks; this indicates that the structure is not well defined, and this is supported by the PCA results of this ligand in Table 4, indicating some instability in the types of motion of the protein.

Table 1: XMGrace generated scatter plot projects of RMSF results for 50 ns production MD simulations.

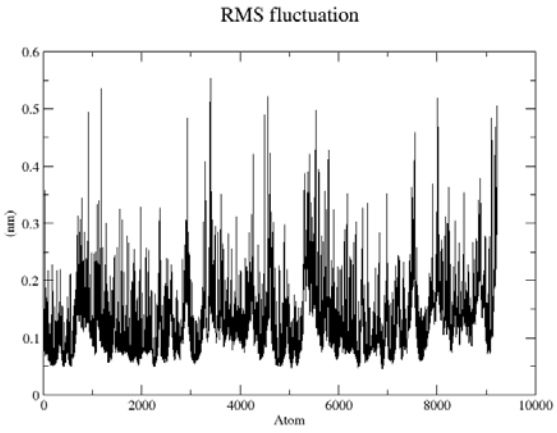
 48356	 387305
 396939	 397136



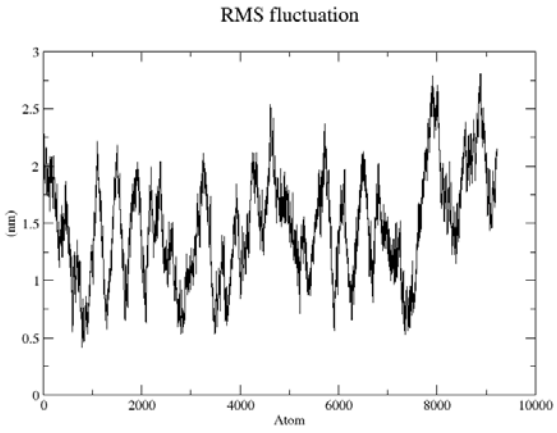
397730



402091



117238



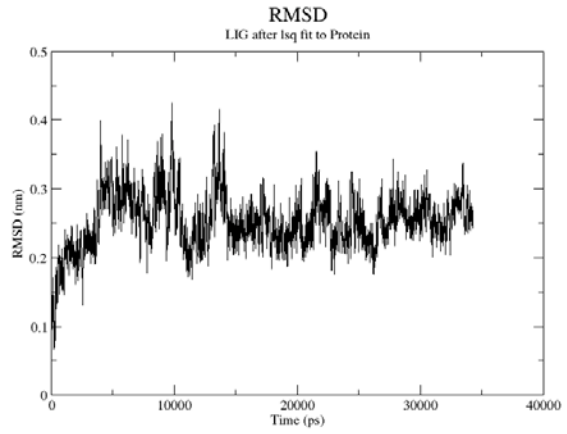
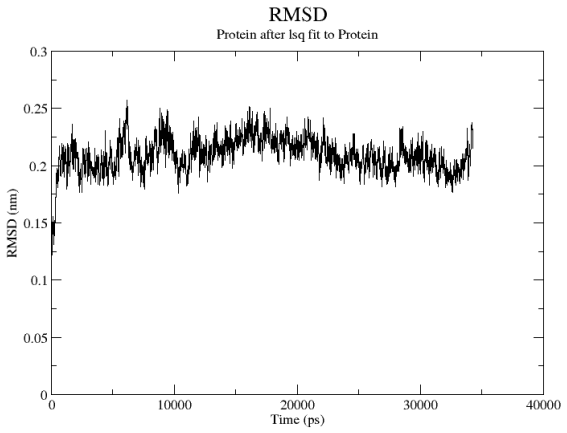
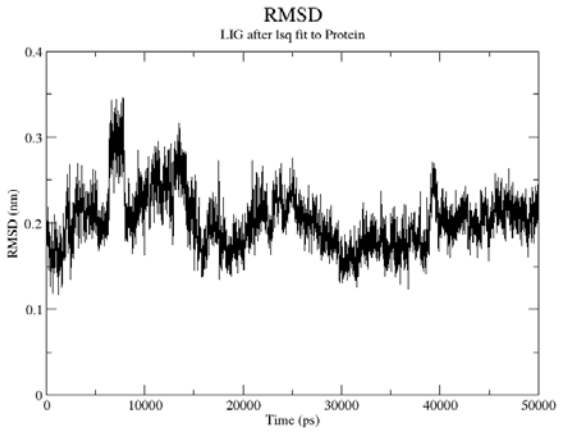
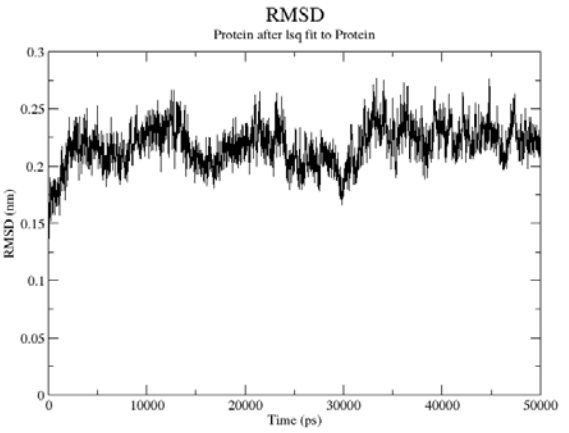
403456

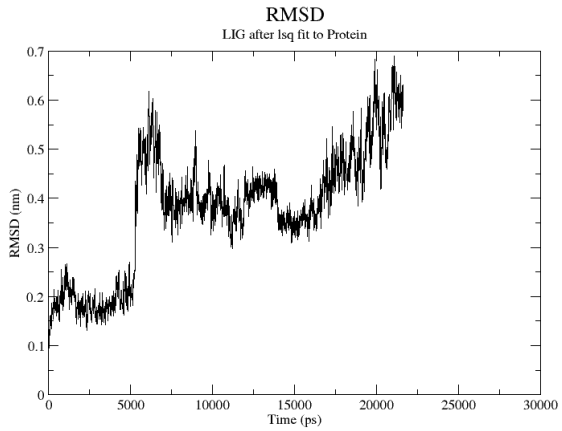
3.4.1.2 RMSD

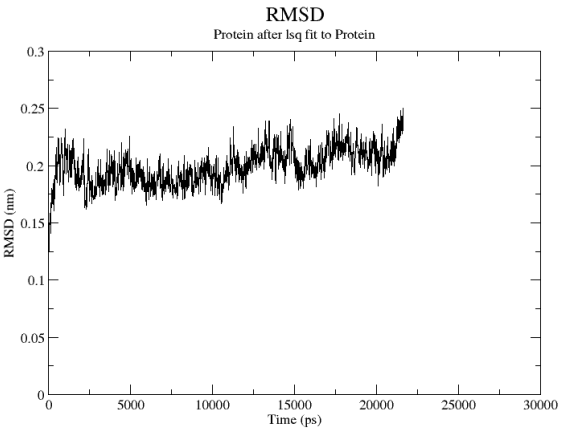
The root mean square deviation (RMSD) was calculated to measure the atomic position deviation and display any structural changes that occurred throughout the MD simulation. Generally, observing Table 2, it is notable that the ligands deviate more than the protein in all simulations. Ligands 48356, 117238, 387305, 397136, 397730, and 402091 seem to be stable complexes in terms of RMSD. The variations generated during the simulation of the protein can be used to gauge its stability in terms of its conformation. The RMSD plot's growth indicates that the protein/ligand gradually departs from its initial structure throughout the simulations. It is interesting to note the jumps in RMSD in ligands

396939 and 403456; these higher values are not due to variation from the initial structure. This is due to the periodic boundary conditions, which indicate that the ligands are crossing the PBC during the simulation.

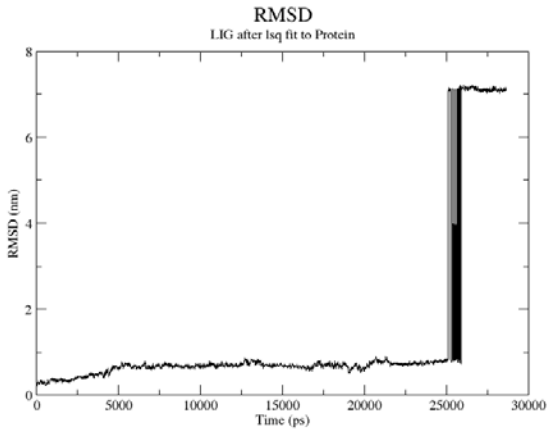
Table 2: XMGrace generated scatter plot projects of RMSD results for 50 ns production MD simulations. With plots on top analyzing the deviations of the ligand and the plots at the bottom analyzing deviations from the protein,

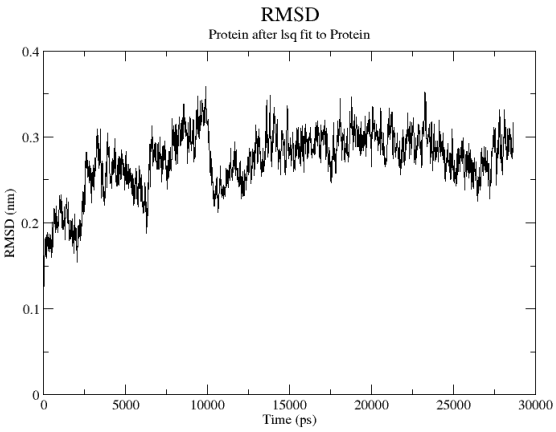
  48356	  117238



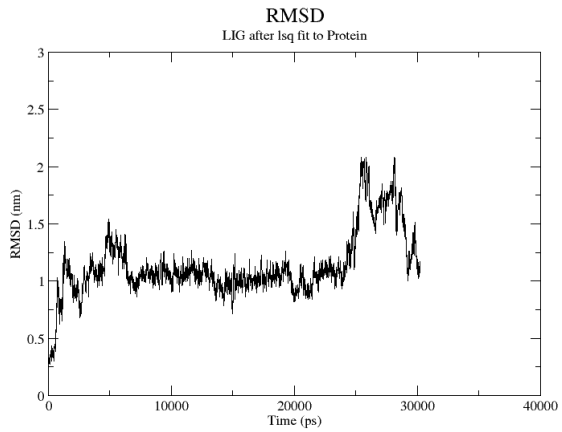


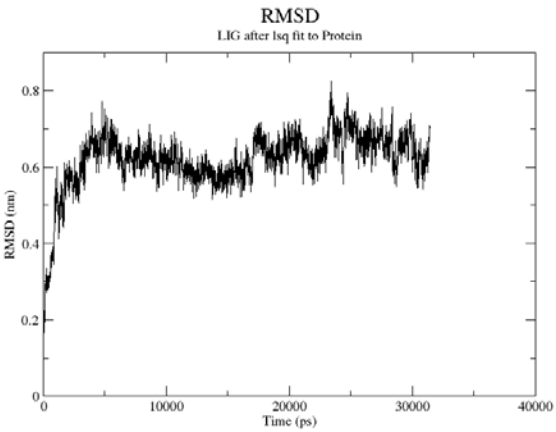
387305

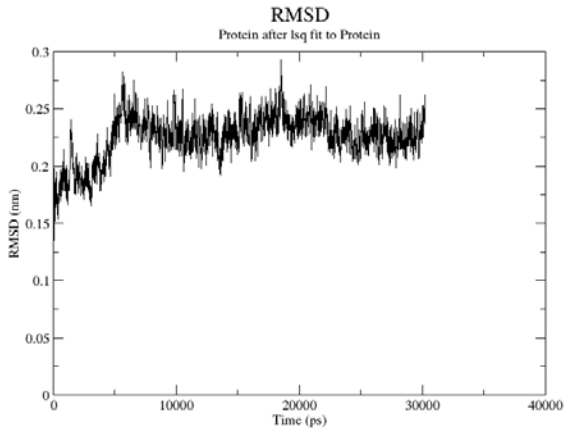




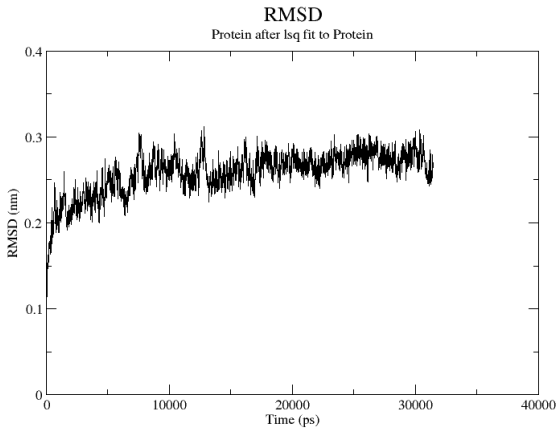
396939



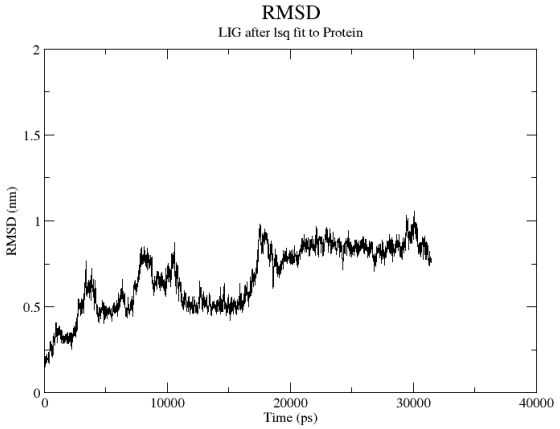


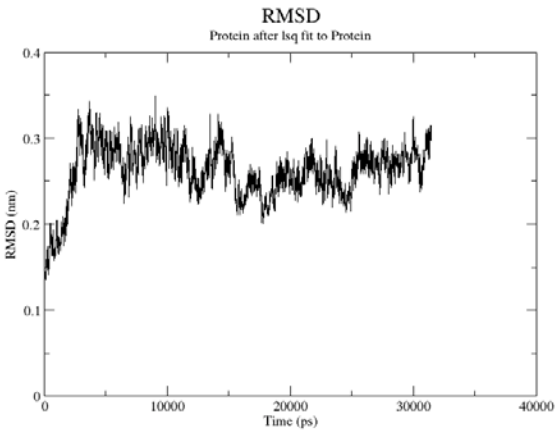


397136

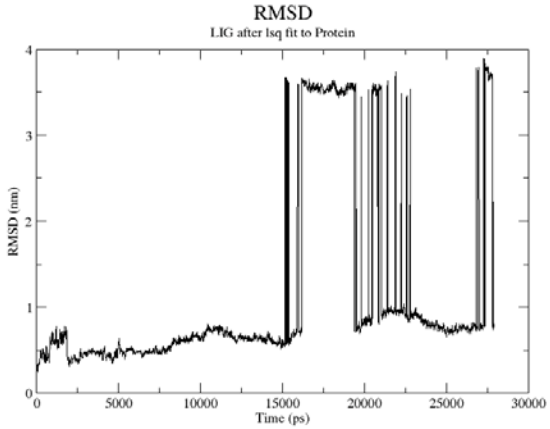


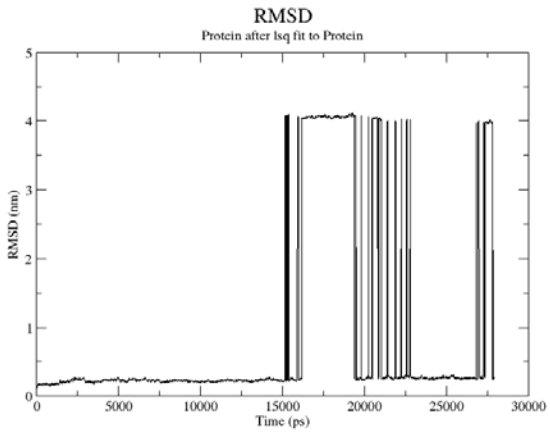
397730





402091



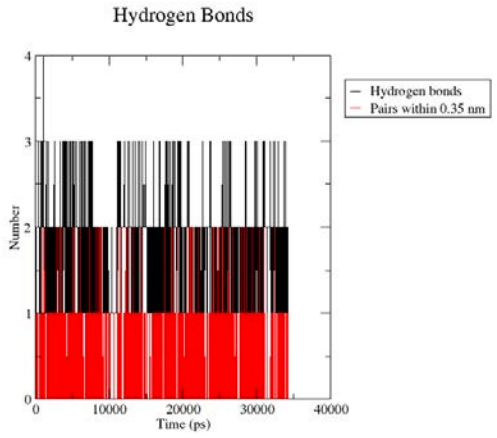
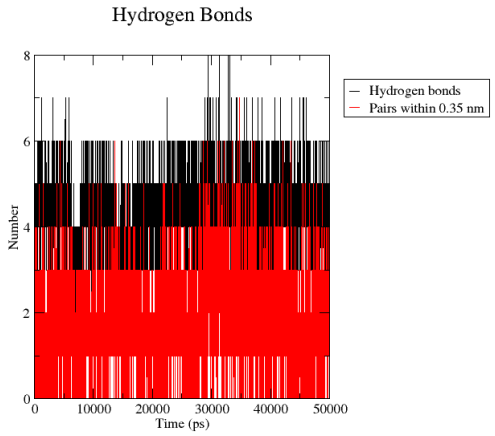
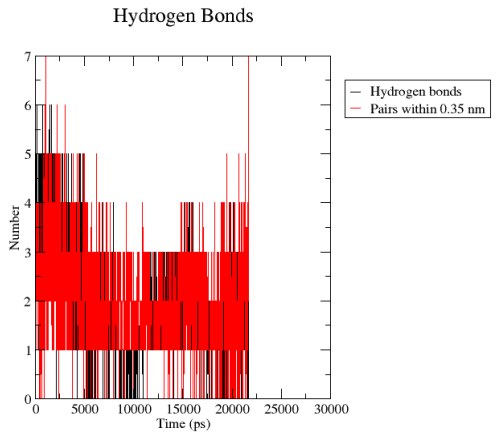
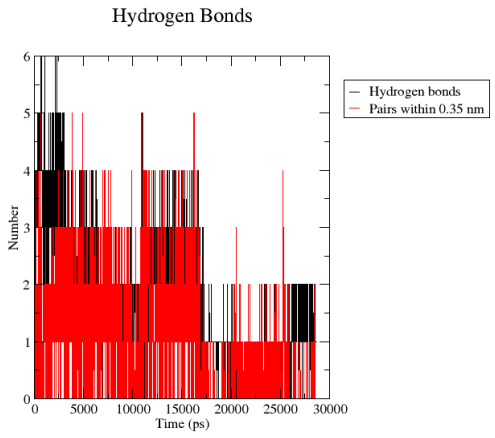


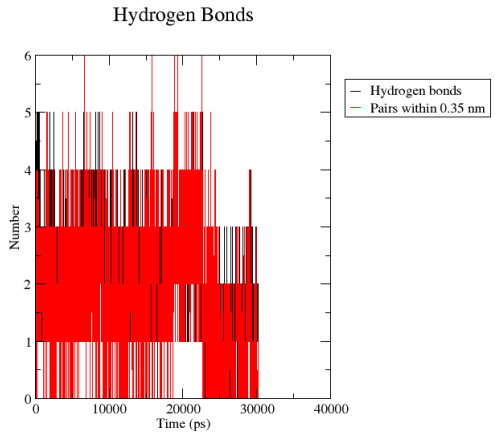
403456

3.4.1.3 Hydrogen Bonding

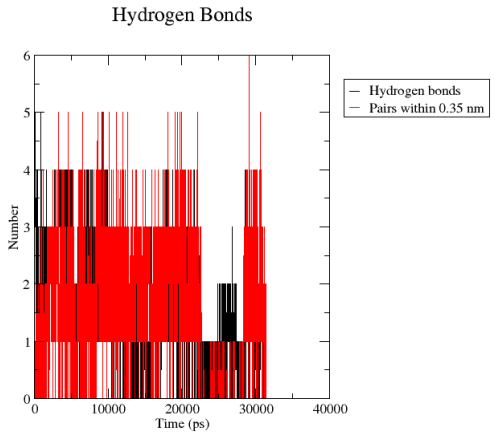
In general, hydrogen bonds are thought to facilitate interactions between proteins and their ligands, which is why hydrogen bond analysis is thus a crucial post-MD consideration. To analyze the complexes in the simulation, hydrogen bonds were calculated. To calculate hydrogen bonds between all possible donors D and acceptors, the program gmx hbond was used. Table 3 shows the variation in hydrogen bonding through dynamics. Ligand 117238 showed a greater number of hydrogen bonds through the simulation as compared to other ligands.

Table 3: XMGrace generated scatter plot projects of hydrogen bond results for 50 ns production MD simulations.

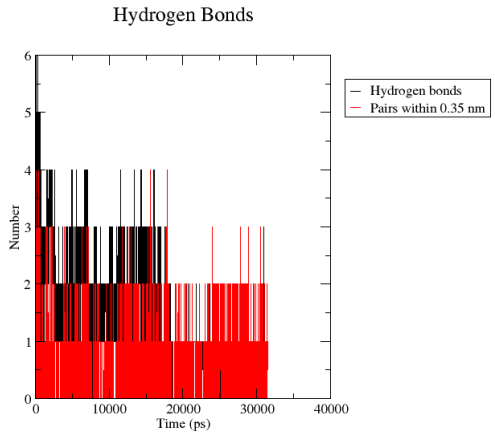
 48356	 117238
 387305	 396939



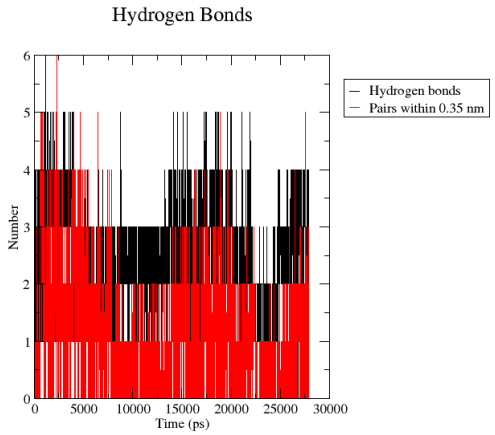
397136



397730



402091

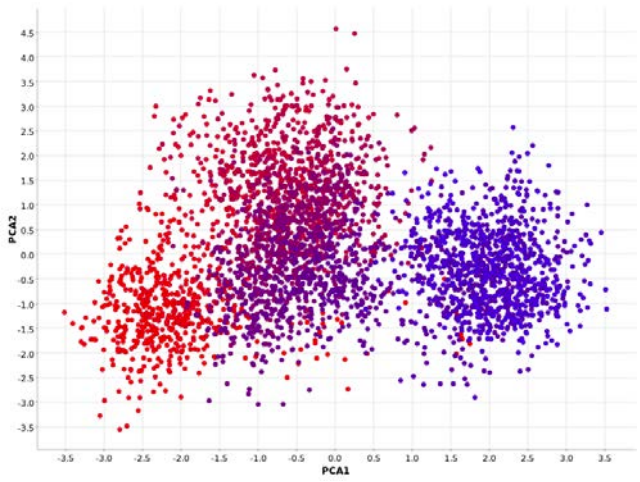
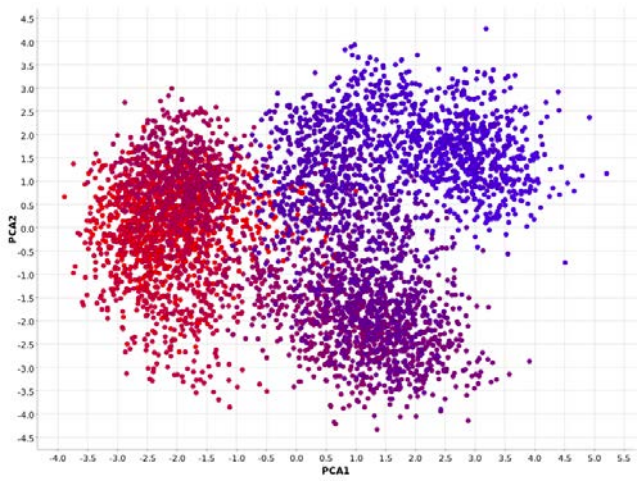
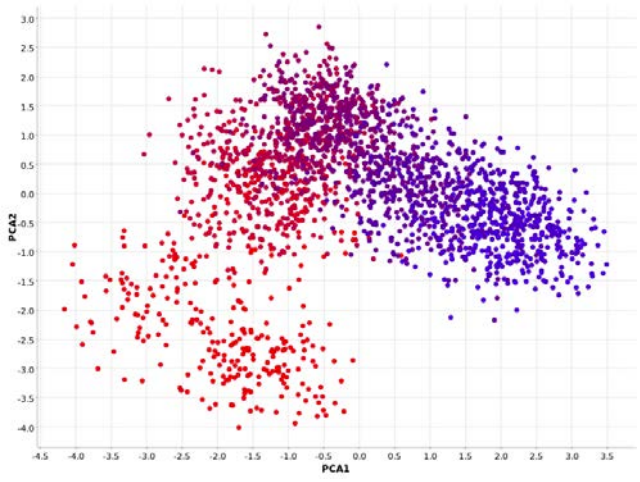
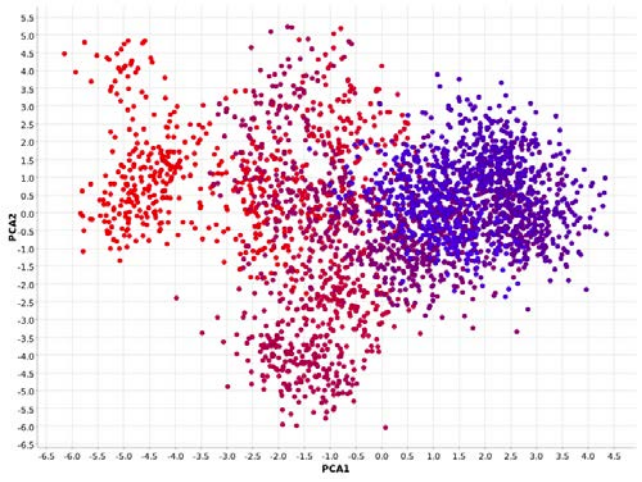


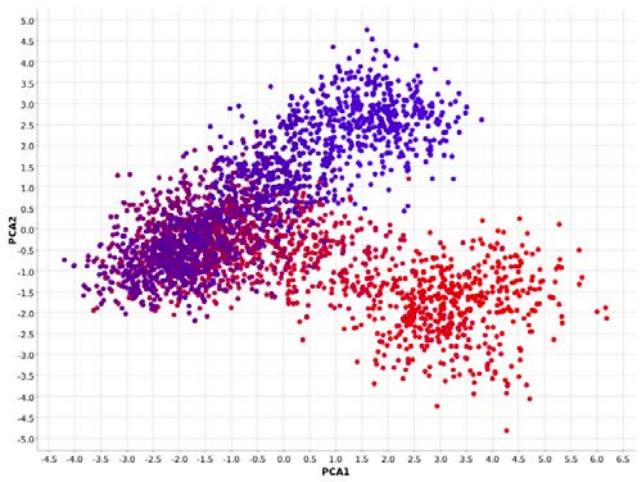
403456

3.4.1.4 Principal component analysis (PCA)

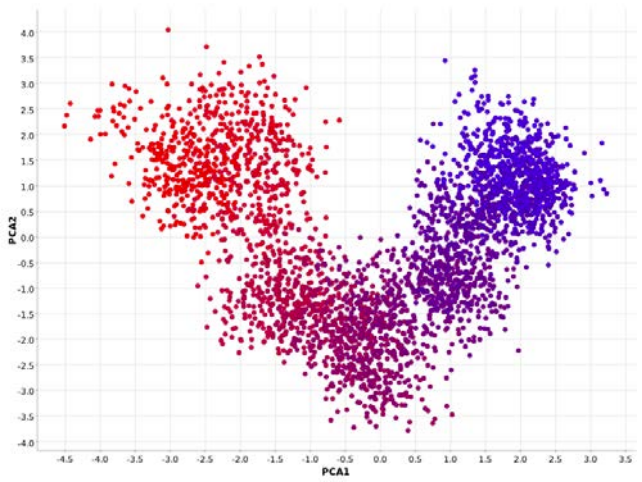
The proteins were subjected to principal component analysis for non-covalent bonding for every one of the gromacs' trajectories. Table 4 shows the results of this analysis, where red indicates the start and blue is the end of the trajectory. It is interesting to note that with ligand 397730, the protein settles into a defined motion, with little variation in this at the end of the trajectory. For 402091 and 403456 in particular, the range of the principal components is still large, indicating some instability in the types of motion of the protein.

Table 4: Schrodinger KNIME generated scatter plot projects of PCA results performed on protein 3CLpro for each of the trajectories from GROMACS.

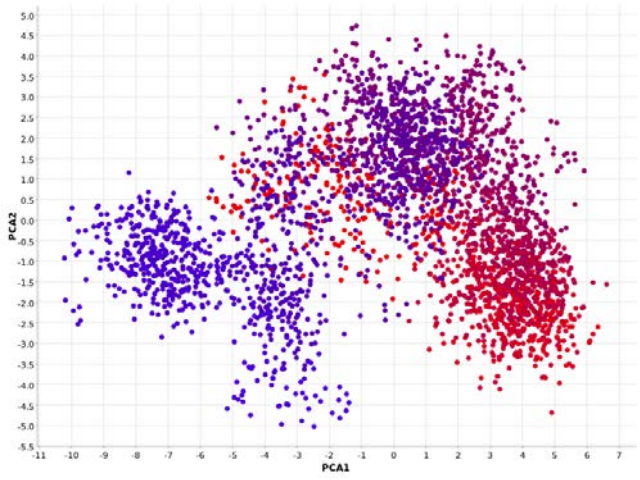
 48356	 117238
 387305	 396939



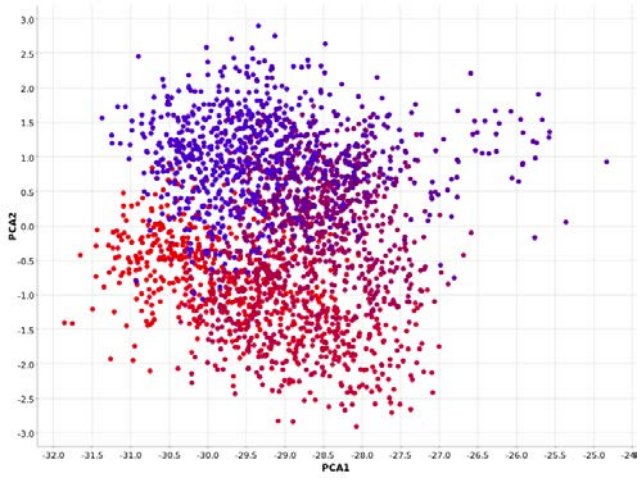
397136



397730



402091



403456

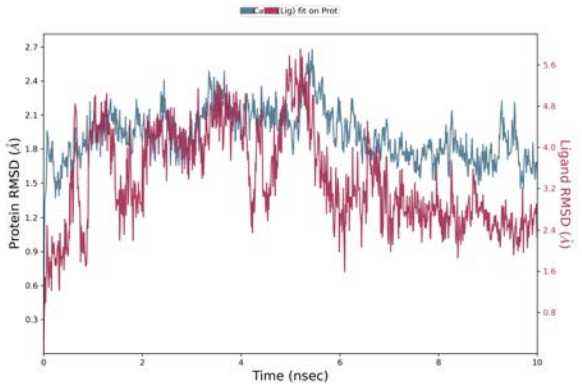
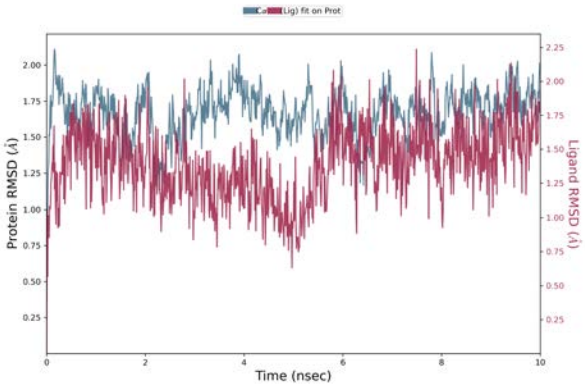
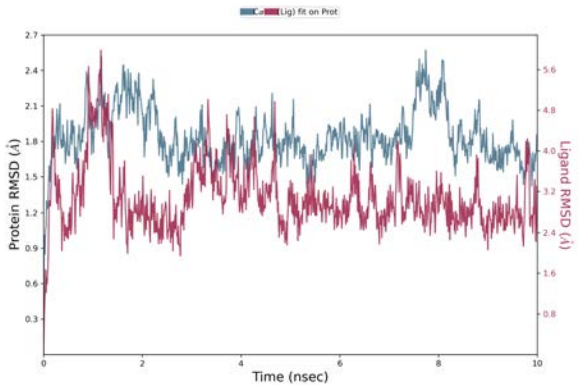
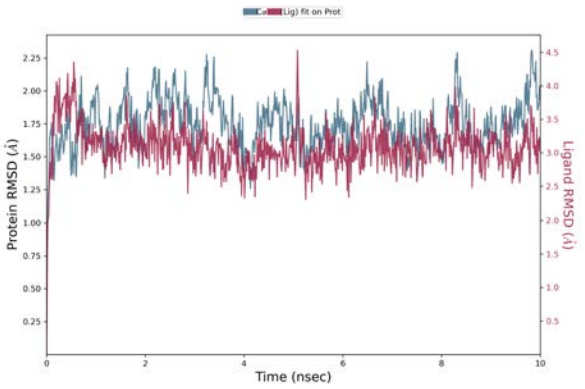
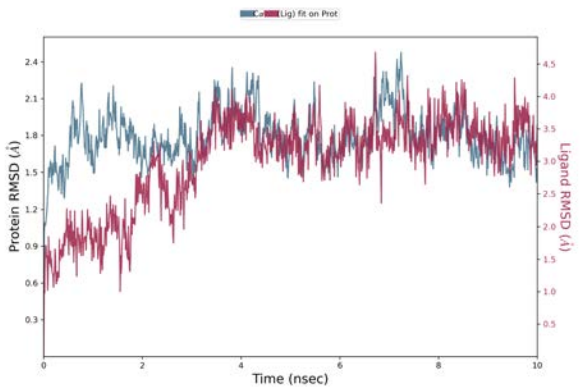
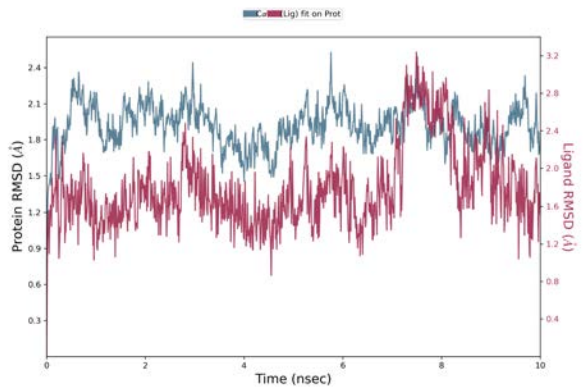
3.4.2 ANALYSIS OF COVALENT MD TRAJECTORIES

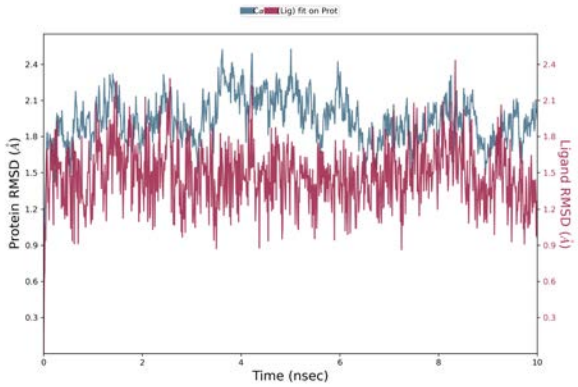
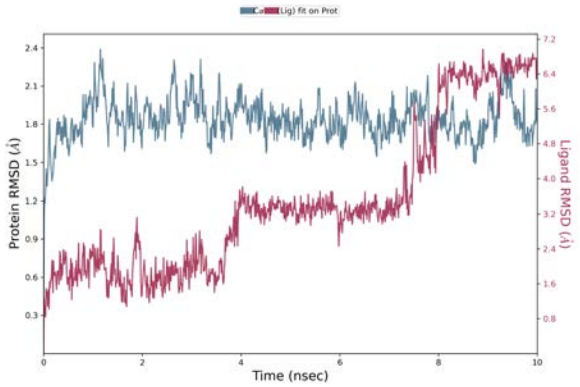
The covalently docked systems were set up for dynamics within Maestro; this included solvation, neutralization, and assignment of the OPLS force field. The results of a short 10 ns molecular dynamics simulation are highlighted here.

3.4.2.1 Protein and ligand RMSD

A dynamic environment was employed to analyze the ligand-protein complexes using 10 ns simulations. The aim of the study was to improve the choice of a 3CL pro inhibitor for possible in vivo research by extracting information on protein conformational changes and ligand dynamic interactions from the acrylonitrile-based ZINC-derived compound databases. In order to quantify the atomic position deviation and show the structural alterations that take place over the course of the 10 ns MD simulation, RMSD was also computed for covalent simulations. It is noted that the RMSD calculations of the protein are higher than those of the ligand. However, the deviations of the ligands from their first conformation appear to be higher compared to those of the protein, with the ligand 403456 in Table 5 being the best example. It is interesting to note the hike in ligand peaks on this ligand 4403456 from 8 ns to 10 ns. There are questions raised about the stability of this complex because the smaller the deviations from its original conformation, the more stable the structure, and the larger the deviation from its original conformation, the less stable the structure. The condition of the remaining complexes overall appears to be stable and can be used for further work.

Table 5: Schrodinger Maestro generated RMSD results of the protein together with the ligand. With RMSD of ligand in red and of protein in blue.

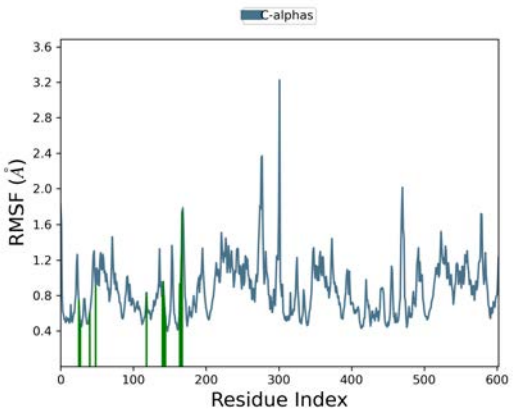
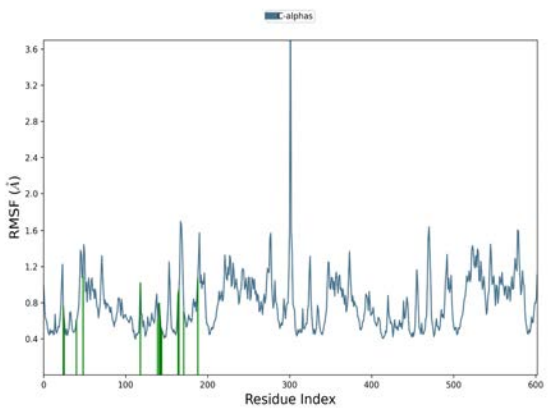
 48356	 117238
 337222	 387305
 396939	 397136

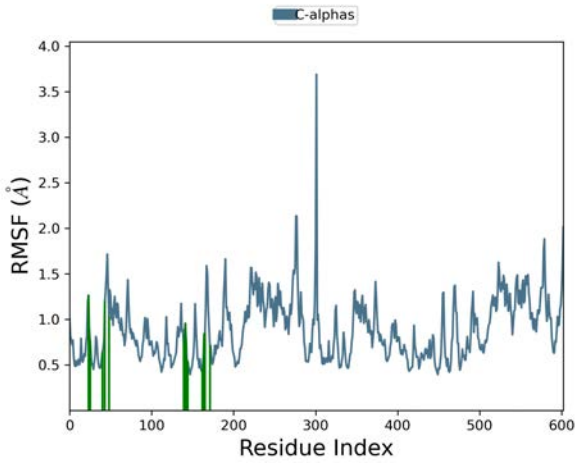
	
397730	403456

3.4.2.2 Protein RMSF

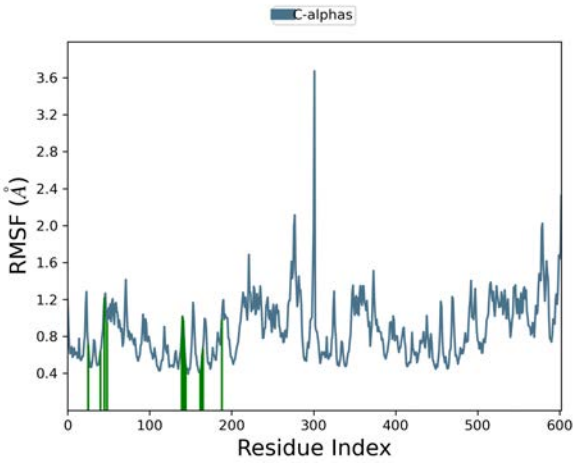
The root mean square fluctuations (RMSF) were calculated on the proteins for each of the trajectories from gromacs to capture the mobility/fluctuation about its position for each atom; it shows the flexibility of regions of the molecule. In Table 6, higher levels of flexibility of the protein are seen by the high peaks displayed in the plots. It is notable in all RMSF results that the protein is stable around 0.4 to 1.2 Å. It is interesting to observe that in all the plots, there is a notable peak at Residue Index 300.

Table 6: Schrodinger Maestro generated RMSF results of the protein during the 10 ns simulation.

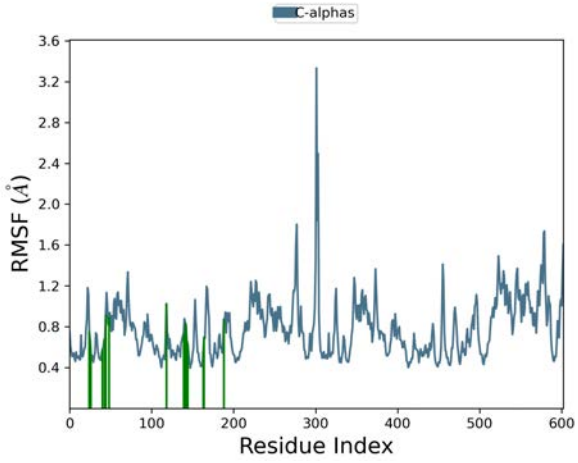
	
48356	117238



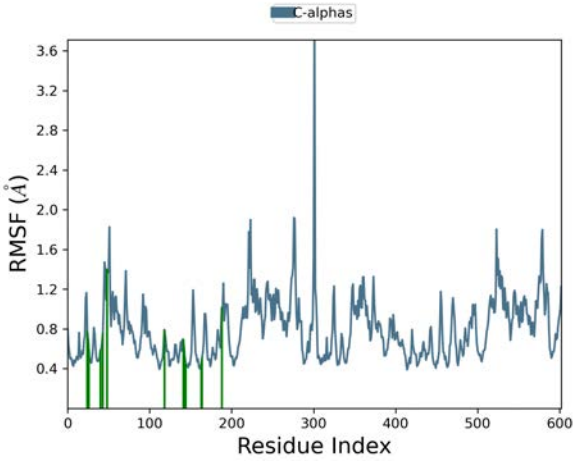
337222



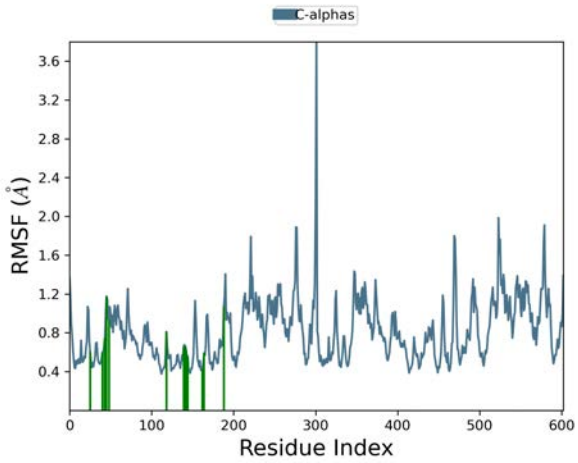
387305



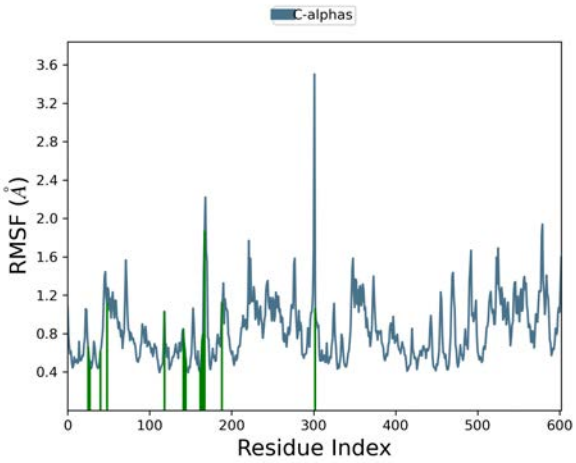
396939



397136



397730

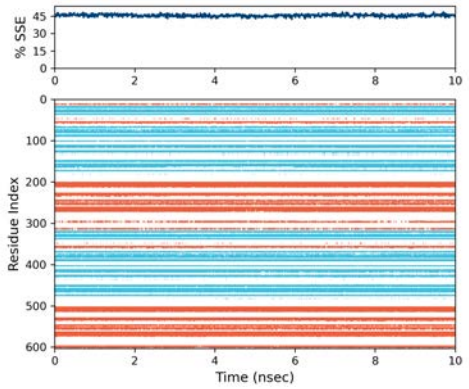
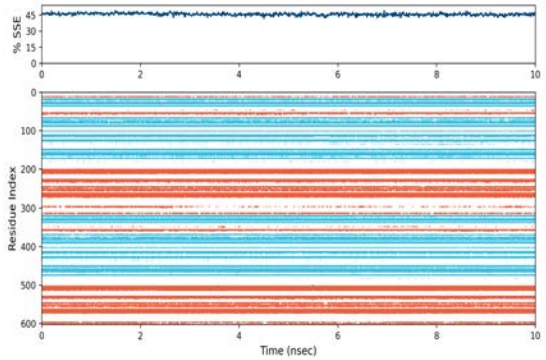
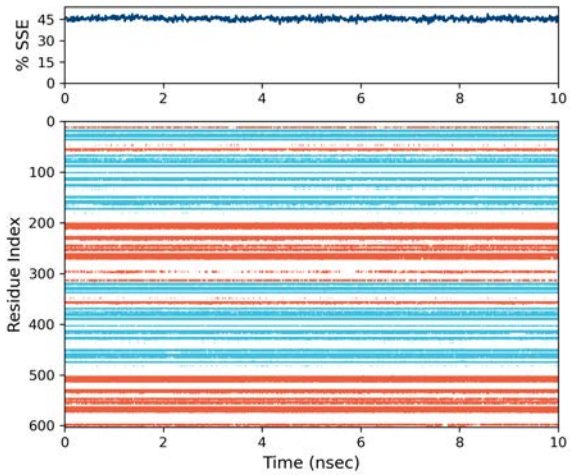
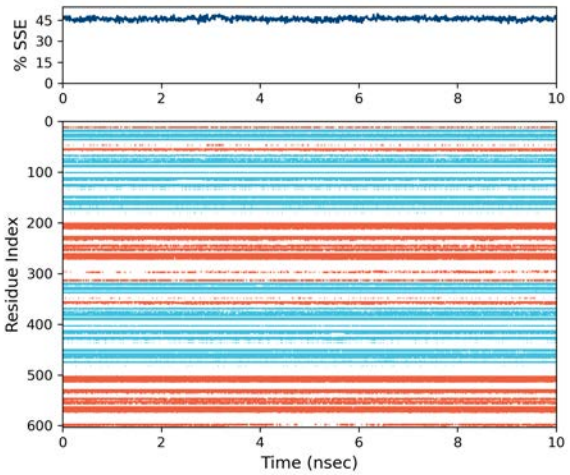


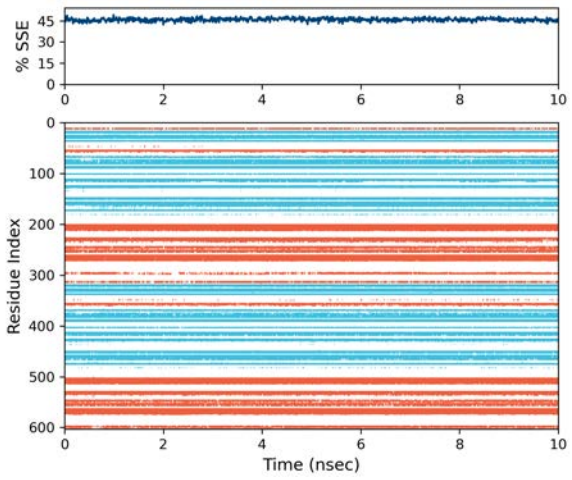
403456

3.4.2.3 Protein Secondary Structure Content Timeline

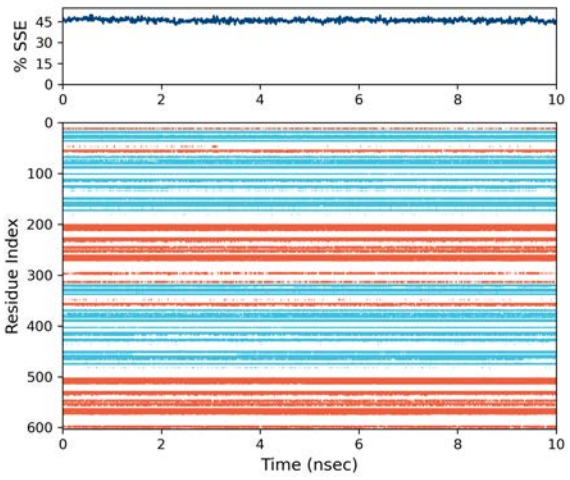
Table 7 shows the protein secondary structure content timeline during the 10 ns covalent md simulation. It is interesting to see that the covalent binding does not disrupt the secondary structure in any of the simulations. What we were aiming for was to find warhead inhibitors that prevent the function of the protein by tying up the active site, and this is done well with little external effect.

Table 7: SSE progression through 10ns dynamics.

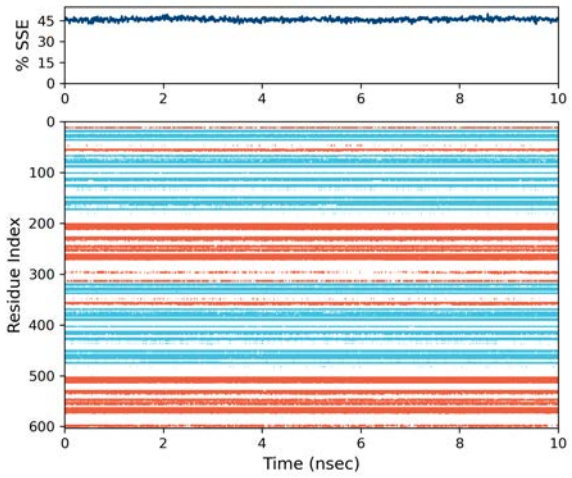
 48356	 117238
 337222	 387305



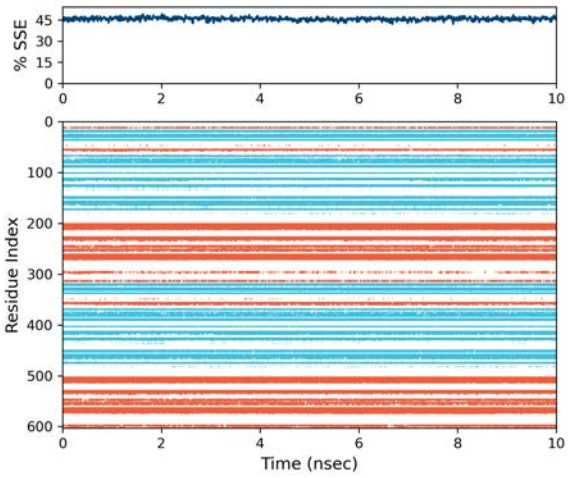
396939



397136



397730

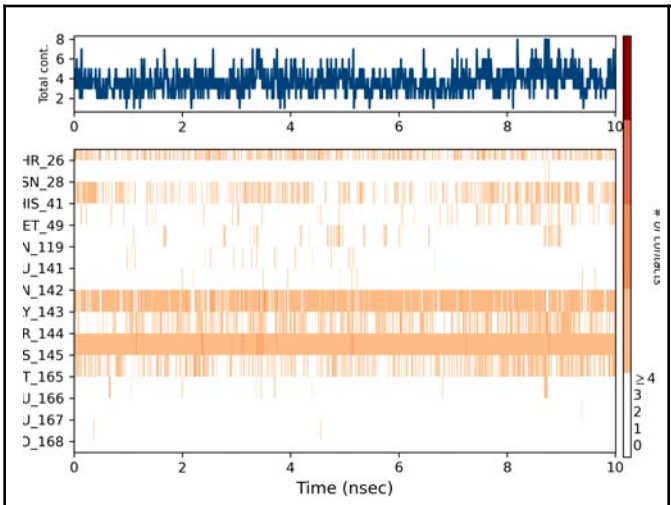


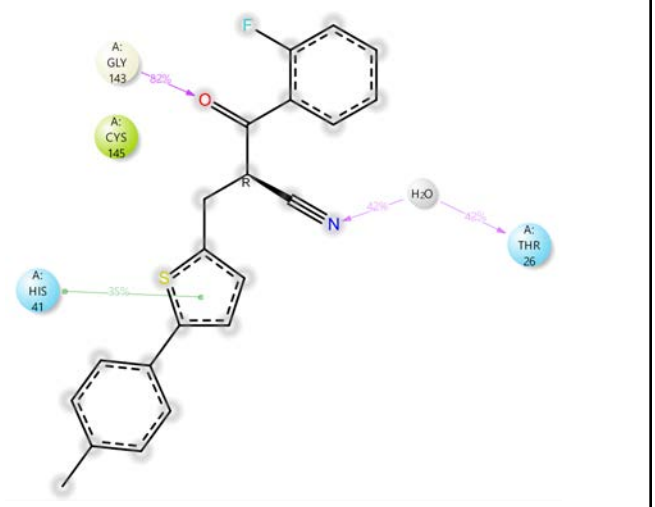
403456

3.4.2.4 Protein-Ligand Contacts (timeline and summary)

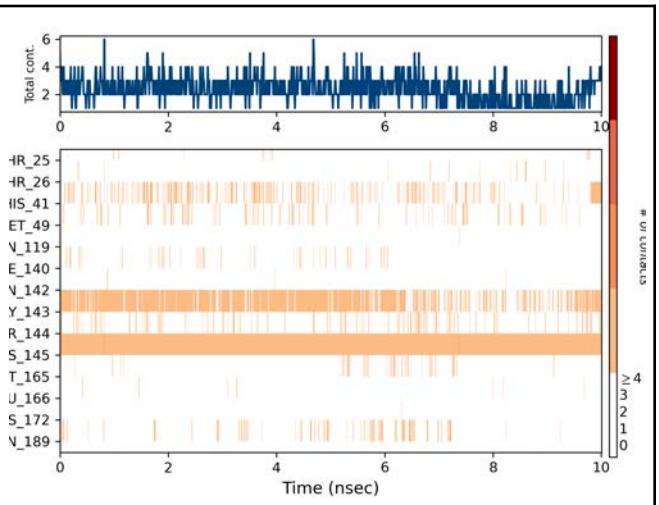
In Table 8, it was interesting to see how the covalently docked systems interacted with the protein during the short dynamics. As expected, due to the covalent bond, these ligands consistently interact with the same residues (such as residues 143 and CYS145 - expected due to the bond) during simulation.

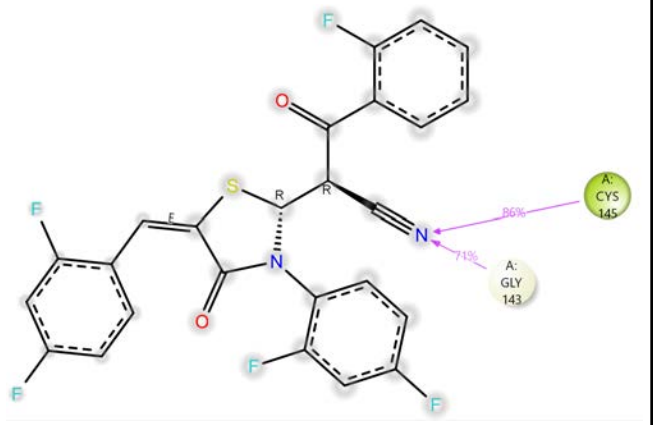
Table 8: shows how and when during the simulations the protein-ligand interactions happen, and at the bottom is a two-dimensional diagram of the protein and the ligand during the simulation.



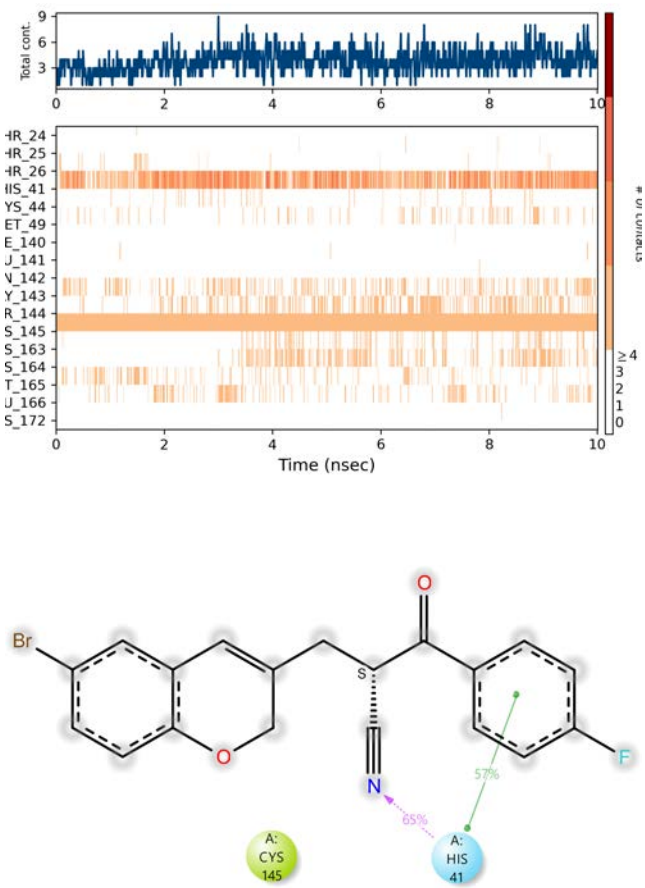


48356

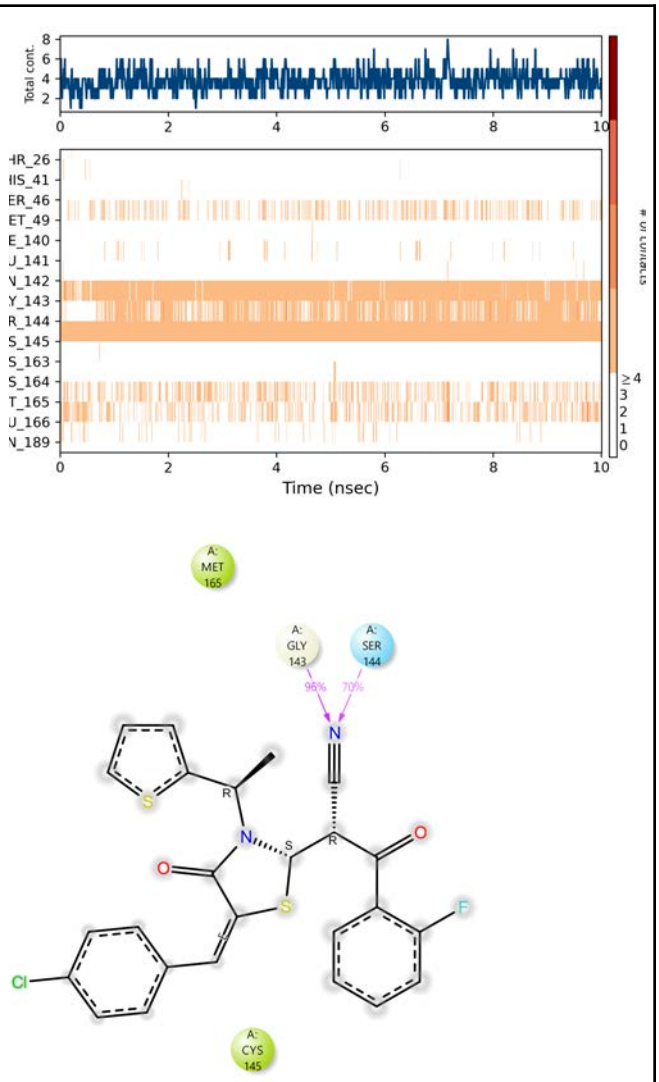




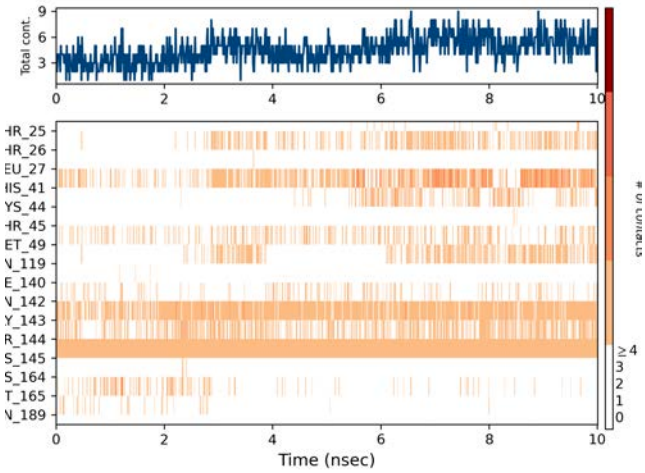
117238

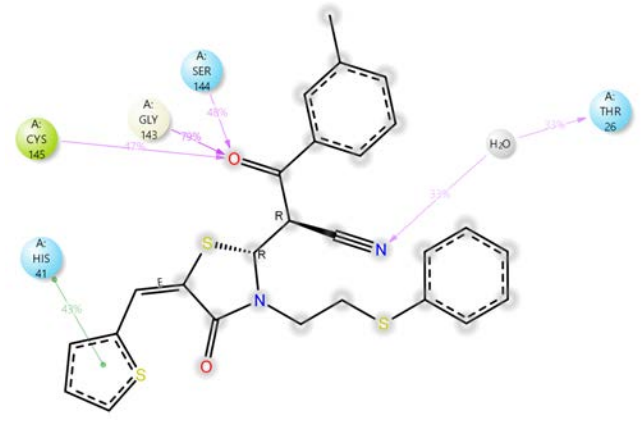


337222

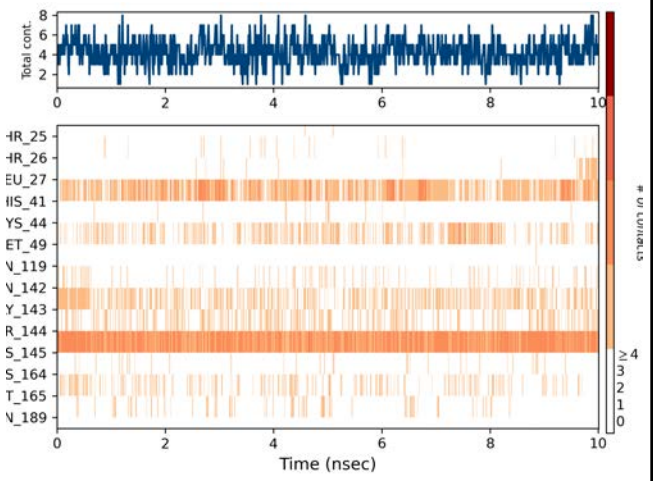


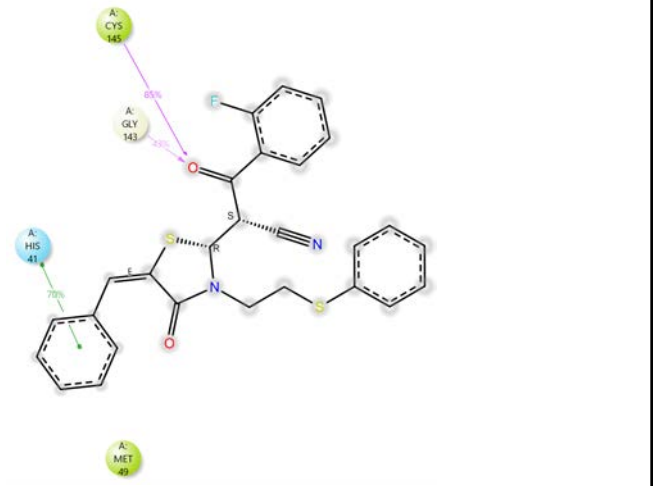
387305



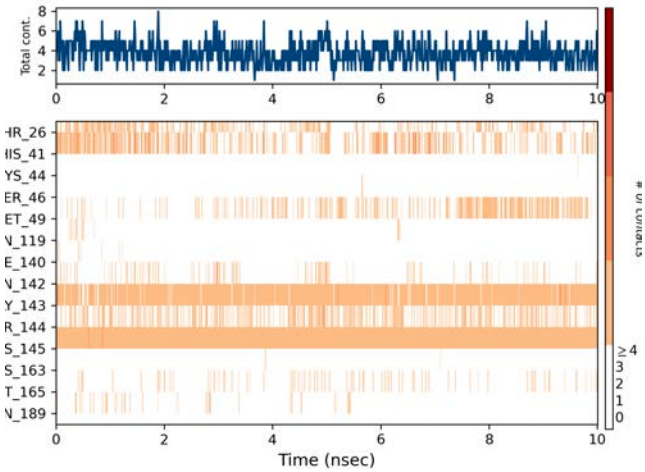


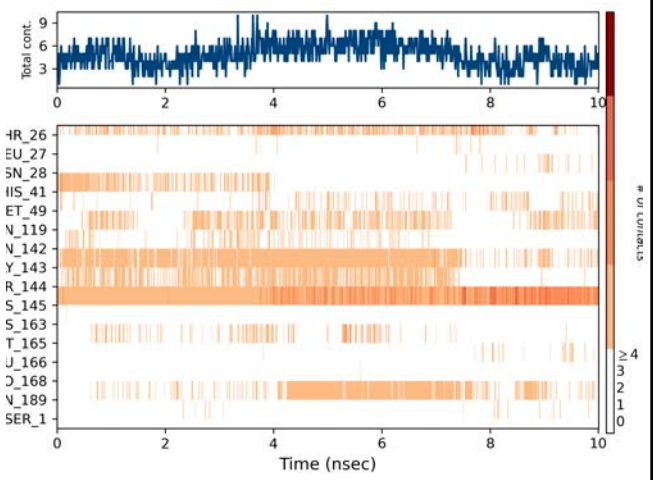
396939

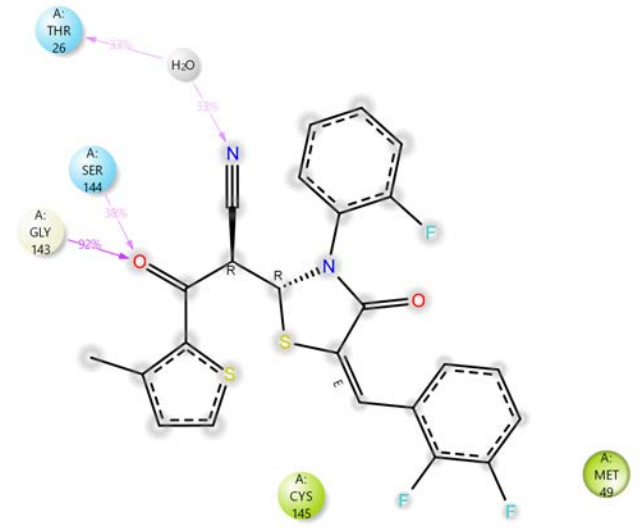
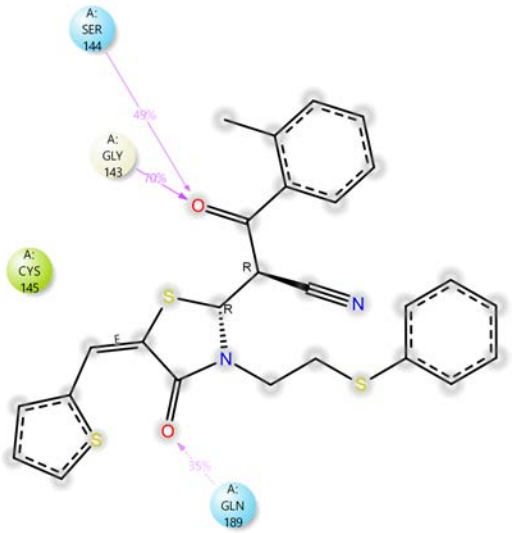




397136



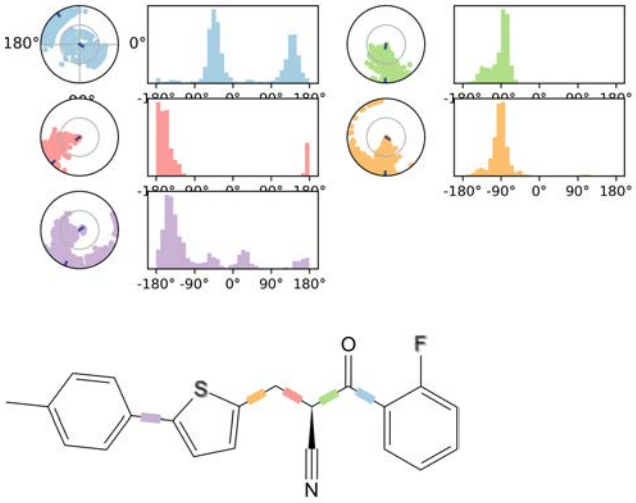
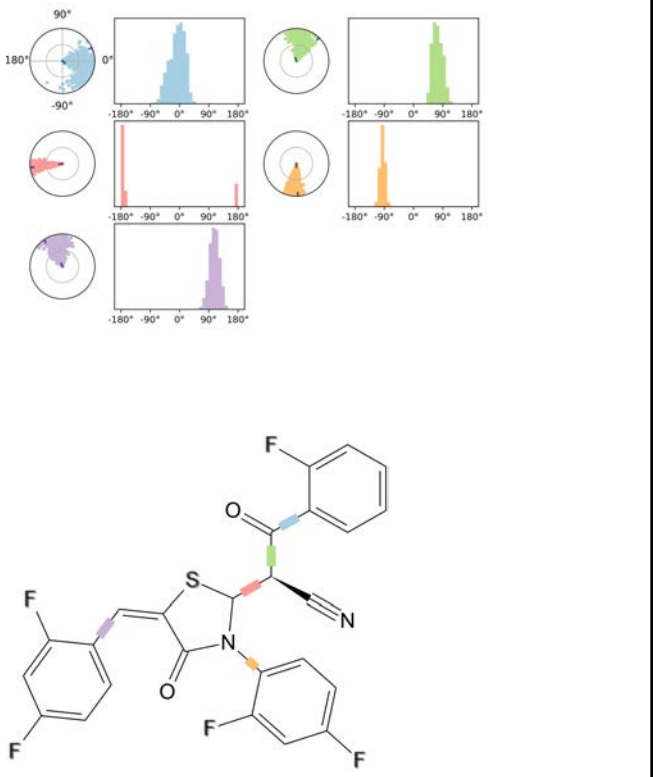


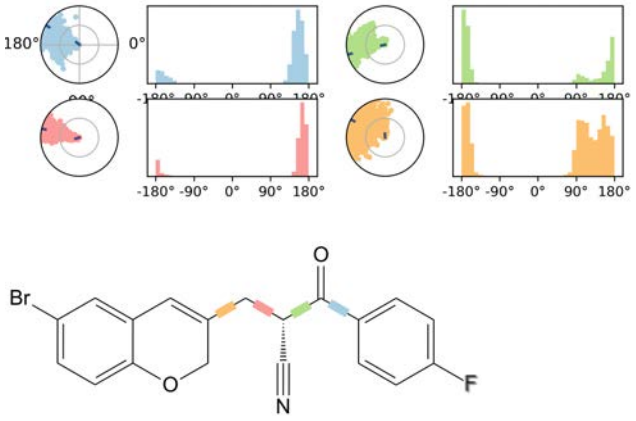
	
397730	403456

3.4.2.5 Ligand flexibility during simulation (ligand torsions)

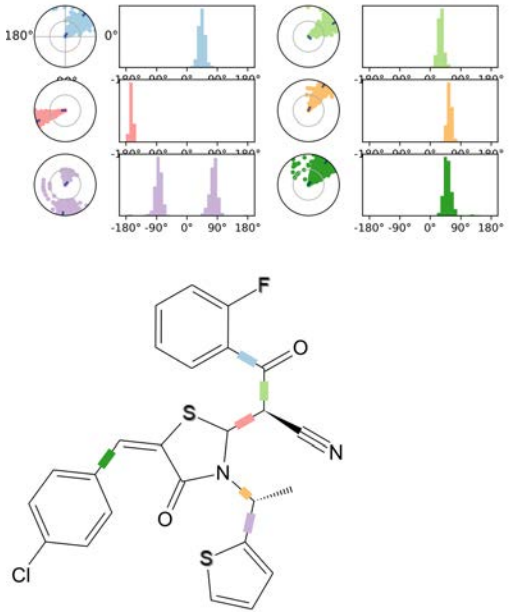
In drug design, torsion-angle scans with force fields or ab initio techniques are commonly employed to estimate the likelihood that a ligand will bind to a target protein in a particular conformation. Table 9 displays the number of occurrences of torsion angles at various torsion values, binned at 90° intervals. This data illustrates the ligands' flexibility during the 10-ns molecular dynamics simulation. It is interesting to note the major flexibility points in ligands such as 396939, 397136, and 403456. The ligand 397136 has too many torsion occurrences at so many different torsion values, and the flexibility in the region where the sulfur atom is bonded with a carbon atom (the region indicated by the color red) is very high, which may raise questions about the stability of the structure.

Table 9: Ligand torsion distributions during the 10 ns molecular dynamics simulation.

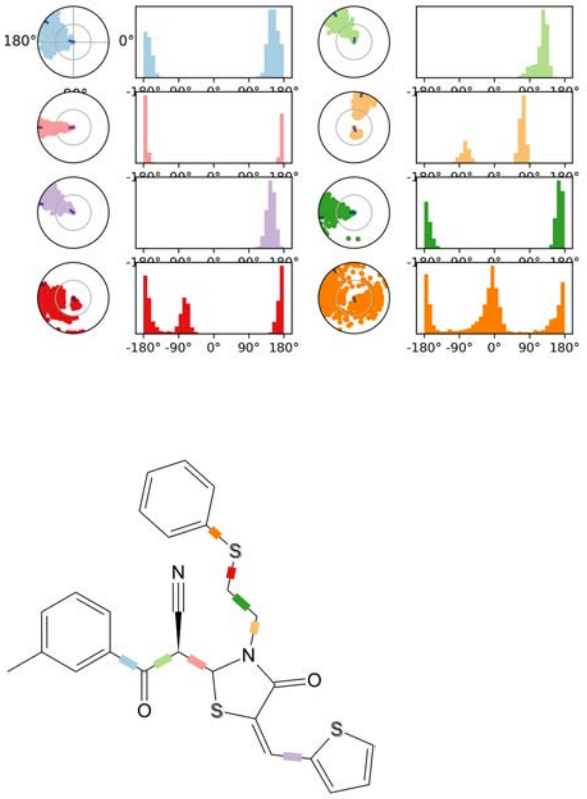
 48356	 117238



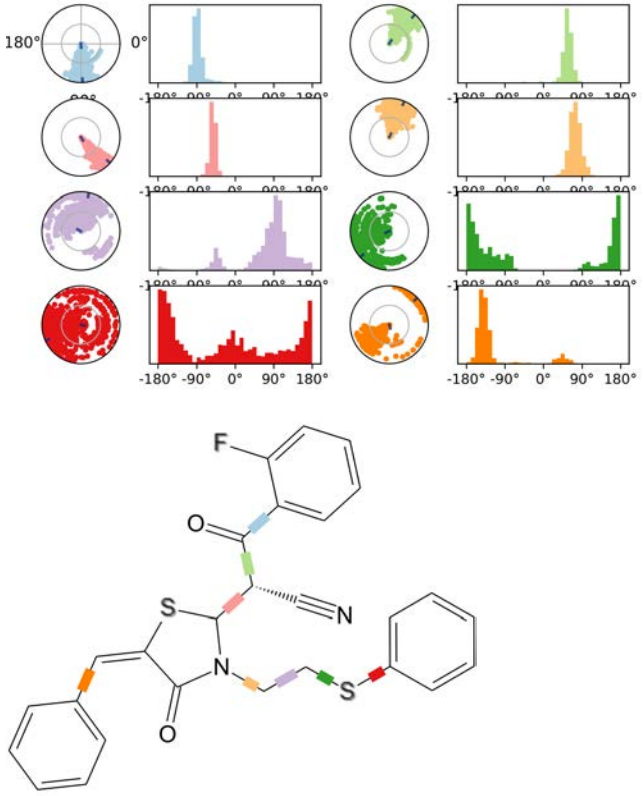
337222



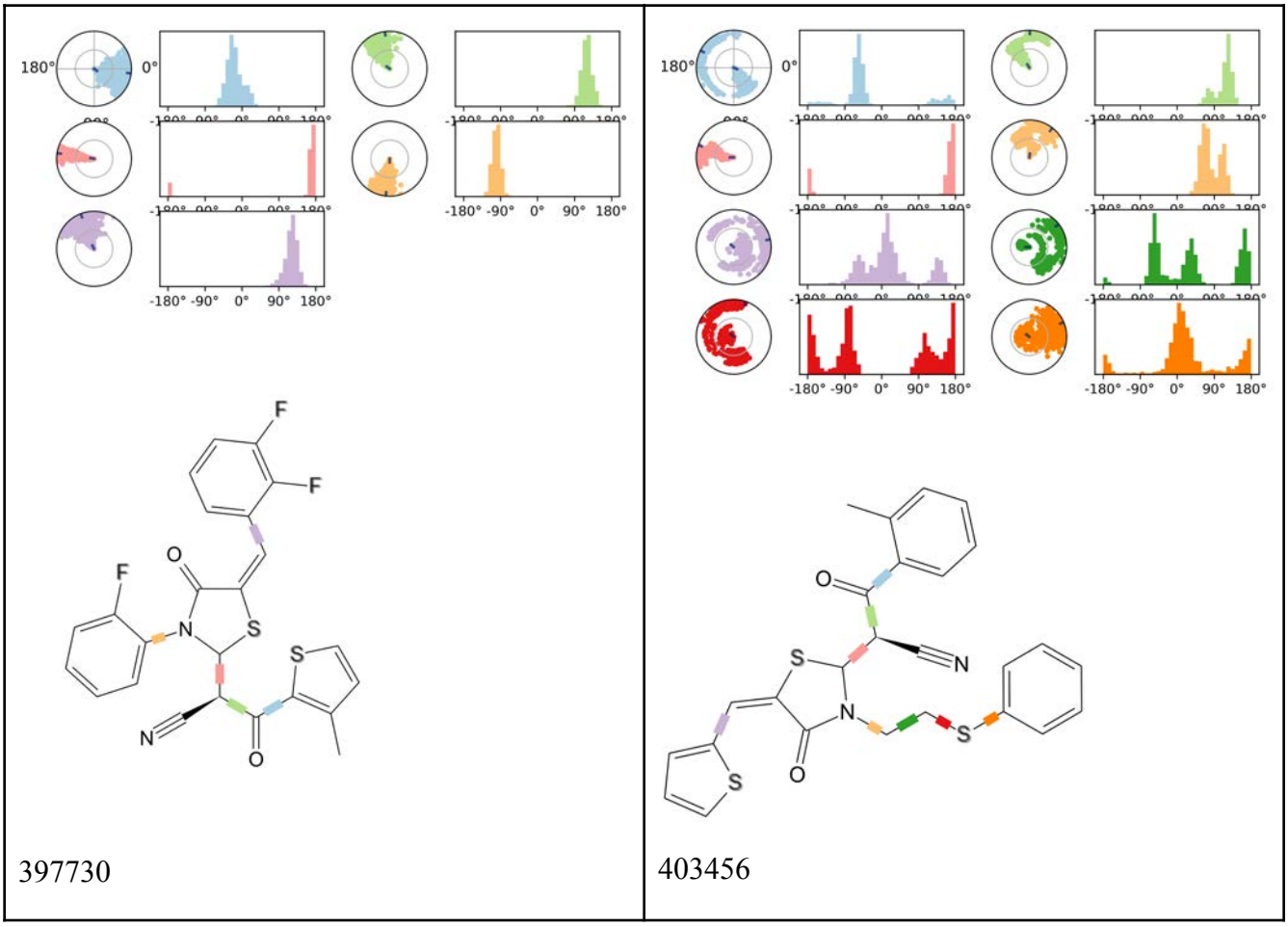
387305



396939



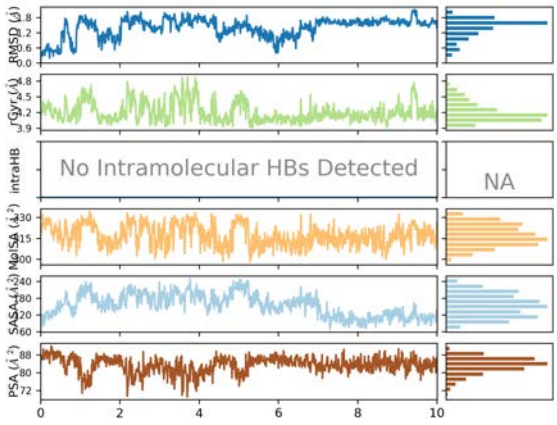
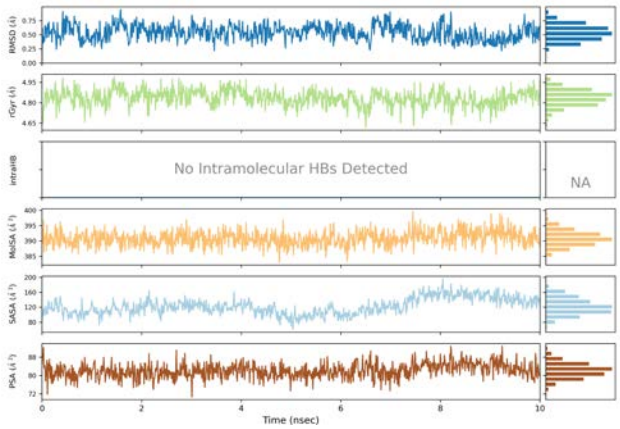
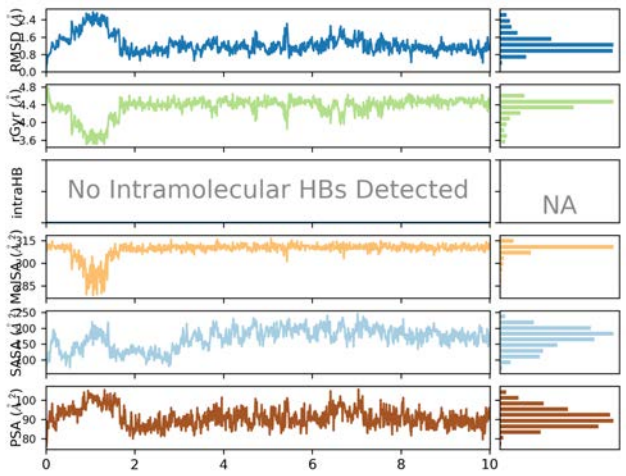
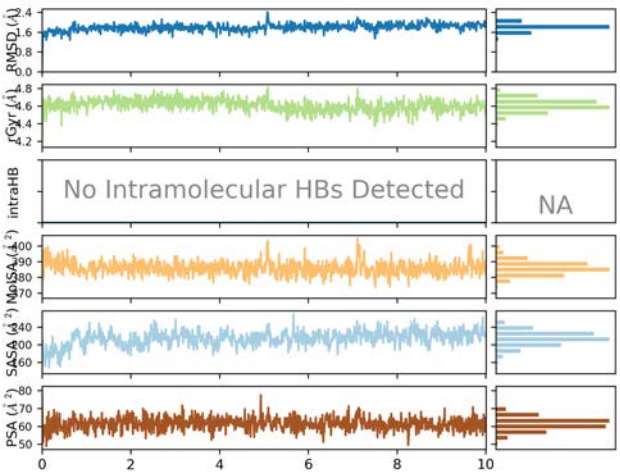
397136

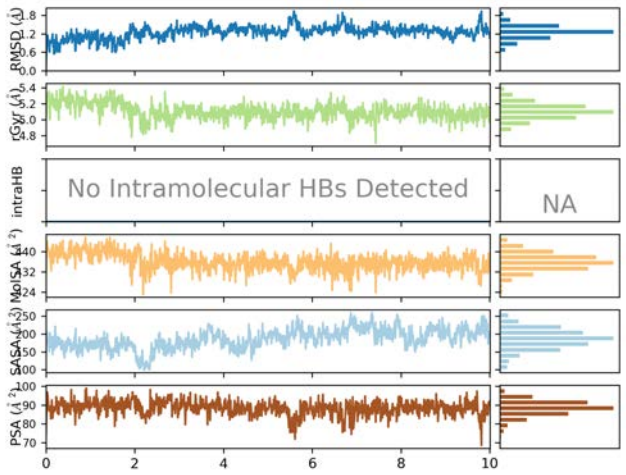


3.4.2.6 Ligand Properties during Simulation

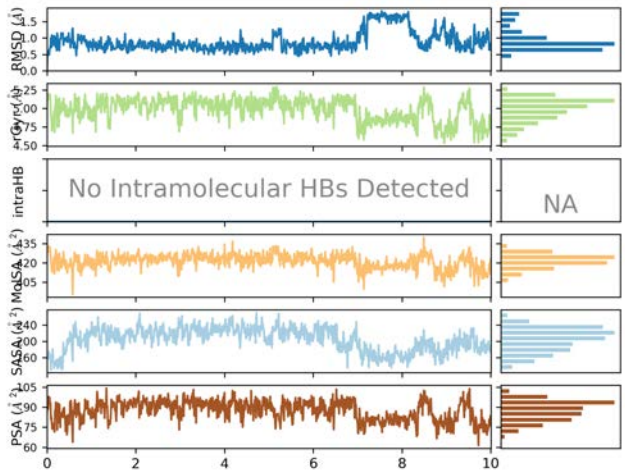
The variations in ligand properties during the Maestro covalent molecular dynamics simulation lasting 10 ns are displayed in Table 10. The properties are RMSD, Gyration, Intramolecular HBs, MolSA, SASA, and PSA, and only intramolecular HBs deviate from their initial conformation with time. Observing these properties helps in determining the stability of the ligand during the 10 ns simulation. An interesting property to note is the radius of gyration (Rg). Rg was used to evaluate the intermolecular compactness of the ligand during the simulation. Rg measures the mean distance of the atom collection from the ligand's center of mass to determine the ligand’s folding and unfolding state. In most ligands, there seems to be not much deviation from the initial conformation, and the difference between the highest andd lowest Rg values is approximately 2nm, which indicates that the ligand is well organized and stably folded (it’s in its compact form).

Table 10: Ligand properties during the 10 ns covalent molecular dynamics simulation.

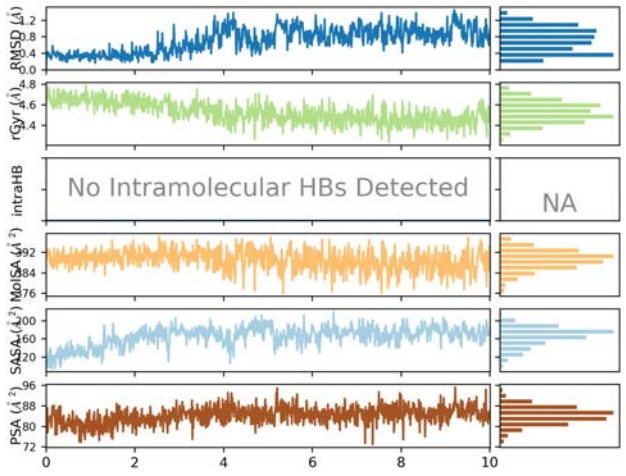
 48356	 117238
 337222	 387305



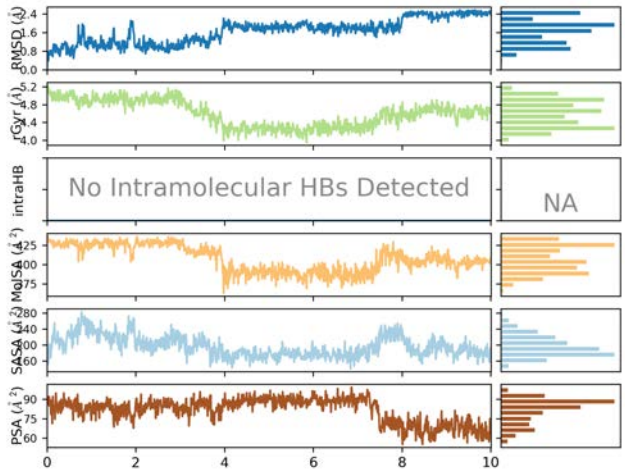
396939



397136



397730



403456

3.5 DISCUSSION

From the molecular docking these 11 molecules have been identified as excellent non-covalent binding systems with good binding energy in the active site. The covalent docking has provided systems where the binding is to CYS145. All covalently and non-covalently bound systems have been taken through to molecular dynamics and analyzed appropriately.

It is interesting to note that during the covalent dynamics, the interaction with CYS145 is clear and expected, but most ligands are held within the proximity of GLY143 and so this interaction is observed right through the molecular dynamics run. There are two notable exceptions to this. Firstly with ligand 337222, the interaction is rather with HIS41, but with ligand 397136, which occupies a huge conformational space during molecular dynamics (covalent) there is little permanent interaction with other proximal residues.

On the other hand for ligands 397730 and 387305, throughout the covalent molecular dynamics, the ligand maintained its position indicated by a small and stable ligand RMSD. In particular for ligand 397730 this is coupled with a consistent number of 6 hydrogen bonds through dynamics in the corresponding non-covalent molecular dynamics.

In terms of molecular dynamics of the non-covalent complexes, it is interesting that the behavior of the protein in PCA is very similar for ligands 397730, 48356, 117238, and 397136, indicating similar effects on the protein motion from each of these. For ligand 403456, the protein RMSF for the non-covalent dynamics was huge, this is coupled with a very compact PCA plot indicating little effect of this ligand on the protein dynamics, even though the number of hydrogen bonds through dynamics was roughly 5 consistently.

During non-covalent dynamics, there is a reorganization with ligands 397136 and 402091 where the number of hydrogen bonds drops during the simulation. An extreme case is with ligand 396939 where the number of hydrogen bonds reduces to zero during simulation. Note in this case, it is not necessarily true that this is not going to be a good inhibitor because even if the non-covalent binding is highly reversible, should a covalent bond form during residence in the active site, the covalent result will result in inactivation of the enzyme.

The variety of interactions through various simulations with just these 11 ligands provides the context for the further development of these compounds into drugs, and this will be further interesting in the context of mutations and variants of the main protease.

CHAPTER SUMMARY

This chapter describes the application of molecular dynamics simulations to ascertain the stability of the ligands in the active site, the effect of ligand binding on receptor conformation, and the inhibitory potential of compounds from docking studies, both covalently and non-covalently docked.
Through covalent and non-covalent docking investigations, lead-acrylonitrile-based compounds from the ZINC library were discovered as potential inhibitors of the 3CL pro. Additionally, they were simulated in a dynamic setting to evaluate the complexes' behavioral and structural changes for 50 nanosecond trajectories. This would improve the choice of inhibitory substances to move on with more research. Trajectory analyses were performed to evaluate and discover protein patterns, stability, interactions, and conformational changes, in addition to ligand conformational changes.

Following analyses of the molecular dynamics simulations produced by the selected conformations, two-dimensional PCA projections were produced, with each point denoting a simulation snapshot in the subspace covered by the eigenvectors. During the simulation, the sampling region along the first two eigenvectors was revealed by the clustering. PCA showed stability in most of the complexes; however, for 402091 and 403456 in particular, the range of the principal components was still large, indicating some instability in the types of motion of the protein.

Protein and ligand RMSDs were calculated for both covalent and non-covalent MD simulations. For non-covalent analysis, it was noted that the ligands deviate more than the protein in all simulations. Ligands 48356, 117238, 387305, 397136, 397730, and 402091 are stable complexes in terms of RMSD.

To enhance the selection of a 3CL pro inhibitor for potential in vivo research, trajectory analyses were carried out to extract data on protein conformational changes and ligand dynamic interactions from the acrylonitrile-based ZINC-derived compound databases.

CHAPTER 4: CONCLUSION

As of December 2023, COVID-19 cases and deaths are on the rise, continuing to put immense strain on healthcare and economies globally as it did when the pandemic was at its peak. Scientists have developed antiviral drugs through clinical trials, but new strains of the virus present a challenge. We need to invest in computational techniques to efficiently find effective drugs to treat patients and overcome this challenge. Using computational methods, this project aims to find acrylonitrile-based inhibitors of legitimate coronavirus drug targets. To accomplish this, target-based compound datasets with acrylonitrile functional groups that bind irreversibly to the SARS-CoV-2 3CLpro main protease for its inhibition were created by downloading the ZINC database locally. Acrylonitrile-based compounds from the ZINC database were identified as possible inhibitors of 3CLpro from covalent and non-covalent docking studies, which in turn provided 11 of the best-docked systems that demonstrated the potential for inhibition, which was confirmed by employing molecular dynamics simulations.

Molecular dynamic simulations applied at 50 nanosecond trajectories were used to validate the inhibitory potential of these systems, both covalently and non-covalently docked, as well as the stability at the active site and the effect of ligand binding on receptor conformation. In addition to ligand conformational changes, protein patterns, stability, interactions, and conformational changes were assessed and discovered through trajectory analyses. PCA projections were created from molecular dynamics simulations. Stability was observed in most complexes, and only 402091 and 403456 showed some instability in protein motion. Protein and ligand RMSDs were calculated for covalent and non-covalent MD simulations. During non-covalent simulations, ligands showed more deviation than proteins. Ligands 48356, 117238, 387305, 397136, 397730, and 402091 form stable complexes based on RMSD analysis, and these are ligands that we would like, in future work, to take forward for testing against the SARS-CoV-2 main protease.

As a result of this study's efforts to identify potential inhibitors of the SARS-CoV-2 main protease, 11 acrylonitrile-based ligands were found to be promising due to their ability to bind to the complex's active site and exhibit realistic complex motions. Despite the short simulation times, these compounds' binding and PCA analysis revealed that they would behave similarly in larger simulations (Yang et al., 2003), thus ensuring their stability.

REFERENCES

Abad-Zapatero, C., (2007). Ligand efficiency indices for effective drug discovery. *Expert opinion on drug discovery*, *2*(4), pp.469-488.

Abad-Zapatero C, Metz J.M., (2005). Ligand Efficiency indices as guideposts for drug discovery. Drug Discov. Today 10(7):464-469.

Abad-Zapatero C., Stamper G.F., Stoll V.S., (2006). Synergistic use of protein-crystallography and solution-phase NMR spectroscopy in structure-based drug design: strategies and tactics. In: Fragment-based Approaches in Drug Discovery. Jahnke W, Erlanson DA (Eds), Wiley-VCH Verlag GmbH & Co., Weinheim, Germany :249-266.

Abagyan, R.; Totrov, M., (2001). High-throughput docking for lead generation. Curr. Opin. Chem. Biol. 2001, 5, 375−382.

Allen, M. P.; Tildesley, D. J., (1989). Computer Simulation of Liquids; Oxford University Press: Oxford, U.K.

Anand, K., Palm, G. J., Mesters, J. R., Siddell, S. G., Ziebuhr, J., and Hilgenfeld, R. (2002). Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. EMBO J. 21 (13), 3213–3224. doi:10.1093/emboj/cdf327

Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R., and Hilgenfeld, R. (2003). Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. Science 300 (5626), 1763–1767. doi:10.1126/science.1085658

Babcock, G. J., Esshaki, D. J., Thomas, W. D. & Ambrosino, D. M., (2004). Amino acids 270 to 510 of the severe acute respiratory syndrome coronavirus spike protein are required for interaction with the receptor. J. Virol. 78, 4552–4560.

Bauer R.A., (2015). Covalent inhibitors in drug discovery: from accidental discoveries to avoided liabilities and designed therapies. Drug Discov Today 20(9):1061–1073. https://doi.org/10.1016/j.Drudis.2015.05.005

Changeux, J.-P.; Edelstein, S., (2011) Conformational selection or induced fit? 50 years of debate resolved. F1000 Biol. Rep. 2011, 3, 19.

Chatterjee, S., (2020). Understanding the nature of variations in structural sequences coding for coronavirus spike, envelope, membrane and nucleocapsid proteins of SARS-CoV-2. SSRN 2020, 1–12.

Chen, C. Y., (2007). Open reading frame 8a of the human severe acute respiratory syndrome coronavirus not only promotes viral replication but also induces apoptosis. J. Infect. Dis. 196, 405–415.

Chen, Y.N.P.; Marnett, L.J., (1989). Heme prosthetic group required for acetylation of prostaglandin-H synthase by aspirin. FASEB J. 1989, 3, 2294–2297.

Chinese, S. M. E. C., (2004). Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science 303, 1666–1669.

Costanzi, E., Kuzikov, M., Esposito, F., Albani, S., Demitri, N., Giabbai, B., Camasta, M., Tramontano, E., Rossetti, G., Zaliani, A. and Storici, P., (2021). Structural and biochemical analysis of the dual inhibition of MG-132 against SARS-CoV-2 main protease (Mpro/3CLpro) and human cathepsin-L. International journal of molecular sciences, 22(21), p.11779.

Cui, J., Li, F., and Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17 (3), 181–192. doi:10.1038/s41579-018-0118-9

Darvas, F., Dormán, G. and Papp, Á., (2000). Diversity measures for enhancing ADME admissibility of combinatorial libraries. Journal of chemical information and computer sciences, 40(2), pp.314-322.

De Vivo, M., (2011). Bridging quantum mechanics and structure-based drug design. Front. Biosci., Landmark Ed. 2011, 16, 1619−1633.

De Vita E., (2021). 10 years into the resurgence of covalent drugs. Future Med Chem 13(2):193–210. https://doi.org/10/4 fmc-2020-0236.

Desmond Molecular Dynamics System, D. E. Shaw Research, New York, NY, (2021). Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2021.

Dömling, A., and Gao, L. (2020). Chemistry and biology of SARS-CoV-2. Chem 6(6), 1283–1295. doi:10.1016/j.chempr.2020.04.023.

Durrant, J.; McCammon, J. A., (2011). Molecular dynamics simulations and drug discovery. BMC Biol. 2011, 9, 71.

Fehr A.R, Perlman S., (2015). Coronaviruses: an overview of their replication and pathogenesis, Methods Mol. Biol. 1282 (2015) 1–23.

Fischer, M.; Coleman, R. G.; Fraser, J. S.; Shoichet, B. K., (2004). Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. Nat. Chem.2014, 6, 575−583.

Frenkel, D.; Smit, B., (2001). Understanding Molecular Simulation; Academic Press, Inc.: San Diego, CA, 2001; p 638.

Gates, B. (2020). Responding to Covid-19—a once-in-a-century pandemic? New Engl. J. Med. 382 (18), 1677–1679. doi:10.1056/NEJMp2003762

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P., and Blinov, V. M. (1989). Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis. Nucleic Acids Res. 17 (12), 4847–4861. doi:10.1093/nar/17.12.4847

Guan, Y., (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302, 276–278 (2003).

Harvey, M. J.; De Fabritiis, G., (2012). High-throughput molecular dynamics: the powerful new tool for drug discovery. Drug Discovery Today 2012, 17, 1059−1062.

Hoffmann, (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor, Cell (2020), PDF Suprewicz et al.: Vimentin binds to SARS-CoV-2 spike protein and antibodies targeting extracellular vimentin block in vitro uptake of SARS-CoV-2 virus-like particles, BioRxiv preprint (2021),

Holmes, K. V., and Lai, M. (1996). Coronaviridae: the viruses and their replication. Fields Virol. 1, 1075–1093.

Hopkins, A., Keserü, G., Leeson, P., (2014) The role of ligand efficiency metrics in drug discovery. Nat Rev Drug Discov 13, 105–121. https://doi.org/10.1038/nrd4163

Hu, B., (2017). Discovery of a rich gene pool of SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLOS Pathog. 13, e1006698.

Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R., (1997). Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol. 1997, 267, 727–748.

Jorgensen, W. L., (2004). The many roles of computation in drug discovery. Science 2004, 303, 1813−1818.

Kan, B., (2005). Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. J. Virol. 79, 11892–11900.

Kitchen, D.B., Decornez, H., Furr, J.R. and Bajorath, J., (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. Nature reviews Drug discovery, 3(11), pp.935-949.

Kollman, P., (1993). Free energy calculations: applications to chemical and biochemical phenomena. Chemical reviews, 93(7), pp.2395-2417.

Kožić, M. and Bertoša, B., 2024. Trajectory maps: molecular dynamics visualization and analysis. *NAR Genomics and Bioinformatics*, *6*(1), p.lqad114.

Kuntz ID, Chen K, Sharp KA, Kollman P.A., (1999). "The maximal affinity of ligands". Proceedings of the National Academy of Sciences of the United States of America. **96** (18): 9997–10002.

Lai, M.M.; Stohlman, S.A., (1981). Comparative analysis of RNA genomes of mouse hepatitis viruses. J. Virol. 1981, 38, 661–670

Lecomte, M.; Laneuville, O.; Ji, C.; Dewitt, D.L.; Smith, W.L., (1994). Acetylation of human prostaglandin endoperoxide synthase-2 (cyclooxygenase-2) by aspirin. J. Biol. Chem. 1994, 269, 13207–13215.

Lee, H. J., Shieh, C. K., Gorbalenya, A. E., Koonin, E. V., La Monica, N., Tuler, J., et al. (1991). The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative proteases and RNA polymerase. Virology 180 (2), 567–582. doi:10.1016/0042-6822(91)90071-i

Levitt, M.; Warshel, A., (1975). Computer simulation of protein folding. Nature 1975, 253, 694−698.

Le, T. M., (2007). Expression, post-translational modification and biochemical characterization of proteins encoded by subgenomic mRNA8 of the severe acute respiratory syndrome coronavirus. FEBS J. 274, 4211–4222.

Li S., Yuan L., Dai G., Chen R.A., Liu D.X., Fung T.S., (2020). Regulation of the ER stress response by the Ion Channel activity of the infectious bronchitis coronavirus envelope protein modulates Virion release, apoptosis, viral fitness, and pathogenesis, Front. Microbiol. 10 (2020) 3022.

Li, W. H., (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature 426, 450–454.

Li, W. H., (2005). Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J. 24, 1634–1643.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The lancet 395 (10224), 565–574. doi:10.1016/S0140-6736(20)30251-8

Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews, 23(1-3), pp.3-25.

Mah R., Thomas J.R., Shafer C.M., (2014). Drug discovery considerations in the development of covalent inhibitors. Bioorg Med Chem Lett 24(1):33–39. https://doi.org/10.1016/j.bmcl.2013.10.003

Masters, P. S. & Perlman, S., (2013). in Fields Virology Vol. 2 (eds Knipe, D. M. & Howley, P. M.) 825–858 (Lippincott Williams & Wilkins, 2013).

McCammon, J. A.; Gelin, B. R.; Karplus, M., (1997). Dynamics of folded proteins. Nature 1977, 267, 585−590.

Meng, X.Y., Zhang, H.X., Mezei, M. and Cui, M., (2011). Molecular docking: a powerful approach for structure-based drug discovery. Current computer-aided drug design, 7(2), pp.146-157.

Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J., AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. J. Comput. Chem. 2009, 30, 2785–2791.

Ohno, K., Nagahara, Y., Tsunoyama, K. and Orita, M., (2010). Are there differences between launched drugs, clinical candidates, and commercially available compounds?. Journal of chemical information and modeling, 50(5), pp.815-821.

Oostra, M., de Haan, C. A. & Rottier, P. J., (2007). The 29-nucleotide deletion present in human but not in animal severe acute respiratory syndrome coronaviruses disrupts the functional expression of open reading frame 8. J. Virol. 81, 13876–13888.

Phan, L.T., Nguyen, T.V., Luong, Q.C., Nguyen, T.V., Nguyen, H.T., Le, H.Q., Nguyen, T.T., Cao, T.M. and Pham, Q.D., 2020. Importation and human-to-human transmission of a novel coronavirus in Vietnam. New England Journal of Medicine, 382(9), pp.872-874.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: progress, challenges and recommendations. Nat. Rev. Drug Discov. 18, 41–58. doi: 10.1038/nrd.2018.168

Qu, X. X., (2005). Identification of two critical amino acid residues of the severe acute respiratory syndrome coronavirus spike protein for its variation in zoonotic tropism transition via a double substitution strategy. J. Biol. Chem. 280, 29588–29595.

Rahman, A. and Sarkar, A., 2019. Risk factors for fatal middle east respiratory syndrome coronavirus infections in Saudi Arabia: analysis of the WHO Line List, 2013–2018. American journal of public health, 109(9), pp.1288-1293.

Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., (1996). A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. 1996, 261, 470–489

Robertson, J.G., (2005). Mechanistic basis of enzyme-targeted drugs. Biochemistry 2005, 44, 8918–8918.

Robertson, J.G., (2007). Enzymes as a special class of therapeutic target: Clinical drugs and modes of action. Curr. Opin. Struct. Biol. 2007, 17, 674–679.

Roth, G.J.; Stanford, N.; Majerus, P.W., (1975). Acetylation of prostaglandin synthase by aspirin. Proc. Natl. Acad. Sci. USA 1975, 72, 3073–3076.

Shereen, M.A., Khan, S., Kazmi, A., Bashir, N. and Siddique, R., (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. Journal of Advanced Research.

Shi, J., Wei, Z., and Song, J. (2004). Dissection study on the severe acute respiratory syndrome 3C-like protease reveals the critical role of the extra domain in dimerization of the enzyme: defining the extra domain as a new target for design of highly specific protease inhibitors. J. Biol. Chem. 279 (23), 24765–24773. doi:10.1074/jbc.M311744200

Simonson, T., Archontis, G. and Karplus, M., (2002). Free energy simulations come of age: Protein−ligand recognition. Accounts of chemical research, 35(6), pp.430-437.

Singh J, Petter RC, Baillie TA, Whitty A (2011) The resurgence of covalent drugs. Nat Rev Drug Discov 10(4):307–317. https://doi.org/10.1038/nrd3410

Smith, A.J.T.; Zhang, X.; Leach, A.G.; Houk, K.N., (2009). Beyond Picomolar Affinities: Quantitative Aspects of Noncovalent and Covalent Binding of Drugs to Proteins. J. Med. Chem. 2009, 52, 225–233.

Song, H. D., (2005) Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc. Natl Acad. Sci. USA 102, 2430–2435.

Sung, S. C., Chao, C. Y., Jeng, K. S., Yang, J. Y. & Lai, M. M., (2009). The 8ab protein of SARS-CoV is a luminal ER membrane-associated protein and induces the activation of ATF6. Virology 387, 402–413.

Trott, O.; Olson, A.J., (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. 2010, 31, 455–461.

Tu, C., (2004). Antibodies to SARS coronavirus in civets. Emerg. Infect. Dis. 10, 2244–2248 (2004).

Vogt, A. D.; Di Cera, E., (2012). Conformational selection or induced fit? A critical appraisal of the kinetic mechanism. Biochemistry 2012, 51, 5894−5902

Wang, M., (2004). [Analysis on the risk factors of severe acute respiratory syndromes coronavirus infection in workers from animal markets]. Zhonghua Liu Xing Bing Xue Za Zhi 25, 503–505 (2004).

Wells, I.; Marnett, L.J., (1992). Acetylation of prostaglandin endoperoxide synthase by n-acetylimidazole Comparison to acetylation by aspirin. Biochemistry 1992, 31, 9520–9525.

Wong, H. H., (2018). Accessory proteins 8b and 8ab of severe acute respiratory syndrome coronavirus suppress the interferon signaling pathway by mediating ubiquitin-dependent rapid degradation of interferon regulatory factor 3. Virology 515, 165–175.

Wong, S. K., Li, W. H., Moore, M. J., Choe, H. & Farzan, M. A., (2004). 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2. J. Biol. Chem. 279, 3197–3201.

World Health Organization, (2020). Pneumonia of unknown cause–China. 2020. Available at: who. int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/. Accessed April, 1.

Xu, H. F., (2004). [An epidemiologic investigation on infection with severe acute respiratory syndrome coronavirus in wild animals' traders in Guangzhou]. Zhonghua Yu Fang Yi Xue Za Zhi 38, 81–83 (2004).

Xu, J. and Stevenson, J., (2000). Drug-like index: a new approach to measure drug-like compounds and their diversity. Journal of Chemical Information and Computer Sciences, 40(5), pp.1177-1187.

Xue, X., Yang, H., Shen, W., Zhao, Q., Li, J., Yang, K., et al. (2007). Production of authentic SARS CoV M(pro) with enhanced activity: application as a novel tag-cleavage endopeptidase for protein overproduction. J. Mol. Biol. 366 (3), 965–975. doi:10.1016/j.jmb.2006.11.073

Yang, H., Yang, M., Ding, Y., Liu, Y., Lou, Z., Zhou, Z., et al. (2003). The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. Proc. Natl. Acad. Sci. USA 100 (23), 13190–13195. doi:10.1073/pnas.1835675100

Zhang, S.; Shi, Y.; Jin, H.; Liu, Z.; Zhang, L.; Zhang, L., (2009). Covalent complexes of proteasome model with peptide aldehyde inhibitors MG132 and MG101: Docking and molecular dynamics study. J. Mol. Model. 2009, 15, 1481–1490.

Ziebuhr, J., Snijder, E. J., and Gorbalenya, A. E. (2000). Virus-encoded proteinases and proteolytic processing in the Nidovirales. J. Gen. Virol. 81 (4), 853–879. doi:10.1099/0022-1317-81-4-853

APPENDIX

RDKit Script for filtering

```
from rdkit import Chem
from rdkit.Chem import Draw
import os
molecules=[]
for file in os.listdir("./"):
   if file.endswith(".sdf"):
      print(file)
      subset=Chem.SDMolSupplier(file)
      for m in subset:
         molecules.append(m)
```

Filtered ligands

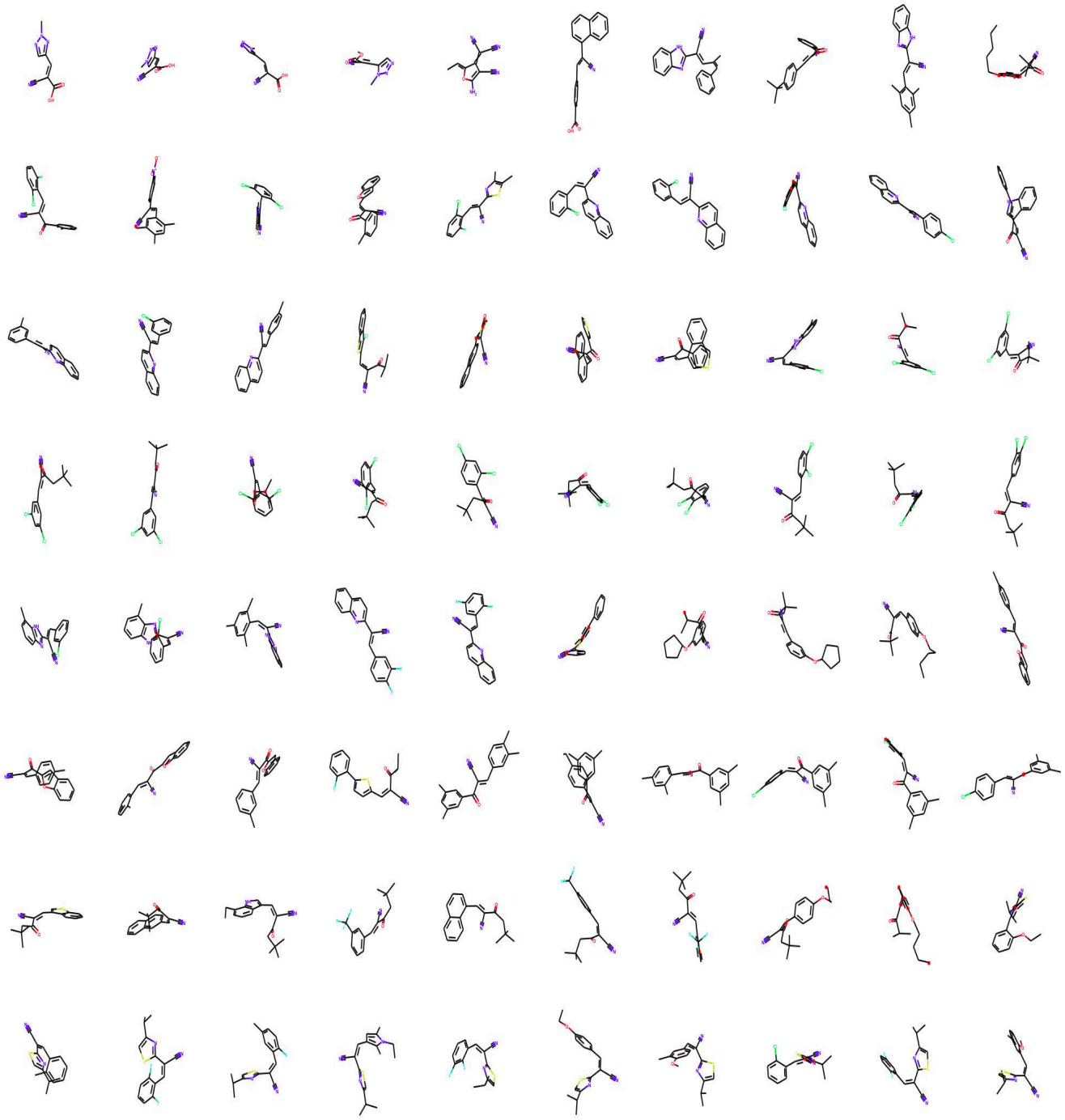


Figure 37: Shows 80 out of the 403804 high-throughput screening compounds with acrylonitrile functional group from the ZINC database before they were filtered.

RDKIT generation of properties:

```
from rdkit.Chem import QED,rdMolDescriptors,rdmolops,AllChem,Crippen
import pandas as pd
##want a dataframe to contain all data
properties=pd.DataFrame(columns=
["molno","mw","alogp","hba","hbd","psa","rotb","arom","atcount",
 "molmr","charge","rings","QED","preSEI","preBEI","alerts"])

#rdkit.org/docs/source/rdkit.Chem.QED.html
ALERTS=7
ALOGP=1
AROM=6
HBA=2
HBD=3
MW=0
PSA=4
ROTB=5

total=len(molecules)

for i in range(total): #not 10 but total for whole set
   props=QED.properties(molecules[i])
   QEDv=QED.default(molecules[i])
   rings=rdMolDescriptors.CalcNumRings(molecules[i])
   charge=rdmolops.GetFormalCharge(molecules[i])
   molmr=Crippen.MolMR(molecules[i])
   count=molecules[i].GetNumAtoms()
   heavy=molecules[i].GetNumHeavyAtoms()
   properties=properties.append(
   {"molno":i,"mw":props[MW],"alogp":props[ALOGP],"hba":props[HBA],"hbd":props[HBD],"psa":props[PSA]

,"rotb":props[ROTB],"arom":props[AROM],"QED":QEDv,"alerts":props[ALERTS],"rings":rings,"charge":charge
   ,"molmr":molmr,"atcount":count, "heavyatm":heavy},
   ignore_index=True)
   #print(i,"molecular weight,",props[MW])



import os,sys
from rdkit import Chem, DataStructs
```

```
from rdkit.Chem import AllChem,QED

ref=Chem.MolFromSmiles("C=CC#N")
unwanted=Chem.MolFromSmiles("CC(C)=CC#N")

rootdir = "/home/zinc"
for subdir in directorylist0:
    #os.system("touch /home/lobb/"+subdir)
    files=os.listdir(rootdir+"/"+subdir)
    for filename in files:
        if "txt" in filename:
            print(rootdir+"/"+subdir+"/"+filename)
            with open(rootdir+"/"+subdir+"/"+filename, 'r') as src:
                lines=src.readlines()
                for line in lines:
                    if not "smiles" in line:
                        words=line.split()
                        x=Chem.MolFromSmiles(words[0])
                        try:
                            if(x.HasSubstructMatch(ref) and not (x.HasSubstructMatch(unwanted))):
                                substructures.append(x)
                                print("Appending, substructures: "+words[0])
                            val=QED.default(x)
                            if val > 0.947:
                                qed.append(x)
                                print("Appending, qed:"+str(val))
                        except:
                            print("error in molecule " + words[0])
```

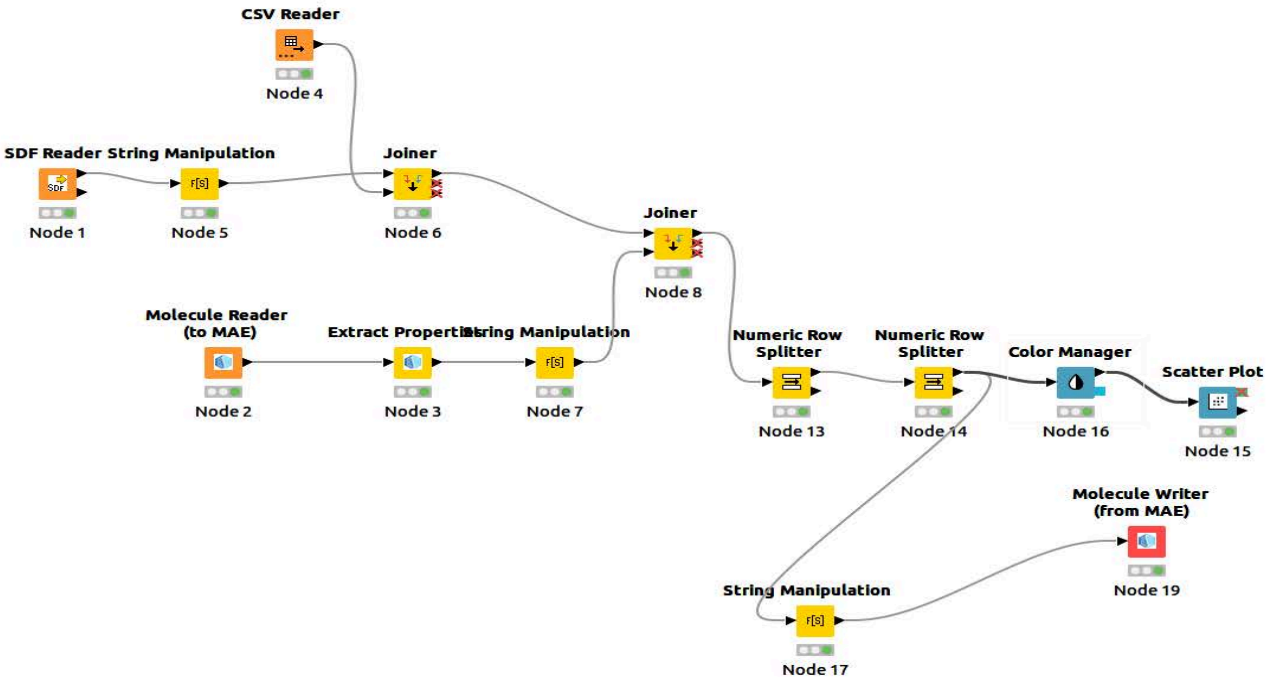


Figure 38: Knime workflow of covalent docking analysis and visualization.